Combined Firm Data for Germany (KombiFiD): matching process-generated data and survey data

Hethey, Tanja; Spengler, Anja

Combined Firm Data for Germany (KombiFiD). Matching Process-Generated Data and Survey Data

Tanja Hethey & Anja Spengler [*]

Abstract: *»Kombinierte Firmendaten für Deutschland (KombiFiD) - Verknüpfung von prozess-generierten und Befragungsdaten«*. In Germany, process-generated data and survey data on firms are collected by different data producers. Each data producer provides access for researchers to its data, but the combination of datasets from different producers is not possible at the moment. A new project (KombiFiD) aims to overcome this limitation: firm data collected by the German Statistical Offices, the Deutsche Bundesbank and the Federal Employment Agency will be linked for the first time. The project aims are twofold: to gauge the possibilities of linking selected datasets beyond the limits of individual labour market data producers and to provide a combined dataset to science, thereby creating new research opportunities. This paper describes the project, the selected datasets and explains potential matching problems. In this context we address e.g. the advantages and disadvantages of survey and process-generated data and some challenges we expect within the project.

Keywords: Longitudinal Analysis, Process-Generated Data, Social Bookkeeping Data, Public Administrational Data, Survey Data, Mixed Methods, Company Data, Record Linkage, Data Fusion.

1. Introduction

Theoretical models are often inadequate to answer politico-economic or socio-political questions. In fact, the political practicability of research findings requires empirical fortification. Obviously the precondition for a successful applied economic research is the availability of manifold business micro data (Hauser et al. 1998).

Access to business micro data has been continuously improved within the last few years. Currently, researchers of non-commercial research institutes have various opportunities to analyse micro data provided by the official statistics and process-generated micro data, too. Nevertheless, there is still a great need for development regarding the possibilities of data access to business

[*] Address all communications to: Anja Spengler, Research Data Centre of the Federal Employment Agency in the Institute for Employment Research, Regensburger Str. 104, 90478 Nuremberg, Germany; e-mail: anja.spengler@iab.de.
Tanja Hethey, Research Data Centre of the Federal Employment Agency in the Institute for Employment Research, Regensburger Str. 104, 90478 Nuremberg, Germany;
e-mail: tanja.hethey@iab.de.

micro data. For example, so far it wasn't possible to link firm data of different data producers.

This challenge addresses the project "KombiFiD – Combined Firm Data for Germany" which will be described below. Within that project, survey and process-generated data of different data producers will be linked for the first time in Germany. Afterwards, this unique new data source will be offered to the scientific community.

The KombiFiD-project will be implemented through cooperation between the Federal Statistical Office of Germany, the German Federal Employment Agency, the Deutsche Bundesbank, the Leuphana University of Lueneburg and the University of Applied Sciences Mainz.

In line with the release of the first small version of the KombiFiD dataset we will organise a kickoff workshop to introduce the dataset to the scientific community. This workshop will be accompanied by a call for papers to select the first research projects that will get access to the new KombiFiD data. Access will be provided via on-site use and remote data access at all three data producers. On-site users from abroad will be offered funds for accommodation costs. At the end of the project there will be a user conference to give researchers the possibility to share their knowledge and first experiences with the data and present their research results.

2. Current Data Producers and Data Access

The official statistics are responsible for the data generation and utilisation in Germany. The official statistics are subordinate to the Ministry of the Interior or the state chancellery or are part of municipal or other institutions like the Federal Employment Agency or the Deutsche Bundesbank respectively. The German Federal Employment Agency, the German Statistical Offices of Germany and the Deutsche Bundesbank collect and prepare comprehensive business micro data.

The above-named data producers work fully independent from each other and the data of different data producers could only be linked in the case of identical underlying regulatory frameworks (Kaiser et al. 2007).

All of the mentioned data producers offer access to a variety of datasets which are generated in different ways like survey or administrative data. These datasets cover a huge number of different topics. The Research Data Centres (RDC) of the above-named data producers offer access to business micro data based on applicable law for non-commercial researchers. The access paths are equal at every RDC (Brandt et al. 2007).

Therefore, access to business micro data at the research data centres is given to non-commercial researchers via on-site use, remote data access and scientific use files:

- *Scientific use Files (SUF):* Scientific use files are factually anonymous datasets offered to researchers of scientific institutions for analysis.
- *Remote data access:* Remote data access means that researchers develop programs in SPSS, Stata or SAS on the basis of test data. At the RDC these programs are run with original data. After verification of compliance with data protection legislation, the results are sent to the researcher.
- *On-Site Use:* During a research visit to the RDC researchers may analyze weakly anonymous data autonomously (Kohlmann 2005).

The type of data access is due to the level of detail of the data. The most important aim is the data protection of the underlying units by avoiding disclosure.

The advantage of the current ways of data access is the general possibility for the scientific community to work with process generated data. The possibilities of data access were steadily extended during the last ten years.

However, there is still room for improvement concerning the availability of process-generated data. Currently, researchers only have the possibility to work with the data of one data producer at a time. There is no chance for researchers to combine several datasets of different data producers in order to get a complete coverage of the research subject. The disadvantage of this system is that the research question has to fit exactly to one data producer's data supply. Within the KombiFiD project data of different data producers are to be combined for the first time ever. Thus it is possible to provide the scientific community a combined data set. This dataset will be available at the research data centres of above named data producers. The major advantage is that each researcher can work with the combined data at the location most suitable for him or her.

3. KombiFiD approach

A change of legal foundations now allows for linking business data beyond the limits of the single data producers for the first time in Germany (Konold 2007). An amendment of the Bundesstatistikgesetzes (§13a BstatG) in 2005 provides the opportunity to link information from different economic and environmental statistics (Brandt et. al. 2007). In addition it is allowed by law to link statistical data of single firms on the base of a written agreement of the firms. In such a case the statistical data of firms offered by different data producers are allowed to link within a limited in time project. For that reason selected firms will be asked to give their written agreement for the data linkage.

The KombiFiD project is designed as a feasibility study and its priority objective is to offer a novel dataset to the scientific community including maximum information about the underlying units. Therefore the new data enable e.g. simultaneous analysis of firms and establishments.

Furthermore the data, which are chosen for the linkage, will be adjusted for redundancies. At present German firms are legally obliged to answer statistical questions asked by different data producers several times. In this context, a further ambition of the project is finding and eliminating multiply asked questions in order to reduce respondent burden for firms.

The years 1993 to 2006 are the time period for the linkage. The survey unit is the firm in terms of a legally independent unit as it is defined in European law (see Council Regulation No 696/93, Annex III A). It is possible that a firm may cover several establishments located at different places.

According to the Federal Data Protection Act a written agreement is necessary for those units for which the data should be linked, as mentioned above,. As such a survey isn't realisable for all German firms a sampling procedure including 60.000 firms is planned.[1] The population includes all firms that have declared employment notifications between 1993 and 2006.

The extracted firms will be informed about the project, its objectives, the relevant datasets and the underlying confidentiality regulations. In particular, the firms will be asked, whether they agree with matching their data or not. Within the project no further firm information will be collected.

Following the field work, the data of firms that agree with the procedure will be linked. The starting point for matching the data is the German Business register system. This dataset includes essential identifiers for linking data of different sources.[2] Analyses of consistency and validation follow. These analyses give information about the appropriate type of linking with respect to sensible content outcome. At the start it is planned to offer just selected data to the scientific community within a research visit at the research data centres of the involved institutions. Therefore, very important subtasks of the project are anonymisation and documentation of the resulting data. In addition, the setting up of a commission to legislate the permanent matching of the data is planned within the second project phase (Brandt et al. 2007).

4. Selection Bias

The quality of analyses based on the new KombiFiD dataset is affected by the response rate and for this reason by the unit non-response, the rate of linking and of possible inconsistencies that may exist in the data (Bender et al. 2007).

Within the KombiFiD project, survey and process-generated data will be linked. Beside other advantages, the linking of these data may adjust weak-

[1] For detailed information regarding the single data sets have a look at chapter "selected datasets".

[2] For detailed information regarding the German Business Register have a look at chapter "selected datasets".

nesses of the single types of data and therefore increase the informative value and the validity of research results.

4.1 Advantages of Process-Generated Data

Process-generated data are collected for administrative purposes. Hence the costs for collecting those data are low. In addition, process-generated data are not affected by typical survey related difficulties like unit non-response, socially desired answering or wrong answers because of limited capacity for remembering. Therefore, it is assumed that process-generated data are of higher validity than survey data in terms of special issues. Another advantage of process-generated data is the fact that these data often cover the population. In this case the number of units is very high and detailed analyses are possible (Hartmann et al. 2007).

4.2 Advantages of Survey Data

There are also some remarkable advantages of survey data. For example, it is feasible to collect those characteristics researchers are mostly interested in. In relation, it is possible within surveys to collect data that are not interesting for administrative purposes and therefore not included in process-generated data. The linkage of process-generated and survey data combines the advantages of the single data sources and includes much more information (Hartmann et al. 2007).

4.3 Selection Bias When Matching

As mentioned above, the linkage of data of different data producers is allowed only with a written agreement of the persons concerned. For that reason, it can not be excluded that there is any non random selectivity concerning the answering units.

The agreement of the units – in that case the firms – depends on a variety of factors. In particular cost benefit analysis done by the respondents should be considered. The degree of benefit a company associates with the agreement to match their data is due to the expected level of positive outcome for their personal situation. For that reason the firms get the information that one aim of the project is reducing respondent burden. Also the expected costs have an effect on the agreement. The costs will be negatively appraised if there is any doubt of data confidentiality (Hartmann et al 2007). Therefore the letter for the firms includes the advice that the data will handled as strictly confidential considering data privacy regulations. Moreover, the decision to agree to the data linking also depends on the attitude to the research objective and the institutions behind (Hartmann et al. 2007). Regarding possible non-response difficulties, comprehensive analyses are planned within the KombiFiD project.

A further challenge of the KombiFiD project is the underlying reporting unit in every single dataset. Indeed, the reporting unit in the resulting dataset is the firm, but some of the single datasets are on establishment level. These datasets have to be aggregated on firm level. For those datasets that are on establishment level but cover the whole number of possible units, like the Establishment-History-Panel does, this aggregation isn't very difficult. Much more challenging is the aggregation of datasets which are samples. An example for such a dataset is the IAB-Establishment-Panel. In this case, a complete coverage of all establishments of a firm can't be expected. Therefore the resulting KombiFiD-dataset is limited in terms of data gaps.

After a successful conclusion of the KombiFiD project, the legislation of a permanent matching of the data is a further challenge.

5. Selected datasets

The following paragraphs introduce the selected datasets for the KombiFiD project ordered by their data producers.

5.1 Federal Statistical Office

Offering access to a wide range of different business datasets, the Federal Statistical Office provides the majority of datasets to be linked during the KombiFiD project. Most of those datasets can be classified into three groups:

1) *The German Business register system (Unternehmensregister (URS) 95)* was implemented in the late 1990s. Its implementation was regulated by European law in 1993 (Council Regulation No 2186/93). According to this Council Regulation, each country in the European Union had to implement a register containing all firms contributing to the gross domestic product at market prices. The register contains information about firm name, address, firm size and a set of variables concerning economic activity and financial performance. Beside the unique business register ID, that is used for identification on firm level the register also lists all corresponding establishment numbers and tax numbers for each firm. Using the business register as a master file those identifiers will help us to aggregate and match other KombiFiD datasets that provide information on different reporting levels as for example the more disaggregated establishment level.
2) *The second group is a set of official business surveys.* These include sample as well as full sample surveys. Although this set covers a heterogenous group of business surveys, concerning content and target population two main groups can be identified:

- The first group are the annual *Cost structure surveys (Kostenstrukturerhebungen)*. Those sample surveys comprise information about all kinds of firm costs (labour costs, costs of materials, expenditures for research and development) and therefore serve as an important basis for the national accounting system . The surveys are conducted separately for each economic sector. The sample and reporting unit is always the firm.

- The second group are the *Annual surveys / Annual reports* . Those surveys are full sample surveys providing information about turnover, investment, wages and salaries. Again all surveys are conducted separately for each economic sector using the firm as reporting unit.

3) *The third group consists of tax data.* The KombiFiD project uses the *German Corporation tax statistics (Körperschaftssteuerstatistik)*, *the German trade tax statistics (Gewerbesteuerstatistik)* and the *German turnover tax statistics (Umsatzsteuerstatistik)*. For a brief description of the latter see Vogel/Dittrich (2008). All three statistics report information on the level of tax numbers.

Table 1 on the previous page gives an overview of all above mentioned datasets.[3] At the moment access to all single datasets is provided via the Research Data Centres of the Federal Statistical Office and the statistical offices of the Länder.

Table 1: Datasets of the Federal Statistical Office

	Full sample sample	Reporting unit
German Business register system (URS 95) Unternehmensregister (URS) 95	full	firm
Cost structure surveys:		
Cost structure survey in manufacturing, mining and quarrying Kostenstrukturerhebung im Bereich Verarbeitendes Gewerbe, Bergbau und Gewinnung von Steinen und Erden (KSE-VG)	sample	firm
Cost structure survey in the building industry Kostenstrukturerhebung im Baugewerbe (KSE-Bau)	sample	firm
Annual surveys/reports:		
Annual survey in wholesale and retail trade Jahreserhebung im Handel (sowie in der Instandhaltung und Reparatur von Kfz und Gebrauchsgütern)	sample	firm

[3] For more information about the several datasets see http://www.forschungsdatenzentrum.de (only in German).

Fortsetzung Tabelle 1		
Annual survey incl. survey of investments in the building industry proper and in the finishing trade Jahreserhebung einschl. Investitionserhebung im Bauhaupt- und Ausbaugewerbe	sample	firm
Annual report on enterprises in manufacturing, mining and quarrying Jahresbericht für Unternehmen im Bereich Verarbeitendes Gewerbe, Bergbau und Gewinnung von Steinen und Erden	sample	firm
Other official surveys:		
Monthly report incl. survey of orders received for local units in manufacturing, mining and quarrying Monatsbericht einschl. Auftragseingangserhebung für Betriebe im Bereich Verarbeitendes Gewerbe, Bergbau und Gewinnung von Steinen und Erden	full	establishment
Survey of investments in manufacturing, mining and quarrying Investitionserhebung im Bereich Verarbeitendes Gewerbe, Bergbau und Gewinnung von Steinen und Erden	full	firm
Structure of earnings survey Gehalts- und Lohnstrukturerhebung im Produzierenden Gewerbe und im Dienstleistungsbereich	sample	establishment
Structural survey in the service sector Strukturerhebung im Dienstleistungsbereich	sample	firm
Tax statistics:		
Trade tax statistics Gewerbesteuerstatistik	full	tax number
Corporation tax statistics Körperschaftssteuerstatistik	full	tax number
Turnover tax statistics Umsatzsteuerstatistik	full	tax number

5.2 Federal Employment Agency (BA) and Institute for Employment Research (IAB)

On the one hand the BA offers access to its process-generated data that originate from the notification procedure of the social security system. On the other hand the IAB which is the BA's research institute offers access to survey data. For the KombiFiD project one dataset out of each group has been selected:

1) The first group is represented by the *Establishment-History-Panel (Betriebs-Historik-Panel (BHP)):* The starting point for this dataset is the annual data from the employment notification process. In Germany, every employer has to provide an annual notification of all employees liable to social insurance. Those notifications contain information about each person's age, sex, wage, education and qualification level as well as the identification code of the working-place (establishment number). By using the establishment number this individual data are aggregated on the establishment level. The resulting dataset is a full sample of all establishments with employees subject to social security contribution. It provides information about the establishment's employee structure (age, sex, education and qualification level) as well as the wage structure. Using the unique establishment number the annual data can also be linked over time. For a detailed description of the BHP see Spengler (2008).

2) In the second group of datasets the KombiFiD project uses data from the *IAB Establishment Panel (IAB-Betriebspanel):* This annual panel with about 15.000 establishments in 2006 is conducted by the IAB since 1993. It primary implies information about the labour market's demand side. Additionally the panel consists of annual changing questions related to various topics such as foreign investments (1998), skilled worker demand (2000) or job security (2006). For more information about the IAB Establishment Panel see Kölling (2000) and Fischer et al. (2008). At present access to both datasets is provided by the RDC of the BA in the IAB.[4]

5.3 Deutsche Bundesbank

Both datasets that come from the Deutsche Bundesbank are based on process-generated data:

1) *Microdatabase Direct Investment (Mikrodatenbank Direktinvestitionen (MiDi)):* The Deutsche Bundesbank collects annual firm data on foreign direct investment stocks due to the Foreign Trade and Payment Regulation from 1976. The MiDi dataset was set up using this full sample data up from 1989. Time series for individual firms are only available up from 1996. The dataset covers information about domestic investments abroad as well as information about foreign investments in Germany. For a detailed description see (Lipponer 2003).

2) *Corporate balance sheet statistics (Unternehmensbilanzen):* This dataset contains balance sheet statistics of non-financial enterprises. "The current Bundesbank's corporate balance sheet statistics which was created in the mid sixties are based on the annual (unconsolidated) accounts submitted in

[4] For codebooks of both datasets as well as information about data access please see: http://fdz.iab.de.

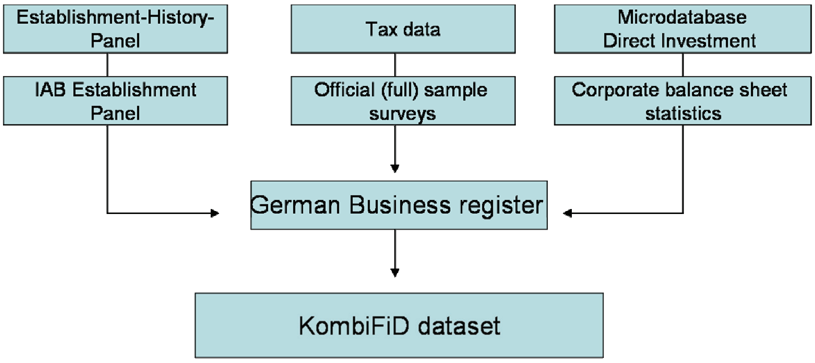
the context of the rediscount business" (Stöß 2003:132). As the importance of this form of financing varies by firm size and economic sector (and time) the dataset mainly covers larger firms from the following economic sectors: manufacturing, construction, wholesale and retail trade. For a brief description see (Stöß 2003). At present access to both datasets is provided via the Research Centre of the Deutsche Bundesbank.[5]

6. The KombiFiD dataset

Figure 1 gives an overview of the above mentioned (groups of) datasets. It highlights the central role of the German business register during the whole matching process.

The strategy is to start the matching using the business register data and two further datasets from the Federal Statistical Office. This will lead to a contemporary release of a first small KombiFiD dataset. After that, all other datasets are matched one by one during the rest of the project time.

Figure 1: Selected datasets grouped by data producers


Establishment-History-Panel
Tax data
Microdatabase Direct Investment
IAB Establishment Panel
Official (full) sample surveys
Corporate balance sheet statistics
German Business register
KombiFiD dataset

Further information about the project, the selected single datasets as well as the project's key dates and deadlines can be found on the project's website: www.kombifid.de

[5] for more information see http://www.bundesbank.de.

References

Bender, Stefan / Wagner, Joachim / Zwick, Markus (2007): KombiFiD – Kombinierte Firmendaten für Deutschland. *Working Paper Series in Economics No. 60, University of Lüneburg*. http://www.uni-lueneburg.de/vwl/papers/.

Brandt, Maurice / Oberschachtsiek, Dirk / Pohl, Ramona (2007): Neue Datenangebote in den Forschungsdatenzentren. Betriebs- und Unternehmensdaten im Längsschnitt. *FDZ-Methodenreport Nr. 7/2007*.

Fischer, Gabriele/ Janik, Florian/ Müller, Dana/ Schmucker, Alexandra (2008): The IAB Establishment Panel – from Sample to Survey to Projection. *FDZ Methodenreport 1/2008*. http://doku.iab.de/fdz/reporte/2008/MR_01-08_en.pdf.

Hartmann, Josef / Krug, Gerhard (2007): Verknüpfung von Befragungs- und Prozessdaten. Selektivität durch fehlende Zustimmung der Befragten? *IAB Discussion Paper No. 13/2007*.

Hauser, Richard / Wagner, Gert G. / Zimmermann, Klaus F. (1998): Erfolgsbedingungen empirischer Wirtschaftsforschung und empirisch gestützter wirtschafts- und sozialpolitischer Beratung. In: *Allgemeines Statistisches Archiv 82*. 369-379

Kaiser, Ulrich / Wagner, Joachim (2007): Neue Möglichkeiten der Nutzung amtlicher Personen- und Firmendaten. *Working Paper Series in Economics No. 48, University of Lüneburg*. http://www.uni-lueneburg.de/vwl/papers/.

Kölling, Arndt (2000): The IAB-Establishment Panel. In: *Schmollers Jahrbuch 120 (2)*. 291-300.

Kohlmann, Annette (2005): The Research Data Centre of the Federal Employment Service in the Institute for Employment Research. In: *Schmollers Jahrbuch 125(3)*. 437-447.

Konold, Michael (2007): New Possibilities for Economic research through Integration of Establishment-level Panel Data of German Official Statistics. In: *Schmollers Jahrbuch 127(2)*. 321-334.

Lipponer, Alexander (2003): Deutsche Bundesbank's FDI micro database. In: *Schmollers Jahrbuch 123 (4)*. 593-600.

Spengler, Anja (2008): The Establishment History Panel. In: *Schmollers Jahrbuch 128 (3)*. :501-509.

Stöß, Elmar (2001): Deutsche Bundesbank's Corporate Balance Sheet Statistics and Areas of Application. In: *Schmollers Jahrbuch 121 (1)*. 131-137.

Vogel, Alexander / Dittrich, Stefan (2008): The German Turnover Tax Statistics Panel. *Working Paper Series in Economics No 92, University of Lüneburg*. http://www.uni-lueneburg.de/vwl/papers/.