ifo WORKING PAPERS

409
2024
May 2024

Machine Learning Based Linkage of Company Data for Economic Research: Application to the EBDC Business Panels

Valentin Reich

ifo INSTITUTE
Leibniz Institute for Economic Research
at the University of Munich

Imprint:

ifo Working Papers
Publisher and distributor: ifo Institute – Leibniz Institute for Economic Research at the University of Munich
Poschingerstr. 5, 81679 Munich, Germany
Telephone + 49(0)89 9224 0, Telefax +49(0)89 985369, email ifo@ifo.de
www.ifo.de

An electronic version of the paper may be downloaded from the ifo website
www.ifo.de

Machine Learning Based Linkage of Company Data for Economic Research: Application to the EBDC Business Panels

Abstract

This article presents a comprehensive approach to probabilistic linkage of German company data using Machine Learning and Natural Language Processing techniques. Here, the long-running ifo Institute surveys are linked to financial information in the Orbis database by addressing the unique challenges of company data linkage, such as corporate structures and linguistic nuances in company names. Compared to a previous linkage, the approach achieves improved match rates and is able to re-evaluate existing matches. This article contributes best practice advice for company data linkage and serves as a documentation for the resulting research dataset.

JEL-Codes: C81, C88
Keywords: record linkage, company data, orbis, survey data

Valentin Reich
ifo Institute for Economic Research
at the University of Munich
Poschingerstr. 5
81679 Munich, Germany
Phone: +49(0)89/9224-1311
reichv@ifo.de

May 06, 2024
I thank Jean-Victor Alipour, Thomas Fackler, Oliver Falck, Moritz Goldbeck, Anna Kerkhof, Valentin Lindlacher, Heike Mittelmeier, and Sebastian Wichert. I further thank the participants of the *ZEW-FDZ Data User Workshop 2019* and of the *Quality Aspects of Machine Learning Workshop 2022*. Lena Abou El-Komboz provided excellent research assistance. All errors are my own.

1 Introduction

In the age of *Big Data*, the speed of information generation increases more and more. Thus, the possibilities for academics to use these data for economic research increase as well. Additionally, data get particularly valuable when different datasets with different kinds of info can be combined: Administrative data, survey data, proprietary data, and more can be linked to each other on a micro level to help to answer questions more timely, to correct for data errors, or to enable the study of completely novel questions. Thus, linking data is often an important task for economic research and policy advice. For example, Meyer and Mittag (2019) link survey and administrative data to overcome measurement error in household income, allowing for an improved evaluation of anti-poverty programs.

Linking entities from different data sources, called *Record Linkage* (RL) or *entity resolution*, is straightforward when the data sources have a common unique identifier. Without a common identifier, records can still be linked via probabilistic matching: The more similar records are in attributes such as name or address, the higher the probability that they refer to the same entity (Fellegi and Sunter, 1969; Newcombe, 1988). Linkage errors, especially wrong matches, can introduce systematic measurement error, and thus bias, to downstream regressions that use linked data (Bailey et al., 2020). While the linkage of natural persons is already a nontrivial task[1], with records of non-natural persons there are additional complications: First, firms often belong to a corporate group of firms, potentially with near identical name and addresses. Second, there are many changes that can occur over time such as reorganizations, name changes, or mergers.

The goal of this paper is to link the *ifo EBDC Business Panels* from the *LMU-ifo Economics and Business Data Center* (EBDC)[2] using Machine Learning (ML) and Natural Language Processing (NLP) methods. Here, I link the responding German firms from the long running surveys of the ifo Institue to their financial information from the commercial *Orbis* database[3] via probabilistic matching. While a linkage not based on ML has previously been conducted several years prior (Hönig, 2010), I now apply these newer techniques to achieve an improved match rate and re-evaluate older matches. This is motivated by the availability of more balance sheet records and because these methods can help overcome some of the challenges of company linkage.

[1] Problems that can arise here are for example typographical errors, different spellings, nicknames, or name changes such as after a marriage (see e.g., Christen, 2012, p. 42ff).

[2] The EBDC then offers the resulting linked dataset for research at their premises. The EBDC is a Munich based accredited research data center at the ifo Institute and it provides secure access to company micro data for academic research at their workstations. Their well documented data include subjective micro data from the *ifo Business Surveys* alongside a version of this data enriched with companies' objective balance sheet data from *Hoppenstedt* and the *Bureau van Dijk* Databases such as *Orbis*. These linked datasets are called the *EBDC Business Panels*. Details about data access can be found in appendix section C. For more info see their website: `https://www.ifo.de/ebdc`

[3] Orbis is an industry standard which is also used and linked for example by the research data centers of the *German Bundesbank* and the *IAB*, the research institute of the German Federal Employment Agency.

For the linkage, I compute a matrix of various similarity metrics for pairs of records which I then use as an input for a supervised ML classification. I use comparison metrics that work well with the challenges of company data and apply NLP methods which are uniquely applicable when dealing with company records: Because the words or *tokens*[4] in company names have a linguistic meaning, pre-trained embedding vectors (Mikolov et al., 2018) allow to extract this information.

The linkage results in a relatively high rate of matched entities, in particular for companies added more recently to the surveys. There also appears to be heterogeneity across sectors and surveys, with the construction survey having the lowest match rate. At the same time, the investment survey for manufacturing has a high match rate despite the long survey run time. Matches with lower predicted match probability were manually corrected, revealing that false positives were almost exclusively cases where a firm was matched with a related entity like its holding. Linkage was particularly difficult when reorganizations within a corporate group occurred. This highlights that corporate structures and relations are a key challenge for company linkage.

This paper contributes to the literature introduced in section 2 by highlighting and addressing key challenges of company data linkage and giving some best practice advice. Additionally, I expand the growing literature of applications of ML methods in applied linkages. A further contribution is that this paper serves as a documentation for the linkage of the final research datasets available at the EBDC.

The next section shows related literature and linkage applications. Then, section 3 describes the specific challenges one faces when linking company data and section 4 explains to what extent NLP methods can support here. Section 5 describes the data used for the linkage which is detailed in section 6. The results of the linkage are then presented in section 7 and the discussion in section 8 lists avenues for further improvements. Finally, the paper concludes with section 9.

2 Related literature

The term *Record Linkage* is said to be coined by Dunn (1946), and Newcombe et al. (1959) proposed an automatic algorithm for linkage without common identifier based on agreement of other fields. These ideas were formalized by Fellegi and Sunter (1969) in an unsupervised framework that computes field specific match weights given how frequently pairs of records agree in the respective field. To determine matches, it then relies on an arbitrarily chosen cutoff for a similarity function that incorporates these weights. An advantage of this method is that it requires no training data. However, because the Fellegi-Sunter framework relies on rarely satisfied assumptions such as conditional independence of fields, supervised ML methods like support vectors machines, random forests and neural

[4]The tokens of "Petra Mayer Sales GmbH" are "Petra", "Mayer", "Sales", and "GmbH".

networks were instead proposed in other methodological papers (e.g., Tejada et al., 2001; Cohen and Richman, 2002; Bilenko and Mooney, 2003; Wilson, 2011; Schild et al., 2017; Cuffe and Goldschlag, 2018; Abowd et al., 2019)[5]

In recent years, the field of methodological RL research evolved further and Hettiarachchi et al. (2014) proposed a *next generation* of linkage using Neural Networks, genetic algorithms, and clustering methods. Thus, modern Deep Learning (DL) neural network architectures, like sequence models and convolutional neural networks are increasingly used for entity linkage (e.g., Gottapu et al., 2016, Ebraheem et al., 2017, Mudgal et al., 2018). Thanks to ML advancements, these applications can make use of *transfer learning*, where models are pre-trained on large datasets and can then be reused for various tasks with less training data. In particular, they make use of NLP methods in the form of pretrained language models, albeit not for company linkage but for example for products and bibliometric data. Mudgal et al. (2018) find that DL benefits only applications with textual or "dirty" data but not those with structured fields. However, by their definition, company names could be considered a dirty field where one can benefit from parsing its informational content using DL.

The particular challenge of linking business data and the need for further research in this field has already been acknowledged by Winkler (1995). However, the methodological literature on company RL appears to be smaller and focused on describing specific linkage cases[6] such as in Peruzzi et al. (2014), Schäffler (2014), Cuffe and Goldschlag (2018), Mason (2018), Moore et al. (2018), Schild (2016), Schild et al. (2017), Abowd et al. (2019), Gschwind et al. (2019), Eberle and Weinhardt (2020), and Doll et al. (2021). Likewise, the present paper is also focused on linkage methodology and serves as a major update to the linkage described in Gramlich (2008).

The original linkage did not use supervised ML but was instead closer to a variant of the Fellegi-Sunter framework. It relied on a set of very likely matches, the *gold standard*, identified via a simple heuristic, to compute field specific weights. This gold standard consisted of pairs that had identical phone numbers, fax number, or email addresses. Since this information is often not available, for the remaining pairs, string similarity metrics for different fields were computed and aggregated in a linear combination with the field specific weights. A match decision was then made based on an arbitrarily chosen threshold on this linear combination. There are a few notable limitations of the original linkage: First, it pre-selected potential matches by requiring an overlap in location information. This can introduce false negatives if the location is erroneously recorded. Instead, I opt for a combination of different pre-selection strategies that together can overcome some of their individual shortcomings. Second, because whether or not phone and email address are present and overlapping may be nonrandom. Therefore, there can be selection into

[5] A survey of the evolution of RL can be found in Binette and Steorts (2022)

[6] Potentially this is because there is a lack of standardized benchmark data.

the gold standard set such that the computed weights may be less representative for other firms. Using hand labelled training data drawn at random, such as I do, alleviates this problem. Third, the previous linkage relied only on a single string similarity metric, whereas my approach employs different methods such that specific errors from individual metrics have a lower impact.

Applied empirical research shows the value of linked company data in economics: Gumpert et al. (2022) use data from administrative German social security records where employees' respective establishment is linked to firms from the Orbis database. This linkage allows to identify establishments belonging to the same firm to analyze how the managerial organization across establishments is interdependent for multiestablishment firms. Additionally, they can estimate how organization is affected by distance to headquarters due to geographic frictions. Aside from this, several papers previously used the *EBDC Business Panels*, i.e., the datasets that are being overhauled in this paper: For example, Huber (2018) analyzes the effect of bank lending cuts on firms and the local economy exposed to such cuts. Therefore, the author uses ifo survey information on the willingness of banks to grant loans and further matches this to a dataset about relationship banks from the credit rating agency Creditreform.[7] Furthermore, Enders et al. (2022) use the EBDC Business Expectation Panel to estimate the effect of firm expectations on later realized production and prices via survey questions. Here, they need the linked balance sheet data for propensity score matching to compare firms that have different expectations but the same fundamentals.

3 Challenges of company linkage

There are some general concerns that apply to any probabilistic linkage application such as tradeoffs between computational feasibility, accuracy, and coverage.[8] Additionally, for RL supported by supervised ML, it is usually required to manually label training data which is time intensive.

Linking non-natural persons such as companies comes with specific complications, in particular through (i) hierarchies, (ii) a lack of standards, and (iii) history which will be explained in the following:

Hierarchies Companies are hierarchical objects in two ways: First, firms can be part of larger corporate groups with separate entities for different business operations such as producing entities, sales entities, or holding companies.

[7] Firms are linked via the *Crefonummer*, a firm identifier that can be recovered from the balance sheet data source of the EBDC Business Panels.

[8] It is usually not computationally feasible to compare all entities of one dataset with all entities from another. Thus, to reduce the computational burden, practitioners need to make some assumptions about potential matches, thereby risking to make false negatives, i.e., worse coverage.

These entities can have very similar or even identical names and addresses. Additionally, there can be various reorganizations both within and across corporate groups due to acquisitions, mergers, fusions, internal activity shifts, renaming, or relocation.

Second, entities in different databases can be at different levels of aggregation. For example, Schild (2016) links establishment level to firm level data.

Standards The company name is a collection of *tokens*, e.g., {*Petra, Mayer, Sales, GmbH*}, and it should be a unique and common identifier. In reality, however, individual tokens can be excluded, included, or replaced across databases. This is a concern because individual words might be crucial to differentiate entities within a corporate group. Additionally, even though company names shall have discriminatory power, Schild (2016) finds duplicate names for around 10% of companies in the Orbis database. Schäffler (2014) further identified that firms with identical names often belong to the same corporate group.

History If the identifying variables have been collected at different points in time across databases, information can differ even if it contains no errors for example due to relocations or name changes. Furthermore, when dealing with panel data, it is possible that different entities would be a preferred match for different periods. This can occur, when a producing entity must be matched for its historical data but the corporate group has since moved its production to a different entity. Depending on the use case, different entities would then have to be matched in different periods.

4 Natural Language Processing for company record linkage

Natural Language Processing (NLP), i.e., techniques for computers to analyze natural language text, can support company linkage in ways not possible for the linkage of natural persons because company names are made up of actual language words. For humans it is easy to understand, contextualize, and relate these but for machines it needs to be processed by means of modern techniques.

One such method are word embedding vectors where words are represented as fixed vectors in an n-dimensional space capturing semantic relations in language (Mikolov et al., 2013). These vectors are learned from a training data such that words used in similar contexts have similar embedding vectors.[9] A variant of these particularly suited for RL is *FastText* (Bojanowski et al., 2017) because it is trained on pairs of characters rather than full words. This makes it more robust towards prefixes, suffixes, and even typos, all of

[9]For training, an algorithm may try to find a relatively low dimensional vector for a given word such that the vector allows to predict the vectors of surrounding words. Thus, words that are used in similar contexts have similar or the same surrounding words and end up with similar vector representations. See for example Ash and Hansen (2022) for a description of properties and applications in economic research.

which matter in RL. Because company names do not follow proper language rules[10], an algorithm cannot infer meaning of words from context. This challenge can be overcome with *transfer learning*, i.e., by using embedding vectors pre-trained on other data. Other entity linkage research already used FastText transfer learning in Ebraheem et al. (2017), Mudgal et al. (2018), and Kasai et al. (2020), albeit not for company data but e.g., products and bibliometric data. I am using FastText embedding vectors pre-trained specifically for German language on massive text corpora of Wikipedia and Common Crawl by Mikolov et al. (2018)

Word embedding vectors can be used for various subtasks in a record linkage process: First, one can use them to compute similarity measures based on contextual similarity. To measure the similarity, I compute the *cosine similarity*[11] of two embedding vectors $\boldsymbol{v}$ and $\boldsymbol{w}$ according to equation 1, where $\|\boldsymbol{v}\|$ and $\|\boldsymbol{w}\|$ refer to the euclidean norms of vectors $\boldsymbol{v}$ and $\boldsymbol{w}$ respectively.

$$similarity = \frac{\boldsymbol{v} \cdot \boldsymbol{w}}{\|\boldsymbol{v}\| \cdot \|\boldsymbol{w}\|} \tag{1}$$

Table B2 in the appendix shows the three words most similar to a given word one might find in a company name. The method captures the meaning of words for example with sectoral, regional, or personal information well and can even enable to extract and standardize legal forms.

Second, they allow segmentation of company names to improve the quality of a linkage (see e.g., Christen, 2012 p. 55).[12]: (i) It reduces the linkage complexity when only tokens that serve the same purpose need to be compared and (ii) it allows for varying importance of types of words in a supervised classification step. For example, a token describing the industry helps differentiate companies with different roles in a corporate group. Figure 1 shows an example where the individual words in names assigned labels such as *legal form*.

Here, segmentation is achieved via supervised machine learning where a Neural Network sequence model with bidirectional LSTM nodes (Hochreiter and Schmidhuber, 1997) uses the FastText word embedding representation[13] of word as inputs to predict the label for each word.[14] The initial training data is taken from Loster et al. (2018)[15] and I iteratively expand this data by manually verifying or correcting predictions for German company

[10]They are just a concatenation of words without proper context.

[11]This measure depends on the angle between the two vectors in the embedding vector space and is higher for words that are used in similar contexts in the training data. The cosine similarity of embedding vectors is also used as a similarity metric for entity linkage for example in Ebraheem et al. (2017).

[12]Also Loster et al. (2017) and Gschwind et al. (2019) show the usefulness of extracting and using company name segments, in particular *colloquial* names. Here, I extract the *proper name* (e.g., "Siemens") which should be identical to the colloquial name in many cases.

[13]Transformation was supported with the *gensim* software library (Řehůřek and Sojka, 2010)

[14]This was done with *Tensorflow* (Abadi et al., 2015).

[15]They train a linear chain Conditional Random Field algorithm on this data for segmentation and do not make use of embedding vectors.

Figure 1: Example for company name segmentation

Original

Name
Petra Mayer Sales GmbH
ABC Gesellschaft mbH Germany

↓

Segmented

Proper Name	Person first name	Person Last name	Sector	Location	Legal
-	Petra	Mayer	Sales	-	GmbH
ABC	-	-	-	Germany	Gesellschaft, mbh

Note: Segmentation example for two fictitious company names.

names randomly selected from the Orbis database. The confusion matrix in appendix figure A1 shows that the classification works well for most types.[16]

Another useful method proposed for RL in Cohen (2000) is the *term frequency - inverse document frequency* (TF-IDF), an information retrieval technique frequently used in NLP applications. It also transforms texts such as company names into real valued vectors in a high dimensional vector space. Here, a lower weight is given to tokens the more frequently they appear in other names[17] because the similarity of rare tokens is more informative than the similarity of a token shared by many company names (Spärck Jones, 1972).

5 Data

ifo Data The focal data consist of the contact information of participating firms[18] from five surveys of the ifo Institute: the monthly Business Surveys (IBS) for manufacturing (IBS-IND, 2019), retail and wholesale (IBS-TRA, 2019), construction (IBS-CON, 2019), services (IBS-SERV, 2019), as well as the biannual Investment Survey (IVS) for manufacturing firms (IVS-IND, 2019). All surveys regularly inquire business related information which is used for example in Bachmann et al. (2013) who study the effect of uncertainty on firms' economic activity and Link et al. (2023) who compare firms' expectations and information frictions to those from the respondents of a household survey.

The ifo Institute conducts regular surveys since 1949[19] but earliest data are no longer available. Instead, data are accessible for the manufacturing sector from the 1960s onwards

[16]The main errors are classifying some rarer words of the *business details*, *abbreviations*, and *other* as *proper name*.

[17]Additionally, the weight increases the more frequent a word appears within a company name. However, this property is less relevant in RL since words are rarely repeated in a name.

[18]This information is held separately from the survey responses and is otherwise not accessible for researchers using the survey micro data at the EBDC.

[19]The original goal was to provide information more timely than official statistics with the IBS being a monthly and the IVS a biannual survey (Sauer and Wohlrabe, 2020).

in the IVS and the 1980s in the IBS. Other sectors have been gradually added to the IBS, with the service sector being the most recent addition in 2001. Questions of the IBS are more of a qualitative or subjective nature, where for example the business expectation is inquired with a three item likert scale. Table 1 contains the number of entities by

Table 1: Number of entities in ifo survey database

Survey	Sector	Absolute	Share	Since	Frequency
IBS	IBS-CON	4797	0.11	1991	Monthly
	IBS-IND	12633	0.30	1980	Monthly
	IBS-SERV	7842	0.19	2004	Monthly
	IBS-TRA	9041	0.22	1990	Monthly
IVS	IVS-IND	7419	0.18	1964	Biannually
Total		41732	1.00		

Note: *IBS* is the ifo Business Survey and *IVS* is the ifo Investment Survey. *CON*, *IND*, *SERV*, and *TRA* respectively denote the surveys for the construction, manufacturing, services and retail/wholsale sectors. *Frequency* refers to the survey frequency, i.e., how often info is inquired from the participants.

survey and shows that manufacturing firms represent almost half of the more than 40.000 entities.[20]

There are several things to note: First, firms are not unique in the database as they can submit multiple questionnaires in a given month.[21] For example, they can be engaged in multiple sectors or participate in both the IBS and the IVS. Also, different surveys have different levels of observations, either firms or products, and this is not consistent over time.[22] Hence, they may respond for multiple products. It is furthermore possible that individual establishments of a company participate (Sauer and Wohlrabe, 2020). Thus, an *m-to-1* matching can occur, where different entities of the ifo database need to be matched with the same entity in the secondary database.

Second, especially before online submission became the dominant form of participation[23], the questionnaire was not necessarily always sent to the entity of interest. Instead, it is possible that for example the holding company collected questionnaires and forwarded them internally. In that case, the database contains information about the recipient rather than the enquired entity.

And third, the database may not always have been updated for example when a firm already ended participation or participates online, where no address is needed. Additionally, because the panel data shall be linked over time, it is possible that the entity of interest in the secondary database changes over time. This adds to the conceptual challenges mentioned in section 3.

[20] More detailed information about the surveys can be found in Sauer and Wohlrabe (2020).
[21] Nonetheless, these duplicate firms each have unique ID numbers.
[22] see Link (2018) for more details.
[23] By now, around 60% participate via the web (Sauer and Wohlrabe, 2020).

BvD Data The secondary data source, i.e., the data I add to the primary source, is the commercial *Orbis* database from the publisher *Bureau van Dijk* (BvD).[24] It contains objective quantitative financial and other information such as balance sheets, patents, or shareholder structures. BvD receive their data for this global database from various sources such as for example *Creditreform Rating AG* for Germany. The entities are on a company level with some of them being so called *branches* of other companies.[25].

The previous linkage of ifo data described in Gramlich (2008) relied on entities from the *Amadeus* database, another BvD product with a focus on European firms. It is supposed to be a subset of Orbis with the same ID numbers.

I use the *2018-06* snapshot from the Orbis flatfiles and export all German companies. I further remove all companies with an ID not starting with *DE* (0.2%) and those whose ID starts with *DE** (1.3%)[26]. This leads to a final selection of 3,759,447 unique BvD IDs.[27]

Available information I use company names, address, sector identifiers, and other contact information such as telephone numbers. Table 2 shows the share of missing information from both data sources after the preprocessing steps described in section 6.1. By design, the company name is always available, since only entities that provide one are considered. Address information is also available in at least 93% of cases. Additional contact information such as phone and fax numbers are frequently missing and the email address is missing in the majority of cases. Thus, these variables are difficult to use for some applications like indexing.

6 Record Linkage procedure

The linkage follows a typical five step workflow as described for example in Christen (2012), a guidebook for practitioners of RL: (i) *Preprocessing* to make the records as comparable as possible. (ii) *Indexing* to narrow down the search space and determine a set of pairs to consider. (iii) *Comparison* to compute a vector of similarity metrics for each pair. (iv) *Classification* of pairs as matches or non-matches using their similarity vector as inputs. And (v) *postprocessing* which includes filtering out ambiguous matches. The rest of this section describes these steps in detail.

[24]The company databases from *Bureau van Dijk*, have already been used in RL applications for example in Peruzzi et al. (2014), Schild (2016), and Schild et al. (2017).

[25]Table A1 shows that around 3.6 percent of IDs in the data are branch IDs.

[26]These do not contain any relevant financial info and are an artifact of entities that could not be matched to other existing IDs by BvD.

[27]This includes both companies registered in the commercial register and unregistered traders. It also includes the *branches*. Additionally, these companies can be active or inactive. Because historical data shall be linked, inactive companies are relevant as well.

Table 2: Share of missing data

	Orbis	ifo
Name	0.00	0.00
Federal state	0.06	0.00
City	0.04	0.05
Address	0.07	0.05
Street	0.07	0.05
Postcode	0.05	0.05
Address number	0.08	0.08
Sector 1 digit	0.16	0.01
Sector	0.16	0.30
Phone	0.55	0.12
Fax	0.68	0.21
Email	0.70	0.52

Note: Share of missing observations after the preprocessing procedure which includes removal of fields that appear erroneous and filling of some missing fields. Thus, for the ifo data, the one digit sector is more frequently available than the full number because it can frequently be inferred when the firm participates in the trade or construction surveys.

6.1 Preprocessing

Even though the linkage is designed to overcome errors and differences in the datasets, it is important to facilitate this by preparing and cleaning the records of both data sources. The main tasks here are (i) standardization, (ii) filling missing information, (iii) feature generation, and (iv) transformation.[28]

Standardization serves to make the records comparable across databases. This includes case folding and replacing German *Umlaute* ä, ö, ü with a, o, and u respectively. Additionally, legal forms are extracted from company names via regular expressions which are adapted from Schild et al. (2017). In Orbis, there can be multiple variants for the name, city, address, phone number, and fax number. I store these alternatives into sets[29] that allow for comparison via set methods as described in section 6.3.

Filling missing data with the help of other fields is possible only in few specific situations. A table containing all German zip codes, their respective municipality, and other regional information (Deutsche Post Direkt, 2019) allows to infer the zip code from the location or vice versa if uniquely possible. The different ifo surveys have different sector identifiers[30] such that they need to be harmonized to the WZ08 industry classification which is roughly equivalent to the NACE Rev. 2 available in Orbis. This is achieved using WZ03 to WZ08 correspondence tables (Destatis, 2008) and in some cases, a one-digit identifier can be inferred from the survey sector itself.[31]

Feature generation infers attributes for a machine learning model from available data:

[28]The steps were in part inspired by Schild (2016). The full list of measures can be found in appendix section B.1.

[29]E.g., phone numbers: {+123 456789, +987 654321}.

[30]See Link (2018) for a very detailed overview.

[31]E.g., all entities in the construction survey should have a "4" as first digit.

The zip code identifies the federal state and the four digit sector identifier can be aggregated into more coarse categories.[32]

Transformation creates new fields as transformations of existing ones. Here, I use Phonetic encoding to counter different spellings. I encoded attributes with the *Double Metaphone*[33] encoding (Philips, 2000) which is designed to work with a number of different languages, including German. This transforms for example the word "Maschinenbau" (engineering) to "MXNNP"[34]. Because, by removing important differences, phonetic encoding can worsen match rates when used for comparison (Bailey et al., 2020), I only use it for selecting candidate pairs in the indexing step. Additionally, I use the FastText NLP method introduced in section 4 by transforming the *city* field into its embedding vector representation and by segmenting the company names. Due to data protection concerns, ifo data need to be processed on a specially protected computer, whereas the Orbis data could be preprocessed on a different machine. Because of hardware limitations of the protected device, the segmentation could only be executed for the Orbis data. Thus, rather than comparing the same tags[35] of both datasets, I check whether there is an overlap between each tag of the Orbis segments with all of the tokens of the ifo company. Under the assumption that a segmentation of the ifo data would have resulted in the same labels for the same words, this second best approach should not differ much from the optimum. This is plausible because the segmentation relies mostly on the fixed word embedding vectors of company name tokens such that the same words are likely predicted equally in both data sources.

The result of the preprocessing step are two tables, one for each data source, with the cleaned contact information of firms.

6.2 Indexing

The set of all possible pairs is the cartesian product of both data sources, i.e., of size $n \times m$ with n and m respectively being the number of records in both sources. The computational cost of this set can be prohibitively large when one dataset has millions of records as it is the case for Orbis. Thus, the indexing step serves to select a set of potential pairs to consider for further linkage steps. Figure 2 shows an example where the number of pairs is reduced from $3 \times 3 = 9$ to 3.

Because the vast majority of potential pairs is very dissimilar, it makes sense to filter

[32] It is possible to aggregate to 3-digit, 2-digit, 1-digit, to only a differentiation between *manufacturing*, *trade*, *construction* and *service* sector. This is helpful because the more coarse the information, the more likely it is that a true match agrees.

[33] I used a python implementation from
https://github.com/dracos/double-metaphone/blob/master/metaphone.py

[34] Some words can be encoded into a primary and a secondary encoding. In this case, only the primary encoding is utilized.

[35] The *tag* refers to the label of tokens. I.e., when comparing sector tags, this refers to an list of all words with sector information within a company name.

Figure 2: Indexing example

Database A

ID_A	Name	Street
A1	Petra Mayer Sales GmbH	Abc-Str.
A2	ABC Gesellschaft mbH Germany	-
A3	XYZ AG	Ghi-Str.

Database B

ID_B	Name	Street
B1	ABC GmbH	Def-Str.
B2	Petra Mayer GmbH	Abc-Str.
B3	Maier GmbH	Xyz-Str.

↓

Indexed pairs

ID_A	ID_B	$Name_A$	$Name_B$	$Street_A$	$Street_B$
A1	B2	Petra Mayer Sales GmbH	Petra Mayer GmbH	Abc-Str.	Abc-Str.
A1	B3	Petra Mayer Sales GmbH	Maier GmbH	Abc-Str.	Xyz-Str.
A2	B1	ABC Gesellschaft mbH Germany	ABC GmbH	-	Def-Str.

Note: Fictitious example of the indexing step. The tables of database A and B each contain records' cleaned attributes after the preprocessing step. Only the indexed pairs are considered for further linkage steps. The table of indexed pairs contains the attributes of the records from both data sources. The index itself refers to the columns ID_A and ID_B from the table of indexed pairs.

them out using a fast selection method in the form of *blocking* (Newcombe et al., 1959; Newcombe and Kennedy, 1962) and *filtering* first.[36] Blocking requires pairs to perfectly agree on a set of predetermined fields, the *blocking keys*.[37] For example, records can be required to have the same sector code. Filtering is a step applied after the blocking and it requires records to have some minimum similarity score in a given field. The similarity measure is ideally fast to compute such that it can be done for a larger set of pairs.

Both methods come with a trade-off: while stricter rules make the search computationally feasible they can lead to false negatives, for example when there are errors in the blocking key. Thus, it is suggested to use the union of pairs from multiple different blocking and filtering strategies as a final index (Herzog et al., 2007).[38] Here, the basis of most strategies are combinations of sector or location based blocking keys as these are frequently filled. Additionally, the respective Orbis-*branches* of candidates and previously collected ML training data pairs were included in the index.

We have already seen in section 3 that the temporal dimension can introduce challenges for panel data sources. Potentially, an ifo ID needs to be matched to one Orbis ID for older historical information and to a different one for more recent observations for example due to a restructuring of the company. Here, I propose to do two separate linkages, a *pre* and a *post* linkage: The *pre* linkage considers only pairs where the date of incorporation from

[36] The usage of both methods together is for example suggested by Papadakis et al. (2019).

[37] Additionally, I use *Sorted Neighborhood Blocking* (Hernández and Stolfo, 1995) which makes the indexing more robust to noisy data (Papadakis et al., 2019). Here, records are sorted on a predetermined key and rather than requiring a perfect overlap, a fixed size window is moved over the records such that all records that lie within this window are considered as pairs.

[38] Both the indexing and comparison step were mostly executed with the *Record Linkage Toolkit* (De Bruin, 2019) with additional metrics from the *jellyfish* and *textdistance* packages in python.

Orbis was before the ifo survey start, i.e., the company must have existed when its ifo counterpart participated in the survey. Conversely, the *post* linkage considers only pairs where the Orbis date of incorporation was *after* the survey start.[39] The following steps, i.e., comparison, classification, and postprocessing, are then all conducted separately for both the *pre* and *post* pairs.

Table B1 in the appendix shows the blocking and filtering keys by strategy. The union of all 14 strategies leads to a final index of around 4.4 million unique pairs. This is substantially larger than the index any single one of these strategies would achieve but a small fraction of the more than 120 billion pairs of the full index.

The *pre* and *post* indexing steps each result in a correspondence table with the IDs of considered pairs.

6.3 Comparison

The basis for classifying the candidate pairs from the indexing step as matches or non-matches is the matrix of their similarity scores. Cuffe and Goldschlag (2018) suggest that linkages can be more effective by combining many different comparison metrics. The full list of comparison metrics is shown in appendix table B3. Figure 3 exemplifies this step based on the example index from figure 2.

Figure 3: Comparison example

Indexed pairs

ID_A	ID_B	$Name_A$	$Name_B$	$Street_A$	$Street_B$
A1	B2	Petra Mayer Sales GmbH	Petra Mayer GmbH	Abc-Str.	Abc-Str.
A1	B3	Petra Mayer Sales GmbH	Maier GmbH	Abc-Str.	Xyz-Str.
A2	B1	ABC Gesellschaft mbH Germany	ABC GmbH	-	Def-Str.

↓

Similarity matrix

ID_A	ID_B	$ngram_{Name}$	LCS_{Name}	$Jaro_{Street}$
A1	B2	0.739	0.842	1.000
A1	B3	0.391	0.562	0.750
A2	B1	0.276	0.444	0.000

Note: Fictitious example of the comparison step. Here the attributes such as name or street of pairs which given by the index from the indexing step are compared with string similiarity metrics. Thus, $ngram_{Name}$ refers to the ngram similarity between $Name_A$ and $Name_B$. Similarities are computed on the raw strings for this example. Results in the actual linkage are likely more favourable thanks to the cleaning from the preprocessing step which is omitted here for simplicity.

For this study, I choose methods that I expect to work well with the specific challenges of company data: (i) Order robust string comparison methods, (ii) array methods, (iii)

[39] Thus, the comparison of the date of incorporation on the one side to the year of survey start on the other can be seen as a *complex feature* according to Wilson (2011).

TF-IDF based methods, and (iv) embedding methods. Order robust string comparison metrics are useful for company names because they compute the similarity between two strings such that the order of tokens has less impact. Here, I use Longest Common Subsequence (LCSSeq) (Hirschberg, 1977), Longest Common Substring (LCS) (see e.g., Gusfield 1997), Character n-gram similarity[40] (Ukkonen, 1992), Cosine similarity of character n-grams, and Smith-Waterman (Smith and Waterman, 1981). Array or set methods come in different forms: First, it is possible to check if two records have any overlapping tokens or compute the share of overlapping tokens.[41] Thus, one can check for overlapping tokens between {*"Petra", "Mayer", "Sales", "GmbH"*} and {*"Petra", "Mayer", "GmbH"*}. Second, one can compute string similarities for all possible combinations of token pairs across two sets to get the maximum similarity or compute a fuzzy overlap where it is sufficient for tokens to have a minimum string similarity for a binary overlap indicator[42]. For sectors, I use array methods with information on all available sector identifiers provided by Orbis. Array methods are also applied to the company name segments where, for each of a subset of segment categories[43], the overlap of Orbis tokens from this category with all ifo tokens is computed. To weight tokens based on their relative frequency, I apply both cosine similarity on TF-IDF vectorized record fields and Soft TF-IDF. The latter is a measure that often performs very well in RL applications (Cohen et al., 2003) by combining string similarity with frequency weights to also consider similar tokens. Embedding methods use the cosine similarity between embedding vectors of tokens as described in section 4. Here, I use them for location information[44] to capture for example similarities in locations of different geographic hierarchy. Figure B1 shows that the cosine similarity of location info word embeddings is highly correlated to string based similarity metrics.

The comparison step results in a matrix[45] where each row is the vector of similarity metric scores for a given considered pair of records.

6.4 Classification

To differentiate between matches and non-matches, I use the comparison matrix as input to a supervised ML classification[46] with manually labelled record pairs as training data.

As suggested for example in Bailey et al. (2020), the algorithm for classification is an

[40] This method is commonly used for company names, for example in Gramlich (2008) and Schild (2016).

[41] Further set methods I utilized were the cosine similarity between words, the Jaccard index (Jaccard, 1912), and Monge-Elkan (Monge and Elkan, 1996).

[42] The latter is used only for filtering in the indexing step. With this filter, pairs are required to have a token-wise Jaro similarity (Jaro, 1989) of 0.8 or 0.9, depending on the attribute.

[43] Here, I restrict the analysis to the segment categories *location*, *person first name*, *person last name*, *sector*, and *proper name* because these were well classified and I expect them to be the most useful in separating companies.

[44] In a follow up study, this method could also be well applied to the company name, in particular to the name segment that contains sector information.

[45] One for both the *pre* and *post* linkages respectively.

[46] Training and prediction were done using the *scikit-learn* library (Pedregosa et al., 2011).

ensemble of several different estimators, each with different transformations and comparison metrics. Appendix table B5 lists the 24 individual models that make up the ensemble. A stratified 10-fold cross validation helps tuning the hyperparameters of the individual models and their preprocessing pipelines. The final score is aggregated using a logistic regression that takes predicted probabilities of the ensemble models as input. Rather than using all available comparison metrics as inputs, most models use only a subset of features[47] or reduce dimensionality via *principal component analysis* (Pearson, 1901; Hotelling, 1933).

Because the vast majority of pairs are non-matches and because of the bimodal similarity distribution, selecting pairs to label at random can result in a set of many completely dissimilar pairs and a few almost identical ones. Thus, the pairs in the training data likely consist only of obvious extremes and the algorithm cannot learn patterns of the more complex cases. Because labelling is very time consuming, it is not feasible to draw and label a random sample with sufficient support for all the different cases. Thus I opt for an iterative *active learning* approach, where I draw further labelling data given their predicted match probabilities from a previous iteration. An active learning approach is also suggested and used for RL in Tejada et al. (2001), Sarawagi and Bhamidipaty (2002), Isele and Bizer (2013), Qian et al. (2017), and Kasai et al. (2020). With this approach, I can ensure that the number of matches and non-matches is more balanced and at the same time, I oversample difficult cases by drawing relatively more pairs with a predicted match probability of around 30-70%. Appendix table B4 shows how many instances are in the training data. *Training data 1* (8,307 instances) is used to train the individual models of the ensemble and *training data 2* (3,561 instances) is used to train the ensemble aggregator.

A drawback of this active learning approach with oversampling of difficult cases is that it is not straightforward to evaluate the algorithm with an unbiased performance metric. Nonetheless, for transparency, I include table B5 with the classification metrics for each model of the ensemble in the appendix. This table also highlights that the ensemble outperforms any of its individual components with both a comparably high recall and precision.

The classification results in a vector[48] containing the predicted match probabilities for each of the considered pairs.

6.5 Postprocessing

A postprocessing step ensures there is only one Orbis ID per ifo ID. Appendix table B6 shows that for around one quarter of the ifo IDs with any match, there is more than one match after the classification step. Thus, for each ifo ID, I keep only the match with the

[47] Features are selected via an aggregation function, via penalized regression, or via some heuristic.
[48] Actually two vectors, one for the *pre* and *post* linkages respectively.

highest predicted match probability.

Ultimately, a manual review of the remaining matches allows to correct for mistakes. To avoid systematic errors in downstream analyses, it is important to avoid false positives more so than false negatives (Bailey et al., 2020). For this reason and because the review is very time consuming, it is mostly limited to correcting for false positives with a predicted match probability in the range between 50% and 90%[49]. Additionally, some pairs in the range from 40% to 50% were manually corrected to evaluate the extent of false negatives. Figure B2 shows the share of corrected entities by predicted probability in the *pre*-linkage. The error rate at match probabilities of 90% was very low and entities from the manufacturing surveys needed to be corrected the most. The errors made by the classifier are almost exclusively cases where the wrong company within a corporate group has been selected or where it was not possible to manually label a match with certainty[50].

The postprocessing step results in one correspondence table with the matched ifo and BvD ID numbers for both the *pre* and *post* linkages respectively.

7 Results

The linkage results in two correspondence tables: A larger *pre* table, containing only Orbis companies founded *before* the firm started to participate in the survey and a smaller *post* table, containing only Orbis companies founded *after* survey participation.

Match rate Table 3 shows how many ifo IDs are be matched in each survey. The majority of companies has a match and these are primarily coming from the *pre* linkage, as expected. The number of matches from the *post* linkage is smallest for the service sector survey, with only 20 identified matches, and largest for the two manufacturing industries surveys. A possible explanation for this is that the IBS-SERV started in 2004, while data for the IVS-IND and IBS-IND are available since 1964 and 1980 respectively. In such a long time span, reorganizations are more likely. Also other factors influence the match rate: Despite the long time the survey has been running, IVS-IND has the second best match rate. This may be explained for example by the nature of the manufacturing companies in this survey. Here, larger companies are more strongly represented than smaller ones (Sauer and Wohlrabe, 2020) and thus there can be a lower risk of these entities exiting the market (Aldrich and Auster, 1986). Appendix table A2 confirms this by showing that the manufacturing companies in the surveys tend to be larger. On average, the IVS-IND firms have almost seven times as many employees as the IBS-CON firms, potentially explaining why the construction companies have the worst match rate with 65% of IDs matched.

[49]The distribution of probabilities is bimodal with its peaks on the two extremes, i.e., very low and very high probabilities. The middle on the other hand contains comparably few observations such that manual review is feasible. Going beyond 90% is impractical due to the high volume and quality of match pairs.

[50]This occurs when it seems ambiguous which match candidate is the correct one.

Sector differences can also be driven by organizational differences or naming conventions. For example, figure A2 in the appendix displays name changes recorded in the Orbis database by sector and shows that construction companies are subject to substantially more name changes than manufacturing or trade companies.

Table 3: Match rates by survey

Survey	All ifo ids	Matches pre	Matches post	Share of ifo ids with any match
IBS-CON	4797	3074	104	0.65
IBS-IND	12633	8486	395	0.69
IBS-SERV	7842	6187	20	0.79
IBS-TRA	9041	6154	147	0.69
IVS-IND	7419	4813	796	0.72

Note: Table shows the matchrate, i.e., the share of ifo IDs that could be matched to an Orbis entity. *Matches pre* is the number of ifo IDs matched in the *pre* linkage. *Matches post* is the number of ifo IDs matched in the *post* linkage. *Share of ifo ids with any match* is the share of unique ifo firms matched in the pre, post, or both linkages.

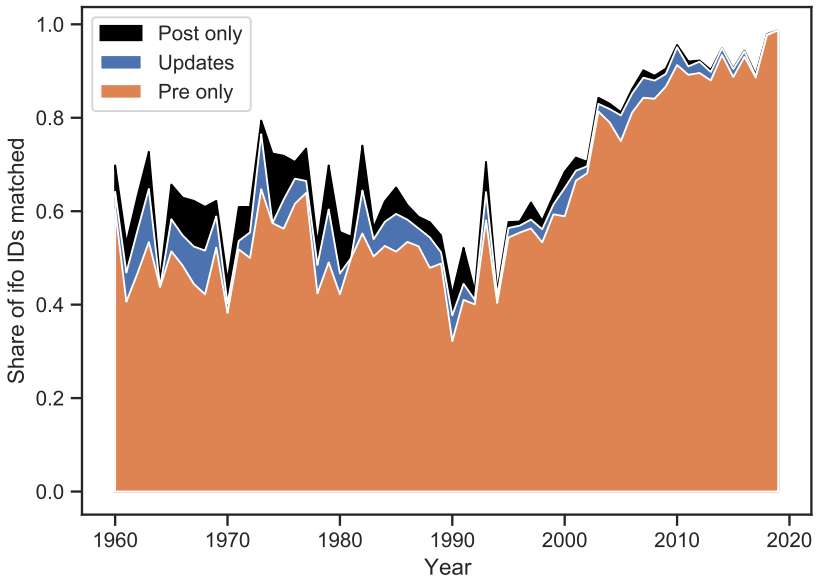
Figure 4 supports the hypothesis that it is primarily older entries that receive updates in the post linkage. It also shows that the linkage rate improves over time and nearly all firms added to the survey in the most recent years can be linked. One can also see that a substantial fraction of IDs from the early years has only a match in the post table. This occurs for example when the original firm ceases to exist after a reorganization and is not listed in Orbis.[51]

A multivariate regression allows to analyze this more systematically by showing how the match rate varies with different attributes holding all others fixed. Figure 5 shows the coefficients from a regression of the match status on various observed firm characteristics: Despite their high overall linkage rate, the linkage appears to be most difficult for service companies when conditioning on other factors. Furthermore, companies from eastern Germany have a lower matchrate than those from western Germany. A very strong predictor of matchrate is when companies still participate or ended participation just recently. There are also some substantial differences between the pre and post linkages: Post linkage rates are higher for eastern Germany and they decrease for more recently added entities. The latter is intuitive since these likely did not experience an organizational shift in the shorter period. Additionally, the coefficients on employment size range dummies are reported in appendix figure B3. They show that the linkage appears to be easier for medium sized companies, while very large companies are harder to match. A potential reason for this is that larger companies can be organized in more complex corporate groups.

Metric importance With the variety of different comparison metrics used, it can be helpful to see which of these are particularly useful in finding matches. This allows to

[51] It is not clear when and under which conditions such inactive firms are not listed in Orbis.

Figure 4: Matchrate by year of survey start



Note: The figure shows the matchrate, i.e., the share of ifo IDs that could be matched to an Orbis entity. The x-axis represents the year an entity was added to the survey and does not need to coincide with its year of incorporation. *Pre only* refers to ifo IDs which could only be linked in the *pre* linkage, i.e., to a company that existed before the survey start. *Post only* refers to ifo IDs which could only be linked in the *post* linkage, i.e., to a company founded after the survey start. *Updates* refers to ifo IDs which could be linked to different entities in both the pre and post linkages.

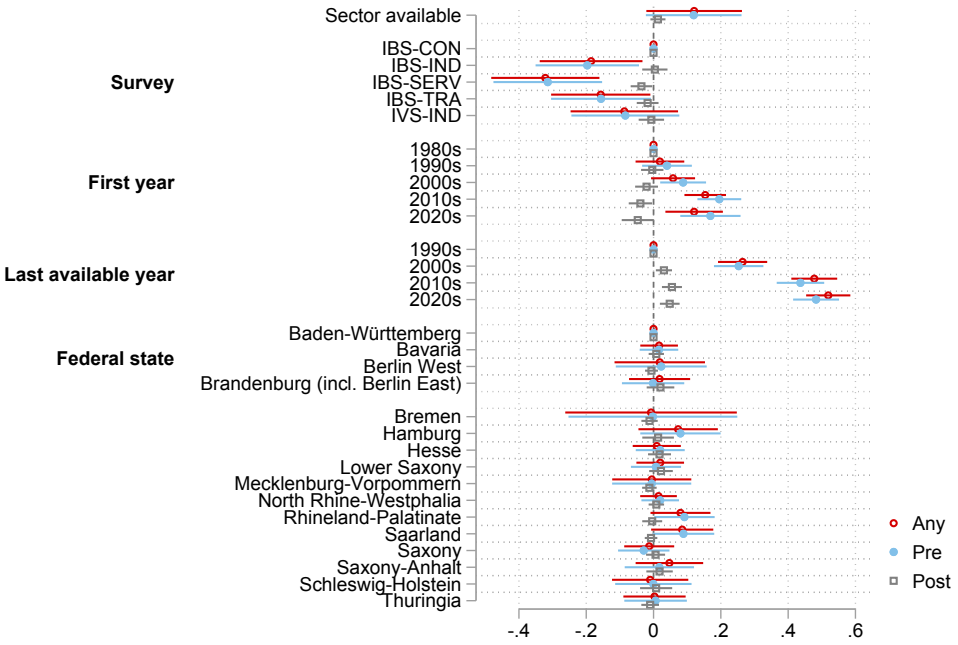
narrow down the metrics and thus decrease computational cost in future applications. Because there are different algorithms in the ensemble and they use different sets of features, it is necessary to evaluate feature importances with a model agnostic framework.

One such method is the recent *SHAP* algorithm by Lundberg and Lee (2017) which they introduced for more interpretability of modern black box prediction models. It does so by computing values informative about the importance of each feature for a given prediction or set of predictions. The method is based on the Game Theoretic concept of Shapley values (Shapley, 1953) which measure the individual marginal contribution to reach a common outcome.

Figure 6 shows the most important features as given by the SHAP[52] method. As is to be expected, name and address are the most relevant pieces of information. Furthermore, the Tfidf and SoftTfidf measures appear to be relatively important, whereas the name segments have a relatively smaller impact. While simple string similarity measures of name, street, and city contribute already much to the predicted probability, one can see that it is also important to incorporate different name variants such as previous names.

[52]I used the official python implementation by the authors via the *shap* package.

Figure 5: Regression of match status on firm characteristics



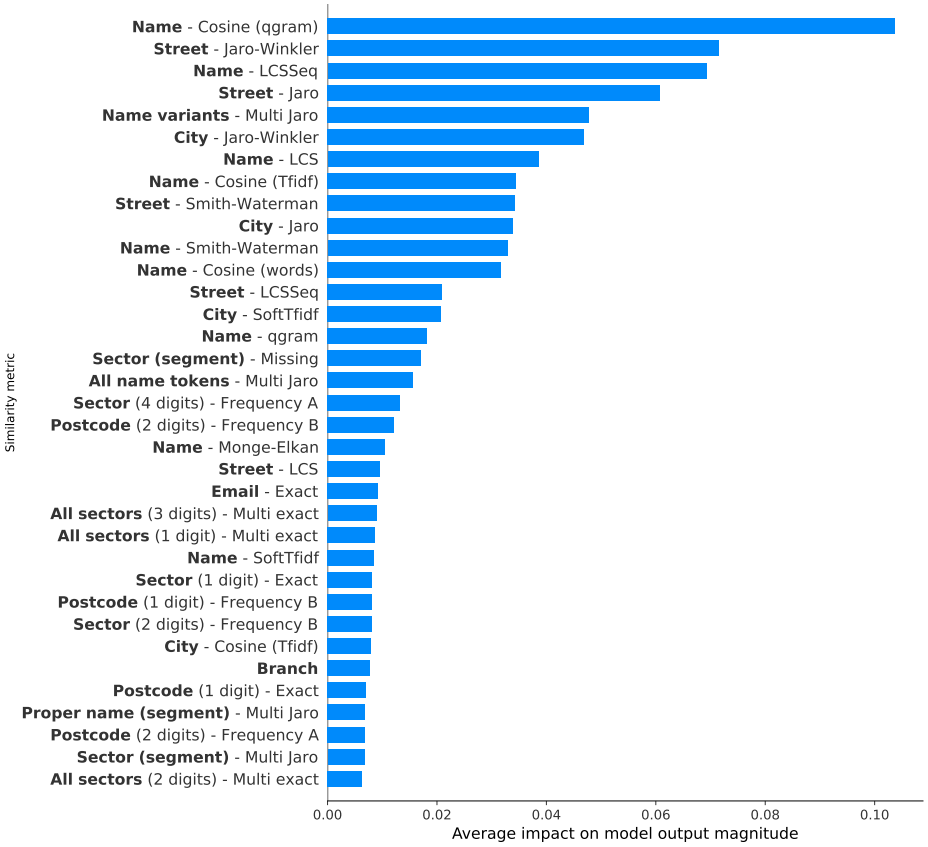
Note: Coefficients of a regression of match status on characteristics from the ifo companies. Employment size range dummies are also included in the regression but their coefficients are only shown in appendix figure B3 for better visibility. Point estimates of a linear probability model are shown with 95% confidence intervals based on robust standard errors. *First year* refers to the first year with available ifo survey responses of the firm and since survey data for the earliest decades is not available any more, it is *1980s* even when the firm started participation before that. *Last available year* refers to the decade of the last response in the survey and it is *2020s* for entities that still participate.

Selection Given the correlation of observed firm properties with the match rate, one may be concerned about how representative the matched sample is or about effects on downstream estimates[53]. Because the Business Panels are used for different types of research questions, it is not straightforward to test for selection bias in a general sense. Instead, in figure 7 I compare the time series of two of the most important questions[54] in the IBS, the assessment of the *business situation* (panel a) and the *business expectation* (panel b). The time series of both the pre and post linkages are close to that of all observations, i.e., linked and not linked, but there is nonetheless a difference which appears to decrease for the more recent periods.

[53]This is a general concern of RL applications and this topic has been the focus of several research papers such as Abowd and Vilhuber (2005), Moore et al. (2018), and Bailey et al. (2020).

[54]These two variables are for example used to compute the *ifo Business Cycle Index*.

Figure 6: SHAP feature importance



Note: Bars indicate the average absolute shapley values, i.e. the average impact of a similarity metric on the predicted probability. Only the 35 most relevant features out of 131 are presented here.

8 Discussion

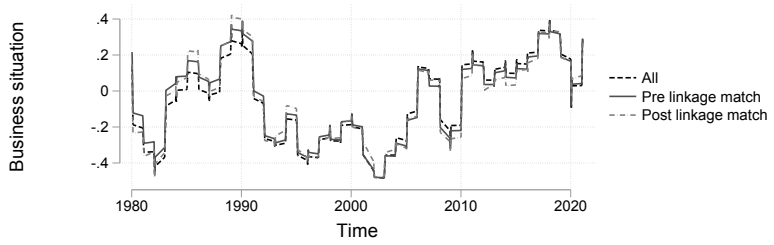
While the linkage is overall successful, there is still room for improvement and possibilities for future linkages.

One challenge comes with the *pre* and *post* linkages: While this paper tries to account for changes in relationships, it is difficult to encode this information into a linked research data set because the timing of the change is unknown. For this reason, the EBDC decided to only use the most recent match for each ID.

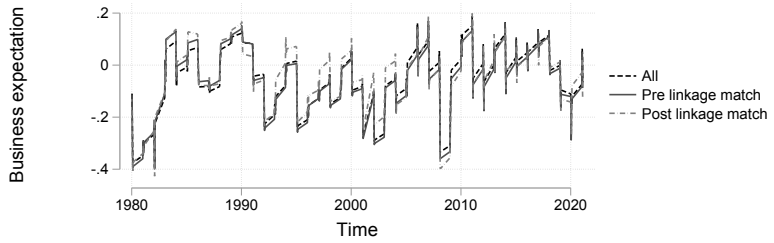
To make the labeling of training data more efficient, I opted for an active learning approach where difficult cases were oversampled. This makes evaluation metrics challenging to interpret and they can also not be compared to other linkages.

Another potential concern is that I used the same trained classifier ensemble for both the *pre* and *post* linkages even though there might be systematic differences. In a subsequent linkage, the pairs from the manual *post* corrections could be used as additional training data either in tandem with a dummy indicating the pre/post status or for a

Figure 7: Time series of business expectation and business situation by match status



(a) Business situation



(b) Business expectation

Note: The time series represent five months moving averages of the business situation and business expectation questions. The three level likert scale question has been recoded to 1 for *good*, 0 for *neutral*, and -1 for *bad*. The dark dashed line represents the time series for all firms, i.e., includes non-matched entities.

separate model.

A further challenge in RL is that it is not possible to compute similarity metrics for fields with missing information. Here, I assigned a field specific similarity of zero for pairs where a field is empty in either data source and additionally included a binary indicator for this value being missing. However, some of the classifiers that make up the ensemble, in particular the tree based models, can make use of this indicator better than others. Alternatively, one could use methods proposed in Ong et al. (2014) to handle these cases.

Despite extensive manual control, it is possible that there are still errors in the linkage given how complicated the task can be. Another challenge lies in choosing the correct BvD ID for each ifo ID if there are multiple predicted matches. Right now, I select the entity with the highest probability, irrespective of other matches. An alternative would be to only match this when it is sufficiently apart from a second potential match and to otherwise not match this at all.

9 Conclusion

Linked data offer great opportunities to work on novel research questions. However, linkage can be challenging, in particular when working with data from non-natural persons. The LMU-ifo EBDC offers researchers access to linked datasets which combine survey

responses with financial information and wants to improve this offering by expanding the data as well as possible. Therefore, this paper combines the respondents of the ifo surveys to their respective records from the commercial Orbis database which contains financial information. Because there is no common identifier, I apply a probabilistic Record Linkage procedure supported by supervised ML for match classification. The process is tailored to the specific challenges of company data linkage via the use of appropriate similarity metrics and the exploration of NLP techniques.

The linkage works particularly well for more recent entries into the database where the entities have a very high match rate. Practically all false positives the classifier produced were cases where an entity was matched to a different but related company. This shows that the key difficulty in company RL is differentiating companies within a corporate group.

Because some aspects of this linkage are specific to the present databases, is is not clear to what extent the method or trained models can be reused in other future linkage applications with different data sources. However, it can be considered as a starting point. Subsequent research should further explore the use of NLP techniques like for example Deep Learning based name similarity metrics which were not possible in the present application due to hardware limitations. Furthermore, because differentiating related companies from each other is the biggest challenge, it could be worthwhile to explore how information about the network for firms can be utilized. Ultimately, the majority of matches are oftentimes cases that are already very similar and a perfect probabilistic linkage will never be achieved such that some of the linkage techniques mostly serve to increase the match rate just a little bit more. Thus, an applied researcher must evaluate how much to invest in improvements into the linkage.

References

Abadi, Martin, A Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng (2015) " TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," URL: `https://www.tensorflow.org/`.

Abowd, John M., Joelle Abramowitz, Margaret C. Levenstein, Kristin McCue, Dhiren Patki, Trivellore Raghunathan, Ann M. Rodgers, Matthew D. Shapiro, and Nada Wasi (2019) *Optimal Probabilistic Record Linkage: Best Practice for Linking Employers in Survey and Administrative Data*: US Census Bureau, Center for Economic Studies, URL: `https://ideas.repec.org/p/cen/wpaper/19-08.html`.

Abowd, John M and Lars Vilhuber (2005) "The Sensitivity of Economic Statistics to Coding Errors in Personal Identifiers," *Journal of Business & Economic Statistics*, **23** (2), 133–152, URL: `https://doi.org/10.1198/073500104000000677`.

Aldrich, Howard and Ellen R Auster (1986) "Even dwarfs started small: Liabilities of age and size and their strategic implications.," *Research in organizational behavior*.

Ash, Elliott and Stephen Hansen (2022) "Text Algorithms in Economics."

Bachmann, Rüdiger, Steffen Elstner, and Eric R Sims (2013) "Uncertainty and Economic Activity: Evidence from Business Survey Data," *American Economic Journal: Macroeconomics*, **5** (2), 217–249.

Bailey, Martha J., Connor Cole, Morgan Henderson, and Catherine Massey (2020) "How Well Do Automated Linking Methods Perform? Lessons from US Historical Data," *Journal of Economic Literature*, **58** (4), 997–1044, URL: `https://pubs.aeaweb.org/doi/10.1257/jel.20191526`.

Bilenko, Mikhail and Raymond J Mooney (2003) "Adaptive Duplicate Detection Using Learnable String Similarity Measures," in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, 39–48, New York, NY, USA: Association for Computing Machinery, URL: `https://doi.org/10.1145/956750.956759`.

Binette, Olivier and Rebecca C Steorts (2022) "(Almost) all of entity resolution," *Science Advances*, **8** (12), eabi8021, URL: https://www.science.org/doi/abs/10.1126/sciadv.abi8021.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov (2017) "Enriching Word Vectors with Subword Information," *Transactions of the Association for Computational Linguistics*, **5**, 135–146.

Christen, Peter (2012) *Data Matching*, Berlin, Heidelberg: Springer Berlin Heidelberg, 1–270, URL: http://link.springer.com/10.1007/978-3-642-31164-2.

Cohen, William W (2000) "Data Integration Using Similarity Joins and a Word-Based Information Representation Language," *ACM Trans. Inf. Syst.*, **18** (3), 288–321, URL: https://doi.org/10.1145/352595.352598.

Cohen, William W, Pradeep Ravikumar, and Stephen E Fienberg (2003) "A Comparison of String Metrics for Matching Names and Records," *Kdd workshop on data cleaning and object consolidation*, **3**, 73–78, URL: www.aaai.org.

Cohen, William W. and Jacob Richman (2002) "Learning to match and cluster large high-dimensional data sets for data integration," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 475–480.

Cuffe, John and Nathan Goldschlag (2018) "Squeezing More Out of Your Data: Business Record Linkage with Python," (18-46), URL: https://ideas.repec.org/p/cen/wpaper/18-46.html.

De Bruin, J (2019) "Python Record Linkage Toolkit: A toolkit for record linkage and duplicate detection in Python," URL: https://doi.org/10.5281/zenodo.3559043.

Destatis (2008) "Klassifikation der Wirtschaftszweige, Ausgabe 2008," *German Federal Statistical Office*.

Deutsche Post Direkt (2019) "DATAFACTORY BASIC 2019 Q4."

Doll, Hendrik, Eniko Gábor-Tóth, and Christopher-Johannes Schild (2021) "Linking Deutsche Bundesbank Company Data," *Technical Report 2021-05 – Version v2021-2-6. Deutsche Bundes-bank, Research Data and Service Centre.*

Dunn, Halbert L (1946) "Record linkage," *American Journal of Public Health and the Nations Health*, **36** (12), 1412–1416.

Eberle, Johanna and Michael Weinhardt (2020) "Record Linkage of the Linked Employer-Employee Survey of the Socio-Economic Panel Study (SOEP-LEE) and the Establishment History Panel (BHP)," *SSRN Electronic Journal*.

Ebraheem, Muhammad, Saravanan Thirumuruganathan, Shafiq Joty, Mourad Ouzzani, and Nan Tang (2017) "DeepER – Deep Entity Resolution," **11** (11), URL: http://arxiv.org/abs/1710.00597%0Ahttp://dx.doi.org/10.14778/3236187.3236198.

Enders, Zeno, Franziska Hünnekes, and Gernot Müller (2022) "Firm Expectations and Economic Activity," *Journal of the European Economic Association*, **20** (6), 2396–2439.

Fellegi, Ivan P and Alan B Sunter (1969) "A theory for record linkage," *Journal of the American Statistical Association*, **64** (328), 1183–1210.

Gottapu, Ram Deepak, Cihan Dagli, and Bharami Ali (2016) "Entity Resolution Using Convolutional Neural Network," *Procedia Computer Science*, **95**, 153–158, URL: http://dx.doi.org/10.1016/j.procs.2016.09.306.

Gramlich, Tobias (2008) "Beschreibung der Verknüpfung der ifo-Konjunkturdaten mit der kommerziellen Firmendatenbank."

Gschwind, Thomas, Christoph Miksovic, Julian Minder, Katsiaryna Mirylenka, and Paolo Scotton (2019) "Fast Record Linkage for Company Entities," *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*, 623–630.

Gumpert, Anna, Henrike Steimer, and Manfred Antoni (2022) "Firm Organization with Multiple Establishments," *The Quarterly Journal of Economics*, **137** (2), 1091–1138.

Gusfield, D (1997) *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*, EBL-Schweitzer: Cambridge University Press, URL: https://books.google.de/books?id=Ofw5w1yuD8kC.

Hernández, Mauricio A and Salvatore J Stolfo (1995) "The Merge/Purge Problem for Large Databases," *SIGMOD Rec.*, **24** (2), 127–138, URL: https://doi.org/10.1145/568271.223807.

Herzog, T. N., F. J. Scheuren, and W. E. Winkler (2007) *Data Quality and Record Linkage Techniques*, **1**, New York: Springer.

Hettiarachchi, Gayan Prasad, Nadeeka Nilmini Hettiarachchi, Dhammika Suresh Hettiarachchi, and Azusa Ebisuya (2014) "Next generation data classification and linkage: Role of probabilistic models and artificial intelligence," *Proceedings of the 4th IEEE Global Humanitarian Technology Conference, GHTC 2014*, 569–576.

Hirschberg, Daniel S (1977) "Algorithms for the longest common subsequence problem," *Journal of the ACM (JACM)*, **24** (4), 664–675.

Hochreiter, Sepp and Jürgen Schmidhuber (1997) "Long Short-Term Memory," *Neural Computation*, **9** (8), 1735–1780, URL: https://doi.org/10.1162/neco.1997.9.8.1735.

Hönig, Anja (2010) "Linkage of Ifo Survey and Balance-Sheet Data: The EBDC Business Expectations Panel & the EBDC Business Investment Panel," *Schmollers Jahrbuch*, **130** (4), 635–642.

Hotelling, H (1933) "Analysis of a complex of statistical variables into principal components.," *Journal of Educational Psychology*, **24**, 417–441.

Huber, Kilian (2018) "Disentangling the Effects of a Banking Crisis: Evidence from German Firms and Counties," *American Economic Review*, **108** (3), 868–898.

IBS-CON (2019) *Ifo Business Survey Construction 1/1991 – 12/2019*, Munich, DOI: 10.7805/ebdc-ibs-con-2019b: LMU-ifo Economics & Business Data Center.

IBS-IND (2019) *Ifo Business Survey Industry 1/1980 – 12/2019*, Munich, DOI: 10.7805/ebdc-ibs-ind-2019b: LMU-ifo Economics & Business Data Center.

IBS-SERV (2019) *Ifo Business Survey Service Sector 10/2004-12/2019*, Munich, DOI: 10.7805/ebdc-ibs-serv-2019b: LMU-ifo Economics & Business Data Center.

IBS-TRA (2019) *Ifo Business Survey Trade 1/1990 – 12/2019*, Munich, DOI: 10.7805/ebdc-ibs-tra-2019b: LMU-ifo Economics & Business Data Center.

Isele, Robert and Christian Bizer (2013) "Active learning of expressive linkage rules using genetic programming," *Journal of Web Semantics*, **23**, 2–15.

IVS-IND (2019) *Ifo Investment Survey Industry 1964-2019*, Munich, DOI: 10.7805/ebdc-ivs-ind-2019: LMU-ifo Economics & Business Data Center.

Jaccard, Paul (1912) "THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.1," *New Phytologist*, **11** (2), 37–50.

Jaro, Matthew A (1989) "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *Journal of the American Statistical Association*, **84** (406), 414–420, URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1989.10478785.

Kasai, Jungo, Kun Qian, Sairam Gurajada, Yunyao Li, and Lucian Popa (2020) "Low-resource deep entity resolution with transfer and active learning," *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 5851–5861.

Link, Sebastian (2018) "Harmonization and Interpretation of the ifo Business Survey's Micro Data," *CESifo Working Paper Series* (December), URL: https://ideas.repec.org/p/ces/ceswps/_7427.html.

Link, Sebastian, Andreas Peichl, Christopher Roth, and Johannes Wohlfart (2023) "Information frictions among firms and households," *Journal of Monetary Economics*.

Loster, Michael, Manuel Hegner, Felix Naumann, and Ulf Leser (2018) "Dissecting company names using sequence labeling," *CEUR Workshop Proceedings*, **2191**, 227–238.

Loster, Michael, Zhe Zuo, Felix Naumann, Oliver Maspfuhl, and Dirk Thomas (2017) "Improving Company Recognition from Unstructured Text by using Dictionaries.," in *EDBT*, 610–619.

Lundberg, Scott M and Su-In Lee (2017) "A Unified Approach to Interpreting Model Predictions," in I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett eds. *Advances in Neural Information Processing Systems 30*: Curran Associates, Inc. 4765–4774, URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Mason, Lowell G (2018) "A Comparison of Record Linkage Techniques," (November), 2438–2447.

Meyer, Bruce D. and Nikolas Mittag (2019) "Using Linked Survey and Administrative Data to Better Measure Income: Implications for Poverty, Program Effectiveness, and Holes in the Safety Net," *American Economic Journal: Applied Economics*, **11** (2), 176–204.

Mikolov, Tomas, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin (2018) "Advances in Pre-Training Distributed Word Representations," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Mikolov, Tomáš, Wen-tau Yih, and Geoffrey Zweig (2013) "Linguistic regularities in continuous space word representations," in *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, 746–751.

Monge, Alvaro E and Charles P Elkan (1996) "The field matching problem: Algorithms and applications," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 267–270.

Moore, Jamie C., Peter W.F. Smith, and Gabriele B. Durrant (2018) "Correlates of record linkage and estimating risks of non-linkage biases in business data sets," *Journal of the Royal Statistical Society. Series A: Statistics in Society*, **181** (4), 1211–1230.

Mudgal, Sidharth, Han Li, Theodoros Rekatsinas, AnHai Doan, Youngchoon Park, Ganesh Krishnan, Rohit Deep, Esteban Arcaute, and Vijay Raghavendra (2018) "Deep Learning for Entity Matching," *Proceedings of the 2018 International Conference on Management of Data*, 19–34.

Newcombe, H B, J M Kennedy, S J Axford, and A P James (1959) "Automatic Linkage of Vital Records," *Science*, **130** (3381), 954–959, URL: `https://www.science.org/doi/abs/10.1126/science.130.3381.954`.

Newcombe, Howard B (1988) *Handbook of record linkage: methods for health and statistical studies, administration, and business*: Oxford University Press, Inc.

Newcombe, Howard B and James M Kennedy (1962) "Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information," *Commun. ACM*, **5** (11), 563–566, URL: `https://doi.org/10.1145/368996.369026`.

Ong, Toan C., Michael V. Mannino, Lisa M. Schilling, and Michael G. Kahn (2014) "Improving record linkage performance in the presence of missing linkage data," *Journal of Biomedical Informatics*, **52**, 43–54, URL: `http://dx.doi.org/10.1016/j.jbi.2014.01.016`.

Papadakis, George, Dimitrios Skoutas, Emmanouil Thanos, and Themis Palpanas (2019) "A survey of blocking and filtering techniques for entity resolution," *arXiv preprint arXiv:1905.06167*.

Pearson, Karl (1901) "LIII. On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **2** (11), 559–572, URL: `https://doi.org/10.1080/14786440109462720`.

Pedregosa, F, G Varoquaux, A Gramfort, Michel V., B Thirion, O Grisel, M Blondel, Prettenhofer P., R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay (2011) "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, **12**, 2825–2830.

Peruzzi, Michele, Georg Zachmann, and Reinhilde Veugelers (2014) "Remerge: Regression-based record linkage with an application to PATSTAT," *Bruegel Working Paper*.

Philips, Lawrence (2000) "The Double Metaphone Search Algorithm," *C/C++ Users Journal*, **18**, 38–43.

Qian, Kun, Lucian Popa, and Prithviraj Sen (2017) "Active Learning for Large-Scale Entity Resolution," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, 1379–1388, New York, NY, USA: Association for Computing Machinery, URL: `https://doi.org/10.1145/3132847.3132949`.

Řehůřek, Radim and Petr Sojka (2010) "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50, Valletta, Malta: ELRA.

Sarawagi, Sunita and Anuradha Bhamidipaty (2002) "Interactive Deduplication Using Active Learning," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, 269–278, New York, NY, USA: Association for Computing Machinery, URL: https://doi.org/10.1145/775047.775087.

Sauer, Stefan and Klaus Wohlrabe (2020) *ifo Handbuch der Konjunkturumfragen ifo Handbuch der Konjunkturumfragen*.

Schäffler, Johannes (2014) "ReLOC linkage : a new method for linking firm-level data with the establishment-level data of the IAB," *FDZ-Methodenreport*, **5**.

Schild, Christopher-J., Simone Schultz, and Franco Wieser (2017) "Linking Deutsche Bundesbank Company Data using Machine-Learning-Based Classification," *Technical Report 2017-01, Deutsche Bundesbank Research Data and Service Centre*, URL: http://dl.acm.org/citation.cfm?doid=2951894.2951896.

Schild, Christopher-Johannes (2016) "Linking "Orbis" Company Data with Establishment Data from the German Federal Employment Agency," *German Record Linkage Center Working Paper No. 2016-02*.

Shapley, L S (1953) "17. A Value for n-Person Games," in Harold William Kuhn and Albert William Tucker eds. *Contributions to the Theory of Games (AM-28), Volume II*, Princeton: Princeton University Press, 307–318, URL: https://doi.org/10.1515/9781400881970-018.

Smith, T F and M S Waterman (1981) "Identification of common molecular subsequences," *Journal of Molecular Biology*, **147** (1), 195–197, URL: https://www.sciencedirect.com/science/article/pii/0022283681900875.

Spärck Jones, Karen (1972) "A Statistical Interpretation of Term Specificity and its Application in Retrieval," *Journal of Documentation*, **28** (1), 11–21, URL: https://doi.org/10.1108/eb026526.

Tejada, Sheila, Craig A Knoblock, and Steven Minton (2001) "Learning object identification rules for information integration," *Information Systems*, **26** (8), 607–633, URL: https://www.sciencedirect.com/science/article/pii/S0306437901000424.

Ukkonen, Esko (1992) "Approximate string-matching with q-grams and maximal matches," *Theoretical Computer Science*, **92** (1), 191–211, URL: https://www.sciencedirect.com/science/article/pii/0304397592901434.

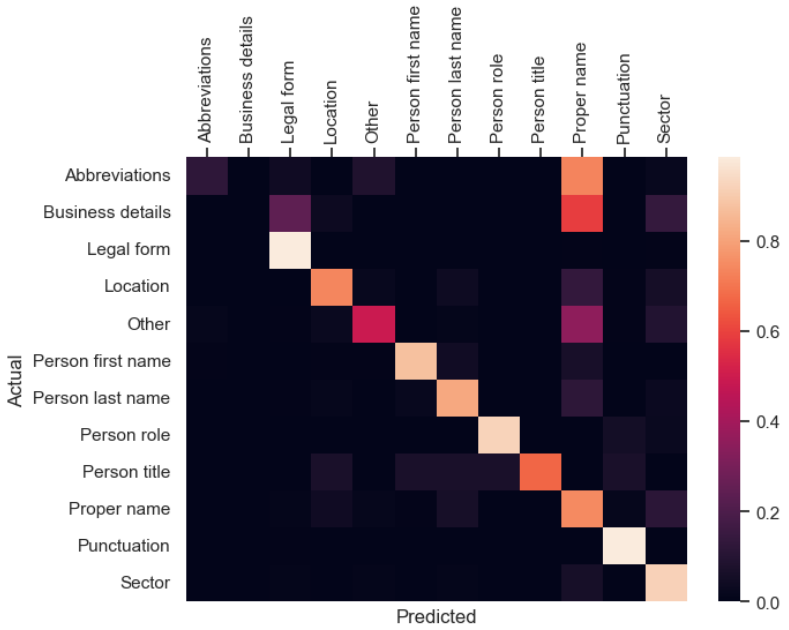
Wilson, D Randall (2011) "Beyond probabilistic record linkage: Using neural networks and complex features to improve genealogical record linkage," in *The 2011 International Joint Conference on Neural Networks*, 9–14.

Winkler, William E (1995) "Matching and record linkage," *Business survey methods*, **1**, 355–384.

Appendices

A Data

Figure A1: Confusion matrix for name segmentation



Note: The confusion matrix is based on cross validation using a held out portion of the labelled data. For each true label, it shows how many instances were predicted to be of each of the labels. The cell values represent shares of the row, i.e., of the actual labels and ideally these values would be all 1.0 on the diagonal indicating that there were no misclassifications. Brighter cells represent higher shares. *Business details* includes tokens such as *i.L.* (in liquidation). *Abbreviations* includes elements such as for example *BMW* as abbreviation for *Bayerische Motoren Werke* and can thus be easily confused with proper or colloqiual names. *Proper name* captures tokens such as *Siemens*, *Microsoft*, etc. *Location* includes words such as *Berlin*, *German*, and *International*.

Table A1: Share of entities by type of Orbis IDs

	mean
Regular IDs	0.949138
DE*-IDs	0.012705
Foreign IDs	0.002176
Branch-IDs	0.036139

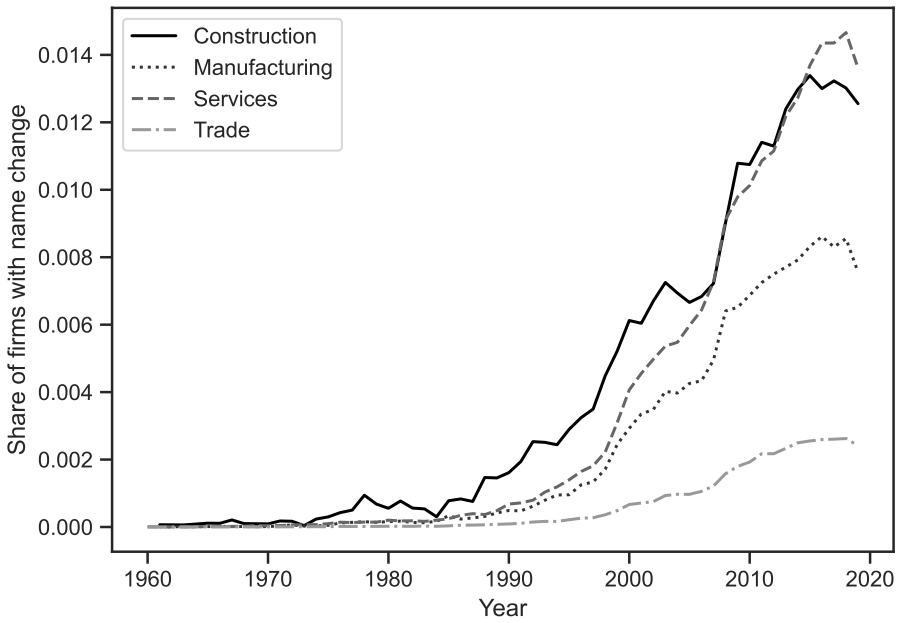
Note: *Regular IDs* take the form *DExxxxxxxx*. *DE*-IDs* indicate entities with some information that seemingly could not be matched to other variables. These usually contain no information that would be valuable for the research data set and are thus excluded. *Foreing IDs* refer to IDs that do not start with the country code *DE* but are nonetheless said to be located in Germany. *Branch-IDs* refer to IDs taking the form *DExxxxxxxxx-yyyy* where the last four digits are a running number starting with 1000 enumerating branches of the respective regular ID.

Table A2: Number of employees by survey

Survey	Employees
IBS-CON	65.28
IBS-IND	258.68
IBS-SERV	131.91
IBS-TRA	94.44
IVS-IND	442.25

Note: Contains the average number of employees by survey. Information about the number of employees is taken from the linked databases of the *BEP* and BIP and thus from the linked company databases. Thus, it is conditional on the entities being successfully linked.

Figure A2: Name changes in Orbis by business area



Note: Values are the number of name changes in a given year relative to the number of existing firms in that year. Includes only firms with available info on their date of incorporation. The increase in the share of name changes does not necessarily indicate a general trend but could be caused by Orbis being more likely to record name changes in the more recent years.

B Linkage details

B.1 Preprocessing

Preprocessing steps:

- Transform information to right data type, i.e. *integer*, *string*, ... (includes transformation sector numbers to strings due to leading 0).
- Case folding (all lowercase)
- Replace German "Umlaute" (*äöü*) and other special letters respectively with *a*, *o*, *u*, and their ascii equivalents.
- Replace special characters
- Unify different spellings of "und" (and) such as *und*, *and*, *&*, and *+*.
- Special treatment for company names:
  - Remove and extract legal form via a set of regular expressions based on Schild (2016).
  - Identify special companies within a group such as *holding* or via regular expressions.
  - Extract a selection of other common terms such as *international*, *group*, *deutschland*, *Niederlassung*.
  - Create different versions of company name:
    1. Original string
    2. No whitespace and no special characters to better deal with compound words
    3. Array of tokens
- Telephone and fax:
  - Remove country code.
  - Parse into area code and number.
- Emails: keep only domain part.
- Location data:
  - Parse addresses into *street*, *number*, and *address supplement*.
  - Standardize different spellings of "straße" (street) and remove special characters.

  - Extract occurrences of zip codes from city.
  - Remove implausible zip codes.
  - Create 1-digit, 2-digit, 3-digit, and 4-digit sector identifiers

- Gather information into arrays:
  - Distinct alternatives (names, cities, addresses, phone numbers, and fax numbers)
  - Ranged address numbers (*"5-8"* $\rightarrow$ $\{5,6,7,8\}$)

- imputations/feature generation
  - Fill potentially missing primary info (e.g. phone number) with secondary info (alternatives).
  - Infer sector section from first two digits of sector identifier.
  - Infer manufacturing, retail/wholesale, construction, and services from sector section.
  - Use Deutsche Post Direkt (2019), to infer federal state, city, zip from other location information where uniquely possible
  - Double metaphone encoding.
  - Location word embedding.

B.2 Indexing

Table B1: Indexing strategies

Strategy	Block	Filter	Pairs pre	Pairs post	Combined
1	plz (5d), legal, extra	sector (1d, multi): exact, name tokens (DM): jaro $\geq$ 0.9	106,122	14,108	120,230
2	plz (4d), sector (section), extra	name tokens (DM): jaro $\geq$ 0.8	756,962	125,804	882,766
3	plz (5d, sorted N=3), legal, sector (group), extra	add. number range: exact, name tokens (DM): jaro $\geq$ 0.8	40,415	3,833	44,248
4	city (DM), legal, sector (4d, sorted N=7)	street (DM): jaro $\geq$ 0.7, name tokens (DM): jaro $\geq$ 0.8	153,480	7,371	160,851
5	email, sector (section), extra	name tokens: exact	567,661	64,002	631,663
6	fed. state, legal, extra, city (DM, sorted N=3)	name tokens (DM)	1,309,197	185,471	1,494,668
7	city, address number	sector (2d, multi): exact, name tokens (DM): jaro $\geq$ 0.9	29,860	3,303	33,163
8	street (DM), plz (3d, sorted N=7)	name tokens (DM): exact	138,800	25,814	164,614
9	sector (section), legal, extra, plz (3d, sorted N=3)	city: jaro $\geq$ 0.9, name tokens (DM): exact	302,188	38,704	340,892
10	sector (section), legal, extra, plz (3d, sorted N=3)	street (DM): jaro $\geq$ 0.8, name tokens: jaro $\geq$ 0.8	58,392	9,382	67,774
11	fed. state, email (sorted N=7)	name tokens (DM): jaro $\geq$ 0.8	352,685	42,489	395,174
12	city (DM), extra, street (DM, sorted N=5)	name tokens: exact	100,930	18,446	119,376
13	street (DM), extra, city (DM, sorted N=3)	name tokens: exact	92,261	15,376	107,637
Extra pairs			782,851	100,685	883,536
Total			3,834,242	533,245	4,367,487

Note: Table shows the blocking and filtering keys for different indexing strategies. *Block* refers to the blocking keys, where an exact match of the entire variable is required for pairs to be considered. *Filter* refers to the variables for the filtering step conducted after the blocking, where a simple comparison metric is computed and pairs are required to have a minimum similarity in this metric or partial overlap in the variable. For computational reasons, an exact overlap in one token of an array variable, here indicated with *multi*, is computed in the filtering rather than the blocking step. Omitted from this table is the additional filtering that separates the *pre* from the *post* linkage which is based on a comparison of the *date of incorporation* from Orbis and the *survey start date* from ifo. *Extra pairs* contains pairs from existing training data pairs, some previous matches and the *branches* from Orbis IDs. *Sorted N* refers to sorted neighborhood matching and the number indicates the window size. *DM* refers to a variable phonetically encoded with with double metaphone. *Name tokens* contains the set of all name tokens from all name variants. *1d, 2d, ...* respectively refer to the number of first digits. *plz* refers to the postcode. *legal* refers to the legal form.

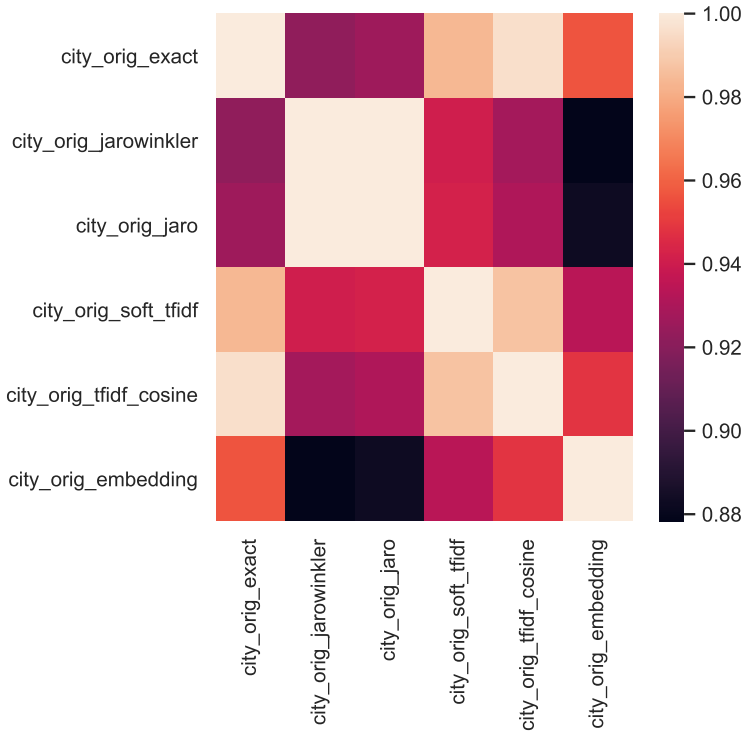
B.3 Comparison

Table B2: Most similar terms with FastText embedding vectors

Category	Term	Most similar
Not firm specific	Banane	Bananen, Ananas, banane
Not firm specific	Auto	Fahrzeug, Motorrad, PKW
Not firm specific	Firma	Tochterfirma, Herstellerfirma, Mutterfirma
Legal form	Aktiengesellschaft	Aktiengesellschaften, Kommanditaktiengesellschaft, Familien-Aktiengesellschaft
Legal form	AG	AG., AG-, AG7
Legal form	Gesellschaft	Gesellschaften, Gesell-schaft, Gesellschaft.
Legal form	GmbH	Co.KG, GbR, GmbHin
Legal form	e.v.	e.V., e.V, e.V.Im
Legal form	e.k.	e.k, ohg, e.K.
Location	Dresden	Chemnitz, Leipzig, Pirna
Location	Berlin	Potsdam, Berlin-Mitte, Charlottenburg
Location	München	Nürnberg, Augsburg, Starnberg
Location	Global	Gobal, global, European
Sector	Holding	Holdin, Holdings, Holding-Gesellschaft
Sector	Verwaltungsgesellschaft	Vermögensverwaltungsgesellschaft, Kapitalverwaltungsgesellschaft, Verwaltungsgesellschaften
Sector	Handelsgesellschaft	Außenhandelsgesellschaft, Handelsgesellschaften, Warenhandelsgesellschaft
Sector	Baugesellschaft	Wohnbaugesellschaft, Wohnungsbaugesellschaft, Union-Baugesellschaft
Sector	Bioscience	Biosciences, bioscience, Therapeutics
Sector	Schweisstechnik	Bewehrungstechnik, Schweißtechnik, Schweißtechnologie
Sector	Fertigteile	Fertigteil, Fertigteilen, Betonfertigteile
Sector	Invest	Investment, invest, Investments
Sector	Seniorenheim	Altenheim, Seniorenwohnheim, Pflegeheim
Name	Peter	Michael, Thomas, Andreas
Name	Meier	Müller, Baumann, Maier
Name	Schlenk	Schlenz, Schlenke, Schlenger
Colloquial name	Optimare	Maximare, Optimax, ComfortCtrl
Colloquial name	Airbus	Boeing, Airbusse, Airbus-Konzern

Note: Table shows which words are most similar to a selection of terms one can find in company names. Similarity is measured and terms are given for the pretrained FastText vectors used in this paper via the *most_similar* method from the python package *gensim*.

Figure B1: Correlation of metrics on the *city* field



Note: Cells show the Pearson correlation coefficient between different similarity metrics on the *city* field. The brighter the cell, the higher the correlation.

Table B3: Comparison features

Attribute	Transformation	Data Type	Method
Company name	Original	string	Exact
Company name	Name tokens	array	Monge-Elkan
Company name	Name tokens	array	Jaccard
Company name	Name tokens	array	Cosine (token)
Company name	Name tokens	array	SoftTFIDF
Company name	Name tokens	array	TFIDF-Cosine
Previous company name	Name tokens	array	Monge-Elkan
Previous company name	Name tokens	array	Jaccard
Previous company name	Name tokens	array	Cosine (token)
Previous company name	Name tokens	string	Smith-Waterman
Previous company name	Name tokens	string	Exact
Also known as company name	Name tokens	array	Monge-Elkan
Also known as company name	Name tokens	array	Jaccard
Also known as company name	Name tokens	array	Cosine (token)
Also known as company name	Name tokens	string	Smith-Waterman
Also known as company name	Name tokens	string	Exact
All company names	Name tokens	array	Multi Exact
All company names	Name tokens	array	Multi Jaro
Company name	Name without spaces	string	Exact
Company name	Name without spaces	string	LCS
Company name	Name without spaces	string	LCSSeq
Company name	Name without spaces	string	Smith-Waterman
Company name	Name without spaces	string	qgram
Company name	Name without spaces	string	cosine (ngrams)
All name variants	Array of all variants	array	Multi Exact
All name variants	Array of all variants	array	Multi Jaro
Company name segments (locations)	Name tokens	array	Multi Exact
Company name segments (locations)	Name tokens	array	Multi Jaro
Company name segments (first names)	Name tokens	array	Multi Exact
Company name segments (first names)	Name tokens	array	Multi Jaro
Company name segments (last names)	Name tokens	array	Multi Exact
Company name segments (last names)	Name tokens	array	Multi Jaro
Company name segments (sectors)	Name tokens	array	Multi Exact
Company name segments (sectors)	Name tokens	array	Multi Jaro
Company name segments (company name)	Name tokens	array	Multi Exact
Company name segments (company name)	Name tokens	array	Multi Jaro
City	Original	string	Exact
City	Original	string	Jaro-Winkler
City	Original	string	Jaro
City	Original	string	Soft-TFIDF
City	Original	string	TFIDF-Cosine
City	Original	string	Embedding cosine
City	Original	string	Frequencies
Postcode	Slices for each 1 digit to 5 digit	string	Exact
Postcode	Slices for each 1 digit to 5 digit	string	Frequencies
Street	Original	string	Exact
Street	Original	string	Jaro-Winkler
Street	Original	string	LCS
Street	Original	string	Smith-Waterman
Street	Original	string	Jaro
Street	Original	string	LSSSeq
Street	Original	string	Frequencies
Address number	Ranges	array	Multi Exact
Address number	Original	string	Levenshtein
Address number	Original	string	Frequencies
Address number	Zusatz	string	Exact
Address number	Zusatz	string	Levenshtein
Address number	Zusatz	string	Frequencies
Phone	Original	array	Multi Exact
Phone	Original	array	Multi Jaro
Email	Domain part	string	Exact
Email	Domain part	string	Jaro
Email	Domain part	string	Frequencies
WZ08	Primary sector classification slices for each 1 digit to 4 digit	string	Exact
WZ08	Primary sector classification slices for each 1 digit to 4 digit	string	Jaro
WZ08	Primary sector classification slices for each 1 digit to 4 digit	string	Frequencies
WZ08	All sector classification slices for each 1 digit to 4 digit	array	Multi-Exact
Legal form	Categories	integer	Exact

Note: Table shows the similarity metrics that were used for each of the company attributes. The datatype *array* refers to a set of strings rather than one consecutive string. An example for this is {"Bayerische", "Motoren", "Werke"}. The methods *Multi Exact* and *Multi Jaro* refer to metrics where the respectively the maximum exact or Jaro score are measured in a cross comparison between all tokens of both records. The method *Frequencies* refers to a feature containing the relative frequency of the respective value of the records.

B.4 Classification

Table B4: Sizes of the training data

	Size	Share
Training data 1	8,307	0.56
Training data 2	3,561	0.24
Test data	2,968	0.20

Note: Training data 1 is used to train the individual components of the model, *training data 2* is used to then train the ensemble aggregator model, and *test data* is used to assess the quality of the model.

Table B5: Ensemble components with classification metrics

Estimator	Description	Accuracy	Precision	Recall	F1-Score
LogisticRegression	feature aggregation, continuous features	0.848	0.777	0.895	0.832
LinearSVC	feature aggregation, continuous features	0.850	0.774	0.907	0.835
MLPClassifier	feature aggregation, continuous features	0.865	0.828	0.857	0.842
XGBClassifier	feature aggregation, continuous features	0.877	0.823	0.901	0.860
LogisticRegression	feature aggregation, continuous features	0.850	0.779	0.896	0.834
LinearSVC	feature aggregation, continuous features	0.849	0.772	0.909	0.835
MLPClassifier	feature aggregation, continuous features	0.876	0.818	0.905	0.859
XGBClassifier	feature aggregation, continuous features	0.878	0.817	0.912	0.862
LogisticRegression	frequency weights, no missing data indicators	0.852	0.785	0.891	0.835
LinearSVC	frequency weights, no missing data indicators	0.849	0.776	0.899	0.833
MLPClassifier	frequency weights, no missing data indicators	0.877	0.817	0.909	0.861
XGBClassifier	frequency weights, no missing data indicators	0.886	0.847	0.888	0.867
LogisticRegression	continuous features	0.850	0.778	0.899	0.834
LinearSVC	continuous features	0.850	0.773	0.907	0.835
MLPClassifier	continuous features	0.865	0.809	0.887	0.847
XGBClassifier	continuous features	0.872	0.819	0.891	0.853
RandomForestClassifier	categorical features, binned continuous features	0.865	0.787	0.927	0.851
CatBoostClassifier	categorical features, binned continuous features	0.890	0.848	0.898	0.872
RandomForestClassifier	no missing data indicators, binned continuous ...	0.871	0.800	0.924	0.857
CatBoostClassifier	no missing data indicators, binned continuous ...	0.866	0.824	0.864	0.843
MLPClassifier	frequency weights, categorical features	0.882	0.848	0.876	0.862
LogisticRegression	PCA	0.861	0.798	0.892	0.843
LinearSVC	PCA	0.860	0.790	0.906	0.844
MLPClassifier	PCA	0.890	0.848	0.898	0.872
Ensemble		0.898	0.852	0.916	0.883

Note: Overview over the individual models that enter the ensemble. Each model has its own pipeline with various transformation and selection steps such as aggregation of all location-based features. These steps are shown in the description column. *MLPClassifier* is a Multilayer Perceptron, i.e., a Neural Network. *XGBClassifier* and *CatBoost* are Gradient Boosting classifiers with the latter supporting categorical features. *LinearSVC* refers to a Support Vector Machines algorithm with a linear kernel. *Accuracy* is the share of correct predictions, *Precision* is the share of correct prediction among the predicted matches, *Recall* is the share of actual matches predicted as match, and *F1* is the harmonic mean of precision and recall. *MLPClassifier* is a Multilayer Perceptron, i.e., a Neural Network. The *Ensemble* is a logistic regression that takes the predictions from the models above as inputs to make the final match prediction.

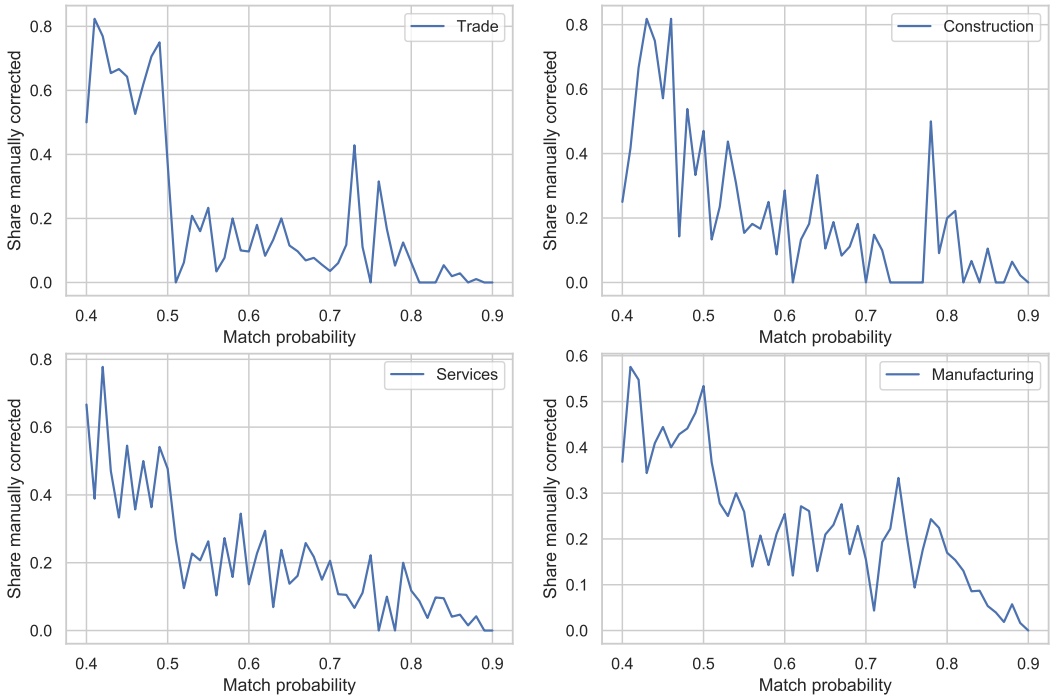
B.5 Postprocessing

Table B6: Multiple matches per ifo ID in the pre linkage

Matches per ifo ID	1	2	3	4	5	6+
Absolute	22210	5243	1230	330	119	134
Share of IDs	0.76	0.18	0.04	0.01	0.00	0.00

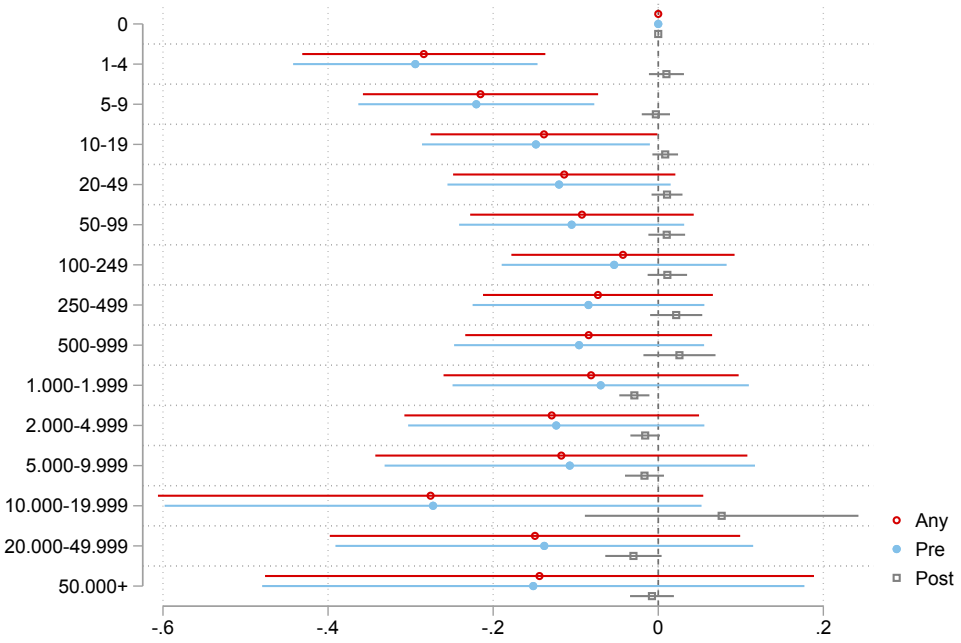
Note: This table shows that, before postprocessing, matches are not unique per ifo ID for around 24% of entities. Hence, this needs to be reduced to one BvD match per ifo ID. These values are not yet indicative of the final match rate after postprocessing. See section 7 for that. *Share of IDs* refers to the share of matched ifo IDs and it sums up to one with some rounding imprecision.

Figure B2: Manual corrections by predicted match probability



Note: This figure shows the share of pairs that were manually corrected. Thus, for probabilities above 0.5, corrections reflect false positives corrected to negatives and for probabilities below 0.5, they reflect false negatives corrected to positives. Corrections were mostly conducted above the 0.5 threshold and the corrections below are based on a very small number of samples. The high number of corrections in the area below 0.5 is in part due to the fact that here, more or less obvious cases were selected, with a focus on false negatives, from a screening without further analysis of more complex cases.

Figure B3: Regression of match status on firm characteristics - size ranges



Note: Coefficients of a regression of match status on other characteristics from the ifo companies. Only coefficients on employment size range dummies are reported here, the others are shown in figure 5 for better visibility. Point estimates of a linear probability model are shown with 95% confidence intervals based on robust standard errors.

C Access

Access to the data can be granted for purely scientific purposes at the *LMU-ifo Economics and Business Data Center* (EBDC) located at the ifo Institute in Munich, Germany. Because the data are confidential, access is only possible on-site on a protected workstation. The data contain no identifiers such as company name or address and it is prohibited to re-individualize individual companies. Only aggregated results, such as regression tables, can be exported and will be controlled by EBDC staff.