



EBDC DATASET

LINKAGE

IFO BUSINESS SURVEY IN MANUFACTURING

AMADEUS

HOPPENSTEDT

31 July 2009

Economics and Business Data Center

Ifo Institute for Economic Research
Poschingerstr. 5
81679 Munich

by

Anja Hönig, Heike Mittelmeier, Andreas Neudecker

Abstract

The Economics and Business Data Center (EBDC), founded as a cooperation of the University of Munich (LMU) and the Ifo Institute for Economic Research in 2008, aims at opening new fields of economic research by providing an innovative, continually updated dataset of German companies. The dataset is based on the extensive micro databases of the Ifo Institute, these being supplemented with external financial statement data and other structural information concerning corporate finance and governance. In general, because of the high levels of confidentiality and data security the Ifo Institute ensures its panel members, the EBDC dataset can only be used for research purposes on the premises of the EBDC. Furthermore, the data is only provided in an anonymised way and with a one-year time lag and its use is subject to strict security precautions. The aim of this paper is to give an overview of the data sources and to describe the scope of and access to the new EBDC dataset. Furthermore, it also provides information on the matching method of probabilistic record linkage.

Table of contents

Table of contents	1
List of illustrations	1
1 Introduction	2
2 Data Sources of the EBDC Dataset	4
2.1 Ifo Business Survey for Manufacturing	4
2.2 Company Database Amadeus	6
2.3 Hoppenstedt Accounting Database	8
3 Probabilistic Record Linkage	9
4 Structure of the EBDC Dataset	13
5 Expansions	16
6 Access	17
7 Bibliography	18

List of illustrations

Fig. 1 Contents Ifo Business Survey in Manufacturing	5
Fig. 2 Scope of Company Database Amadeus (Europe-wide)	7
Fig. 3 Hoppenstedt Accounting Database	9
Fig. 4 Generation of the EBDC Dataset	14

1. Introduction

The LMU Economics and Business Data Center (EBDC) was founded as a cooperation of the economics and business administration faculties of the University of Munich (LMU) and the Ifo Institute for Economic Research at the beginning of 2008. It receives funding from *LMUexcellent*, this being a project within the framework of the *Exzellenzinitiative* of the German federal and state governments for the promotion of science and research at German universities.

The Economics and Business Data Center has the aim of creating an innovative dataset that links the Ifo Business Survey and later also the Ifo Innovation and the Ifo Investment Survey with different external company databases containing financial statement or corporate governance data. By this means, we want to supplement the more qualitative aspects (business expectations, assessments, etc.) contained in the micro databases of the Ifo Institute with accounting and structural information of companies and thus facilitate innovative approaches for empirical economic research. Therefore, the tasks of the EBDC also include the procurement and administration of data sources for research and teaching, the central provision, updating and documentation of external databases, as well as the acquisition of corresponding support tools.¹ Moreover, the EBDC provides a suitable hard- and software technical infrastructure and offers support with regard to software-specific knowledge transfer. Additionally, within the framework of the Ifo Datapool, it offers access to the survey data of the Ifo Institute,² which has conducted regular surveys throughout Germany since 1949.³ Thus, interested researchers and students working on empirical projects may profit from synergy and efficiency effects.

With regard to the new EBDC enterprise dataset, in a first step, we have linked the micro data from the largest sector of the Ifo Business Survey, namely manufacturing (KT VG), with balance-sheet data from the enterprise databases Amadeus and Hoppenstedt. As a result, we can offer an extensive company panel which integrates different databases thus allowing for the simultaneous research of rather qualitative

¹ Accessible external databases at the EBDC include Bankscope and Thomson OneBanker.

² The Ifo Institute and the EBDC also provide access to other external micro and macro data.

³ For more information about Ifo micro data, see Becker, S. O. and K. Wohlrabe (2008).

together with quantitative business factors. The EBDC dataset contains both historical and current data and focuses on the period 1994-2007, however, we also plan to continually update the database.

The aim of this paper is to give interested professors, guest researchers and doctoral students a general overview and to describe the sources, scope of and access to the new EBDC dataset. Furthermore, it also provides information on the matching method of probabilistic record linkage. To this end, Section 2 describes the data sources, i.e. the Ifo Business Survey as well as the company databases Amadeus and Hoppenstedt. An overview of the linkage method is provided in Section 3, while Section 4 explains the structure and components of the EBDC dataset. A short insight in planned extensions is given in Section 5, finally, Section 6 describes the access to the data.

2. Data Sources of the EBDC Dataset

To generate the EBDC dataset described in this article we have linked the micro data from the Ifo Business Survey for manufacturing with external balance-sheet data from the enterprise databases Amadeus and Hoppenstedt. In this section, we briefly describe the individual data sources, while in Section 3, the matching technique of probabilistic record linkage is explained.

2.1. Ifo Business Survey for Manufacturing

In general, there are four regularly conducted standard enterprise surveys at the Ifo Institute: the Ifo Business Survey (KT), the Ifo Investment Survey (IT), the Ifo Innovation Survey (INNO) and the Ifo World Economic Survey (WES) with the KT being conducted monthly. Unlike the other Ifo surveys which concentrate on investment behaviour (IT, semi-annual), on innovation activities (INNO, annual) or the international outlook for economic activity (WES, quarterly), the KT focuses on enterprise-specific appraisals and expectations concerning business as well as market conditions. Furthermore, the monthly published Ifo Business Climate Index which is based on the Ifo Business Survey makes it of particular interest for empirical panel research.

The Ifo Business Survey is structured around four sectors: manufacturing (KT VG), wholesaling/retailing (KT HAN), construction (KT BAU) and service providers (KT DVDL). In each sector, the questionnaires contain standard monthly as well as periodically recurring special questions with the questionnaires relating to a product or product line (KT VG, KT DVDL) or to a sector or business field (KT HAN, KT BUILDING), respectively. For this reason large enterprises whose main sales consist of the production of several products frequently provide information on more than one area and therefore receive several questionnaires. More detailed information on the KT sectors, the topics of the questions posed and numerous studies based on this survey can be found in Becker and Wohlrabe (2008).

With respect to the Ifo Business Survey for manufacturing the focus of the questions is the following:

<i>Standard questions:</i>	<i>Special questions:</i>
<i>Order stocks, capacity utilisation, constraints to production, competitive position (at home and abroad), employment, inventories</i>	<i>Profits, access to credit, innovations, special occasions</i>

Fig. 1: Contents Ifo Business Survey in Manufacturing

In the business surveys of the Ifo Institute, the participants are generally not requested to provide absolute or monetary figures, since experience has shown that a survey of exact purchasing plans does not provide an accurate picture of reality. This is due to the fact that half of the purchase decisions of private households, as well as those of small and medium-sized enterprises, are spontaneous.⁴ Therefore, monthly “appraisal questions” such as the question on current business conditions or the demand situation have proven to be reliable since these implicitly mirror also a company’s actual and expected profit. In addition, as enterprise decisions are influenced by judgements and other subjective factors, the variables also provide an outlook on the probable direction of economic activity, which is published monthly as the Ifo Business Climate. In this regard it is important that the Ifo approach is not limited to Germany, instead it is also supported by the European Commission and the OECD. Therefore, in January 2002 and as a result of the Europe-wide harmonisation of business surveys, there were also some changes in the Ifo Business Survey. Specifically, there is no longer a distinction between survey and questionnaire months.

Both for manufacturing (KT VG) as well as for the other survey sectors, the participating companies are presented with binary or ordinal scaled response categories⁵ i.e., for each standard question there are only two (“1” yes, “2” no) or three different possible responses (“1” better, “2” the same, “3” worse) to choose from. Thus,

⁴ See Goldrian (2004).

⁵ In some instances percentage specifications are required.

only assessment statements on a company's current (and anticipated) business situation are possible. Moreover, the response also depends on the respective interpretation: the question of the "business situation" indeed refers to the overall economic condition of the business, but it is left to the company to determine the basis on which to make this judgement.

As already mentioned, the KT VG contains responses of the enterprises on the basis of the products they manufacture. It can be the case, for example, that an enterprise returns several questionnaires in a month if it produces several products. Hence, in total, the Ifo Business Survey for manufacturing contains 300 product groups which are defined such that they are maximally homogeneous in themselves. Specifically, in order to attain "authentic survey data"⁶ consideration was given both to area representativeness (product variety) as well as to company representativeness (size, legal status, etc.). Since 1991 the KT VG has no structural breaks so that, over the years, there is an average of 3000 responses per month resulting in a return rate of 92%. Intense contacts with the companies maintain the panel size and composition at a representative level.

2.2. Company Database Amadeus

In addition to the Ifo KT VG and the Hoppenstedt database, the Amadeus company database is the main source for the first version of the EBDC dataset. It is a product of the Bureau van Dijk Electronic Publishing GmbH (BvDEP), one of the leading European providers of global enterprise information, and contains business and finance information on more than 11 million, mainly non-quoted enterprises, from 41 countries in Europe. Currently, about 1 million German enterprises are registered.

For the company databases information of market-leading local institutions and well-known businesses of the respective countries are used. The financial closing data for German businesses come from Creditreform or Creditreform Rating AG, which belongs to the Creditreform Group. The Creditreform Group has provided bank credit information appertaining to customer-supplier relationships for more than 125 years and is the European market leader for bank credit information. The key source for the

⁶ See Goldrian (2004).

Amadeus database used by the EBDC dataset is the MARKUS Database, which contains business information of companies in the German Commercial Register with a bank credit index of a maximum of 499 (Creditreform Association) and the DAFNE database including annual accounts, investment data, etc. of all disclosing German firms (Creditreform Rating AG).

Country	Top 250,000	Top 1.5 million	All Companies	Country	Top 250,000	Top 1.5 million	All Companies
Austria	4,204	26,902	123,114	Belarus	825	956	967
Belgium	9,328	45,072	340,484	Bosnia-Herzegovina	188	1,601	2,993
Cyprus	94	182	250	Bulgaria	2,390	22,216	112,116
Czech Republic	6,575	27,636	67,227	Croatia	1,339	7,231	20,007
Denmark	6,075	34,538	170,378	Iceland	236	1,623	18,239
Estonia	728	7,575	58,449	Liechtenstein	26	104	220
Finland	3,578	19,420	83,607	Macedonia	285	804	1,331
France (inc. Monaco)	26,722	175,766	963,947	Montenegro	260	2685	2,835
Germany	33,567	209,918	834,513	Moldova	28	179	987
Greece	2,656	17,954	28,074	Norway	7,473	47,222	193,199
Hungary	2,113	13,945	37,158	Romania	3,967	39,635	513,555
Ireland	3,931	19,137	132,179	Russian Federation	24,257	123,824	497,747
Italy	23,764	180,682	567,600	Serbia	1,490	9,521	40,555
Latvia	833	3,810	6,585	Switzerland	3,888	11,429	32,254
Lithuania	875	5,111	7,842	Ukraine	6,366	20,456	25,978
Luxembourg	566	1,609	3,699	Total no. of companies	276,590	1,593,504	8,619,828
Malta	161	758	2,314	Coverage in November 2006. The numbers of companies is continually increasing. Please visit bvdep.com for the most up to date coverage figures.			
Netherlands	12,599	88,647	359,484				
Poland	8,247	22,777	29,922				
Portugal	3,777	22,241	87,748				
Slovak Republic	1,390	4,813	6,750				
Slovenia	882	5,413	36,218				
Spain	17,764	165,038	852,330				
Sweden	10,644	58,745	258,502				
United Kingdom	42,499	146,329	2,098,471				
Total EU coverage	223,572	1,304,018	7,156,845				

Fig. 2: Scope of Company Database Amadeus (Europe-wide)

Unlike the DAFNE database (raw data format), the data in Amadeus are in a homogeneous, standardised accounting format based on generalised national and/or international accounting rules.

Every enterprise report consists of a total of 23 accounting items, 25 positions of the financial statements, 20 key finance figures and numerous descriptive information such as industry codes, partnership structures, stocks and stock price information. For the enterprises in the EBDC dataset, more than 50 positions have been selected, but

initially not including partnership, stock and stock price information.⁷ In addition, with regard to currentness, Amadeus guarantees that the financial closing data are available in the database after 15 months at the latest.

2.3. Hoppenstedt Accounting Database

The Hoppenstedt Accounting Database is a product of Hoppenstedt Business Information GmbH, which is one of the leading providers of business and industry information in Germany and is a part of the Hoppenstedt Group. Key business areas, in addition to the provision of company information, are the sales of postal addresses, credit and risk analyses as well as the publication of technical journals, so that – depending on customer status and rights of use – choices can be made from a great number of databases. For the company databases, information from external sources such as the Federal Official Gazette, the Commercial Register, the business press or annual reports, is used and if required also derived in direct dialog. According to Hoppenstedt all known changes are updated, evaluated and incorporated daily into the corresponding database, which is why the supplied business information is marked by currentness, quality and data depth.

The information for the EBDC dataset was taken from the Hoppenstedt Accounting Database,⁸ currently containing more than 2.7 million closing statements from more than 1 million German enterprises in the areas of manufacturing, distribution, services, insurance and banks.

Almost all final statements published since 2005 are registered here and the historical information for large firms even dates back in part to 1987. The collected data on accounts and financial statements of individual companies are accessible in varying levels of detail (norm accounting: maximal available positions according to the respective accounting regulations; abridged accounting: ca. 90; short accounting: ca. 30 positions). Moreover, for taking into consideration the different types of final statements separate accounting schemes, which are closely oriented on the respective original, were developed for HGB, IAS and US-GAAP.

⁷ These, however, can be exported via the historical databases in the EBDC.

⁸ At the time of data collection for the EBDC, ca. 120,000 enterprises were included in Hoppenstedt.

The screenshot shows the Hoppenstedt Bilanzdatenbank interface. On the left is a sidebar with search filters like 'Datenbestand', 'Firmenname', 'Ort', 'PLZ', 'Rechtsform', 'Branche', 'Beschäftigte', 'Absatz', 'Max. Abschluss', 'Alt. Abschluss', 'Rechnungslegung', 'Abschlussdatum', and 'Statistik'. The main area is titled 'Trefferliste' and contains a table of search results. Above the table are buttons for 'Alle markieren', 'Trefferliste drucken', 'Selektion speichern', 'Merkliste', 'Alle denartieren', and 'Trefferliste exportieren'.

Firma	PLZ	Ort	Bilanzsumme in TEUR	Abschätze
1. <input type="checkbox"/> DZV Deutscher Zeitungsverlag GmbH	60327	Frankfurt am ...	19	1
2. <input type="checkbox"/> Frankfurter Societäts-Druckerei GmbH	60327	Frankfurt am ...	225.435	20
3. <input type="checkbox"/> Heidelberger Mediengestaltung - HVA Ge...	69117	Heidelberg	2.711	2
4. <input type="checkbox"/> Impuls Verlagsgesellschaft mbH	69167	MANHEIM		2
5. <input type="checkbox"/> iz IMMOBILIENZEITUNG Verlagsgesellsch...	65185	Wiesbaden	1.411	2
6. <input type="checkbox"/> Kühnverlag AG	69623	Lampertheim	21.509	3
7. <input type="checkbox"/> Limburg Land Presse-Vertriebsgesellsch...	65549	Limburg	69	3
8. <input type="checkbox"/> Mannheimer Morgen Großdruckerei und Ve...	68021	Mannheim	14.746	3
9. <input type="checkbox"/> Medienhaus Südbessen GmbH	64201	Darmstadt	39.369	15
10. <input type="checkbox"/> Motac Medien Verlags GmbH	63071	OFFENBACH		2
11. <input type="checkbox"/> Rhein-Neckar-Zeitung GmbH	69117	Heidelberg	43.695	7
12. <input type="checkbox"/> Saarbrücker Zeitung Verlag und Drucker...	66117	Saarbrücken	218.559	6
13. <input type="checkbox"/> Schwetzinger Verlagsdruckerei GmbH	68709	Schwetzingen	11.035	2
14. <input type="checkbox"/> Schwetzinger Zeitungsverlag GmbH & Co...	68723	Schwetzingen	1.269	2
15. <input type="checkbox"/> Union-Druckerei- und Verlagsanstalt Ge...	60444	Frankfurt am ...	35.437	6
16. <input type="checkbox"/> Verlag Frankfurter Stadtanzeiger GmbH	60529	FRANKFURT		2
17. <input type="checkbox"/> Wilhelm Krauth GmbH	69403	Eberbach, Br	1.284	2
18. <input type="checkbox"/> -das inserat. Verlag GmbH	63303	Dreieich	494	2

Fig. 3: Hoppenstedt Accounting Database

3. Probabilistic Record Linkage

To link the different data sources, recourse is taken to the companies' address information contained in each database. By this means, we can generate two allocation tables (Ifo-Amadeus and Ifo-Hoppenstedt) and combine them in the EBDC dataset. In the following, as an example, we give a short overview of the record linkage of the address data of the Ifo Business Survey with that of the Amadeus company database which is exactly the same in the case of the Hoppenstedt or any other database. The only difference is that here, additionally, a so-called "gold standard" was created to determine the match or non-match weights for every address variable. The 28,636 data records of the Ifo Business Survey (including construction, wholesaling/retailing, service providers, status: 03/2008) were linked with the 923,946 German companies in the Amadeus database (status: 01/2008 and 10/2008) by using

the matching software MTB (Merge Toolbox) developed at the Center for Quantitative Methods and Survey Research of the University of Konstanz.

MTB is especially useful when there is no unambiguous key, such as a company code number, in the dataset as it allows the allocation of different datasets containing large amounts of data with different postal-address presentations. Specifically, using the method of Probabilistic Record Linkage which is based on the theory of Newcombe et al. (1959) and was formalised by Fellegi and Sunter (1969), similar name/address data can be linked whereby the degree of agreement, the so-called “similarity” of the variables, is determined by probabilities.

This “similarity” is calculated from the quotient of the probability that the variable x from both datasets is identified as concurrent for similar companies (M probability) and the probability that the variable x from both datasets is evaluated as concurrent for non-similar companies (u probability). Ideally, the quotient amounts to $1/0$. Since deviations between the different address variables are not equally important, we use the logarithm of the M (= match weight) or the u probability (= non-match weight) of the respective variable to weight in case of a variable match or non-match. In sum, these (positive/negative) variable weights make up the overall “quality”, i.e. the degree of similarity of each address linkage.

As these parameters (weights) of the Probabilistic Record Linkage have to be gained empirically, an ensured partial set from the amount of data records is initially formed. This data subset is called a “gold standard”, whereby, in general, a gold standard refers to a linkage that allows for an unambiguous allocation of the data records of two databases. Thus, this first record linkage is performed using telephone, fax numbers and e-mail addresses. Since, however, these variables were not filled in all data records and since the systematics of telephone and fax number entries differed as well, a linkage of all the data was not possible in this way. Nevertheless, by this means, we received 11,225 matches meaning that about 40 percent of the Ifo entries could be directly assigned to an enterprise in Amadeus.

However, a closer look revealed that also data records actually not belonging together were identified as matches. Frequently, this is the case if there are several subordinate

units within a company which have the same central telephone and fax numbers or e-mail addresses – for example administrative units, holding and management companies. Therefore, several of the linked gold-standard pairs were not true matches in the literal sense. They were only true matches in the sense of their affiliation to a larger enterprise group.

Within these linked data records, we then computed weights for the individual address variables. To this end, the number of correct/false concurrences was compared with the total population of comparisons. In addition, in order not to distort the results, the weights were determined for different preparation variants of the variables. By this means, we could avoid giving too positive assessment of the results due to frequently occurring name sequences or to give too negative assessments due to different spellings of the same name: the variant of the match variables with the best differentiation is the one having the greatest number of true agreements (M probability) without displaying an increased number of wrong agreements (u probability).

As already mentioned, the gold standard also contains matches that could not be designated as such in the true sense. Therefore, the correctness of the determined weights was again controlled by hand by means of a random sample, comprising 2000 data records from the gold standard. From the resulting similarity weights, which were allocated to the variables before the actual MTB run, we could then calculate the quality. If positive matches were determined for all variables when linking companies' addresses from both databases, the quality resulted as the maximum total weight from the sum of the individual concurrence weights. For example: $10.6 + 8.6 + 4.9 + 12.9 = 37.2$. On the other hand, if MTB identified one of the variables as a non-match, the (negative) non-match weight was used in the equation, so that the overall quality of the respective address linkage decreased accordingly.

The comparison of the variables from the Ifo and the Amadeus databases was carried out by means of a string similarity function, comparing N grams of length 2 (= bi-gram with blanks before and after all strings) from the respective variables by placing a raster of the length 2 over the string. Differences in the variable characteristics were weighted linearly according to bi-gram similarity, with MTB giving a lower evaluation to cases where there is a high agreement with a shorter name than in the case of a high

agreement with a longer name. In addition, the determination of a Jaro factor for all variables (= weight adjustment) to the value “2” led to a faster allocation of the full agreement weight in the case of high concurrence and thus resulted in a better differentiation of the matches from the non-matches.

Finally, for the evaluation of the MTB run, we had to define a threshold value for the quality variable. In doing so, however, recourse was not taken to an evaluation of the program, which as of a certain quality identifies the linkage as a match. Instead, a higher value was applied from the very start and, once again, we conducted a manual control for a larger quality area. As before, the idea here was to avoid the error of a false positive: pairs do not qualify as true matches only because they are above a threshold value (e.g., a place in a name can increase similarity or there can be the wrong legal form but the same name). On the other hand, some true matches, which were too strongly devalued because of lacking information (e.g. same name, street empty), can still be found below the defined threshold value in certain blocks.

With this procedure, 28,636 enterprise postal addresses from the Ifo Business Survey were compared with 923,046 enterprise postal addresses from the Amadeus company database. This led to a total of 103,377,535 pair comparisons, whereby in the subsequent manual control 9,472 were classified as matches.⁹ Of these, 4,073 matches belong to the Ifo Business Survey for manufacturing, while the rest comes from the other Ifo Business Survey sectors (construction, wholesaling/retailing, service providers). Furthermore, since the Ifo KT VG is conducted on the product level, it can occur that one company in Amadeus is allocated to two or more Ifo entries. The number of linked companies thus amounts to 8,915 or to 3,819 for the sub-group of manufacturing.

In comparison, the linkage of Hoppenstedt with the Ifo Business Survey resulted in a total of 4,811 matches or 4,377 enterprises, of which 2,703 linkages (2,454 companies) are allocated to manufacturing.

⁹These results are primarily attributable to the differing currentness of the two address databases. While Amadeus companies not having reported for more than 3 years are deleted; the Ifo panel contains all the addresses of the past 20 years, even those that have already been cancelled.

4. Structure of the EBDC Dataset

As described in the previous section, we used the identifying address variables to link the companies of the Ifo Business Survey with those of the accounting databases Amadeus and Hoppenstedt. If both linkages are subsequently combined, the result is an allocation table containing, for each company, the key variable from the Ifo KT VG (Ifo-ID) as well as the key of the respective accounting database (Amadeus-ID and/or Hoppenstedt-ID). In this regard, one has to take into consideration that some companies of the Ifo panel can only be found in Amadeus, only in Hoppenstedt or in both enterprise databases. Therefore, there is a total of 3,858 allocations (IDs), however, both balance-sheet and Ifo information in the same year is only available for 2,338 enterprises or 2,473 products.

Every response to the Ifo Business Survey can thus be assigned to an enterprise or, if there also exists a balance-sheet, to the financial statement data from Amadeus or Hoppenstedt and is identifiable via the newly generated variable "EBDC_ID". As already mentioned, the responses of the Ifo KT VG refer to individual products, which is why a company (a balance-sheet) that manufactures several products can correspond with several responses to the Ifo Business Survey. For this reason the EBDC_ID consists of two components: the variable "EBDC_company", a running company number and the variable "EBDC_product" indicating the various products of the company and which is integrated in the last digit of the EBDC_ID.

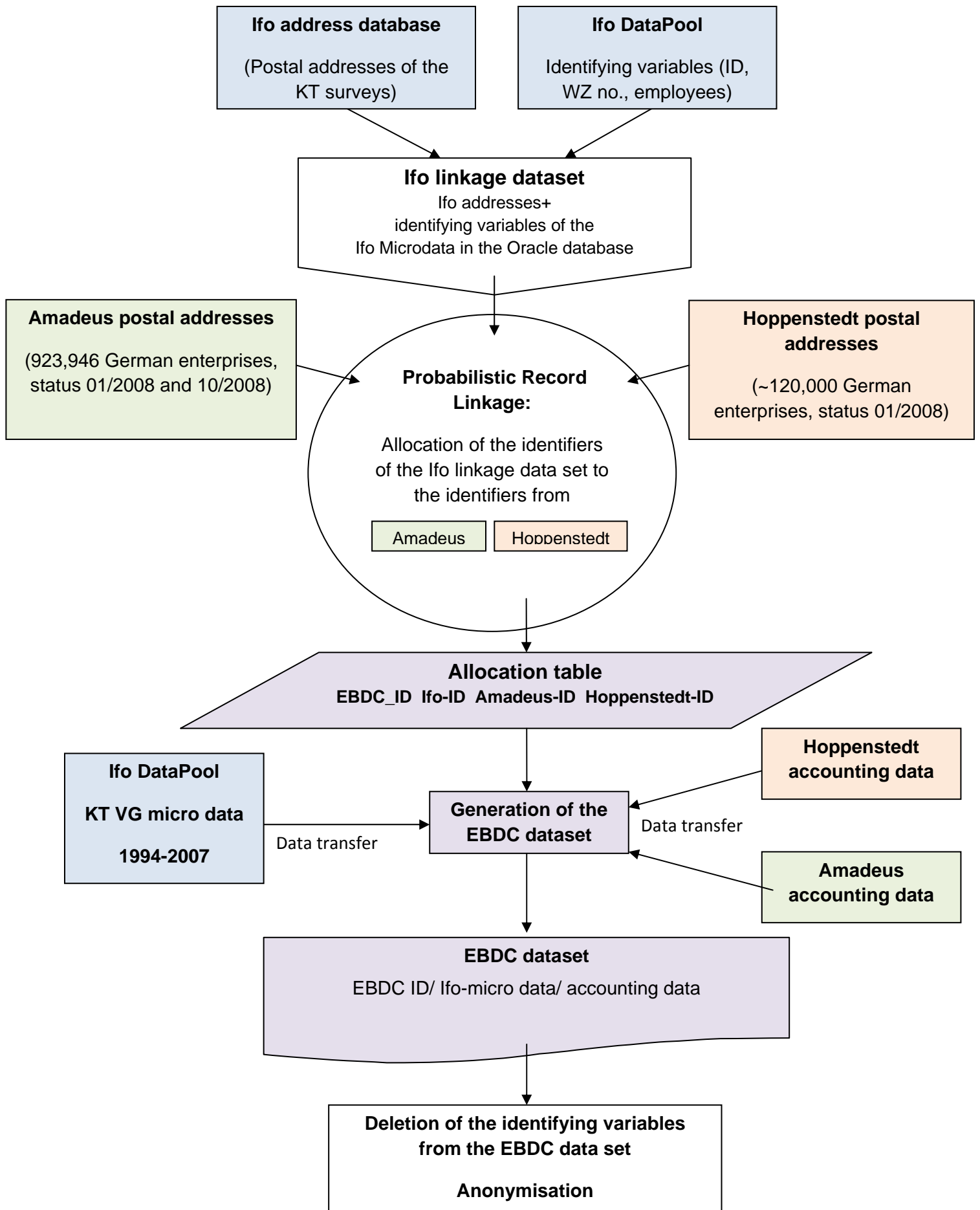


Fig. 4: Generation of the EBDC Dataset

As can be seen in Figure 4, having allocated the respective identifiers, we can finally transfer the micro data from the Ifo Business Survey for manufacturing as well as the accounting data of Amadeus and Hoppenstedt and thus get a combination of monthly and annual data over the period 1994–2007.

In general, when available, the EBDC dataset contains information from individual instead of corporate group accounts,¹⁰ however, we did not take over the accounting schemes from the original databases. Instead, we developed a new EBDC scheme which integrates both Amadeus as well as Hoppenstedt variables and abstracts from the existing differences of the two databases.¹¹ Specifically, the EBDC accounting scheme is based on the accounting and earnings-statement structure of the German Commercial Code (HGB) and in part also contains variables according to total or turnover cost procedures.¹² If there was available both information from Amadeus and Hoppenstedt for a specific company, preference was given to the latter due to a more extensive variable selection and deliverance.¹³ For a detailed and correspondingly structured overview of the available accounting and earnings-statement variables, see the list of variables.

The dataset is sorted according to EBDC_ID, year and month and is presented in a long format, i.e. every response is identifiable via these three variables. Each month contains the survey results from the Ifo KT VG resulting in up to 12 monthly reports per product/ company in each year.

Sorted according to their function, the following variables are included in the columns of the EBDC dataset: identification variables, accounting and GUV as well as Ifo variables. In addition to EBDC_ID, year and month, the identification variables also contain information such as the industry code, company size, German federal state,

¹⁰ The respective type of reported account is indicated by the variable “reporting_basis”. In this context, limited financial data means that the accounting information was not published but requested individually.

¹¹ The conversion scheme employed to transfer the initial variables into the newly generated EBDC accounting variables can be viewed at the EBDC. Furthermore, there is also a detailed standard accounting scheme available that can be used for orientation.

¹² Upon request the EBDC dataset can be made available with the initial accounting variables from Amadeus and Hoppenstedt as well.

¹³ In general, accounting information from Amadeus is in thousands of euros and rounded off; furthermore individual positions are often aggregated. In contrast, Hoppenstedt displays raw data with an accounting-structure being closely oriented to the HGB.

stock exchange quotation and legal form.¹⁴ In this regard, the list of variables does not only give a general overview but also sum up important features of the variables (e.g. questions of the individual surveys, survey frequency, ect.).¹⁵

In total, there were 3,858 allocations (IDs) and 17,850 accounting results linked to 278,947 Ifo responses for the whole period. A first evaluation of the dataset (descriptive statistics, filling of selected variables, etc.) can be seen and duplicated at the EBDC.

5. Expansions

With the EBDC dataset, the Economics and Business Center makes possible new approaches for empirical research in the field of business administration and economics and contributes to the promotion and development of top-level research.

The newly created dataset will be regularly updated and continuously expanded. Currently, an update of 2008 data is being worked on. The next step will then be the expansion of the existing enterprise panel by integrating further companies that have been added to the respective address databases (see Section 3).

Another option is to link the information from the remaining sectors of the Ifo Business Survey with the balance-sheets from Amadeus and Hoppenstedt, i.e., the sectors of construction (KT BAU), wholesale/retail (KT HAN) and service providers (KT DVDL). In this regard, however, it must be taken into consideration that the surveys in construction and wholesale/retail are not directed to various product groups but to individual construction and distribution sectors. By contrast, just as the KT VG, the already mentioned Ifo Innovation Survey (INNO) is aimed at individual products of an enterprise. Specifically, the firms included in the annual innovation survey are a subset of the firms in the Ifo Business Survey for manufacturing. Therefore, linking INNO with Amadeus and Hoppenstedt is a sensible supplement of the existing data basis. The

¹⁴ Due to anonymisation, the federal state information for very large firms (> 10.000 employees) was deleted.

¹⁵ Individual fields and variables originally contained in the KT VG which were only used to care and maintain the dataset or for data requests without having specific content were removed and not included into the EBDC dataset

same holds for the semi-annual Ifo Investment Survey (IT), although the questions posed aim at the companies' investment behaviour with regard to their focus of production.

Finally, we also plan a linkage of the dataset to other external data such as supervisory board or proprietor structures.

6. Access

The EBDC sees itself as a service provider that supports research projects of professors, visiting researchers and doctoral students by providing, among other things, the EBDC dataset. In general, research projects must be non-commercial, high-level projects in economics that can be empirically analyzed using the EBDC dataset.

Due to the high confidentiality and the obligation to maintain the secrecy of survey results as well as panel member identity, the EBDC dataset can only be used on the premises of the EBDC and is made available with a time lag. We provide a computer without access to the Internet, a printer or other external storage media and which can only be used in the presence of an EBDC staff member. This person will ensure, on completion of the researcher's stay, that the anonymised data do not allow the identification of individual firms and that no inferences can be made regarding the panel composition. Moreover, after this examination has been successfully carried out, he/ she will send the results in a Stata format.

Access to the EBDC dataset can be applied for using a form at the Ifo Website.¹⁶ In addition, a short description of the research project and accompanying information as to scheduling must be submitted. Upon request the EBDC will send by e-mail a test package containing an anonymised EBDC dataset in Stata-format as well as the documentation on the original dataset. The EBDC expressly supports empirical research projects and is thus free of charge. Access to the EBDC dataset only depends on the availability of workplaces.

¹⁶ http://www.cesifo-group.de/portal/page/portal/ifoContent/N/data/EBDC_Container/EBDC_Angebot_Container/EBDC_Vertrag.pdf

7. Bibliography

Becker, S. O. and K. Wohlrabe (2008), Micro Data at the Ifo Institute for Economic Research: The "Ifo Business Survey", Usage and Access. *Journal of Applied Social Science Studies (Schmollers Jahrbuch)*, 128(2), 307-319.

Fellegi, I. P. and A. B. Sunter (1969), "A Theory for Record Linkage", *Journal of the American Statistical Association*, 64, 1183–1210.

Goldrian, G. (2004), *Handbuch der umfragebasierten Konjunkturforschung. Ifo Beiträge zur Wirtschaftsforschung 15*, Ifo Institute for Economic Research, Munich.

Newcombe, H. B., J. M. Kennedy, S. J. Axford and A. P. James (1959), "Automatic Linkage of Vital Records", *Science* 130, 954–59.

Oppenländer, K. H. and G. Poser, eds. (1989), *Handbuch der Ifo-Umfragen: 40 Jahre Unternehmensbefragungen des Ifo-Instituts für Wirtschaftsforschung*, Duncker & Humblot, Munich and Berlin.