2011 European Establishment Statistics Workshop – EESW11
(12-14 September 2011, Neuchâtel, Switzerland)

It is always good to have just one number, but it is better to have more than one way to get it – Results from the project „Combined Firm Data for Germany“

Stefan Bender, Anja Gruhl

In Germany a lot of business micro data exist. Most of these data are collected by various data producers who use often different methods of data collection. Some characteristics, like *German federal state (Bundesland)* or *industry*, are included in more than one dataset. However, adjustments of the data concerning those similar characteristics do not exist so far. Improving the coordination of data collection or even generating one single dataset thus bears great potential in terms of data quality and cost savings. Against this background, the project “Combined Firm Data for Germany (KombiFiD)” is a milestone of harmonizing German business micro data.

Within the project we have linked survey and process-generated data of different data producers (Federal Statistical Office (FSO), Federal Employment Agency (BA), Deutsche Bundesbank) in Germany for the first time. At the same time we have also linked different datasets collected by one data producer (FSO) for the first time in Germany.[1]

The KombiFiD project is designed as a feasibility study and its main objective is to offer a novel dataset to the scientific community including the maximum possible information about firms. Furthermore, the data which we have chosen for the linkage will be adjusted for redundancies. Another target of our project is to find and eliminate multiply asked questions in order to reduce respondent burden for firms.

Our paper continues as follows: Section 1 explains the KombiFiD dataset. In Section 2 and 3 we describe our methodology and results. In Section 4 we discuss our findings with regard to the feasibility in the daily practice of the National Statistical Institutes and their separate ways of data collection.

1 Some basic facts about KombiFiD

The survey entity in KombiFiD is the firm in terms of a legally independent unit as it is defined by European law (see Council Regulation No 696/93, Annex IIIa). It is possible that a firm covers several establishments located at different places. Because of the German Federal Data Protection Act a written agreement by the firms is mandatory to link the firm information offered by different data producers.

Within the project we have taken a sample of 54,960 firms, which were asked to give us the permission for linking the data. We have selected firms, which are included in large and relevant surveys of the FSO to make sure that we will have rich combined data set for further analysis. About 16,571 firms gave their written agreement to the linkage, and data from the years 2003 to 2006 was linked.

The selected datasets for the KombiFiD project cover a huge number of different aspects. Most data of the KombiFiD dataset originally was generated by the FSO. Table 1 presents these datasets, the reporting units and the reporting path. Within this context, locally is defined as data which are collected by the Statistical Offices of the Länder (Statistische Ämter

[1] The KombiFiD project is implemented through cooperation between the Federal Statistical Office, the German Federal Employment Agency, the Deutsche Bundesbank, the Leuphana University of Luenburg and the University of Applied Science Mainz. The project has been supported by the Federal Ministry for Education and Research (BMBF).

der Länder) and afterwards reported to the FSO. Centrally means the data are collected by the FSO itself.[2]

Table 1: Datasets generated by FSO and included in KombiFiD

Dataset	**Full sample/sample**	**Reporting unit**	**Reporting path**
German Business register system (URS95)	full	firm	locally
Cost structure surveys:			
Cost structure survey in manufacturing, mining and quarrying	sample	firm	locally
Cost structure survey in the building industry	sample	firm	centrally
Annual surveys/reports:			
Annual survey in wholesale and retail trade	sample	firm	partly locally partly centrally
Annual survey incl. survey of investments in the building industry proper and in the finishing trade	sample	firm	locally
Annual report on enterprises in manufacturing, mining and quarrying	sample	firm	locally
Other official surveys:			
Monthly report incl. survey of orders received for local units in manufacturing, mining and quarrying	full	establishment	locally
Survey of investments in manufacturing, mining and quarrying	full	firm	locally
Structure of earnings survey	sample	establishment	locally
Structural survey in the services sector	sample	firm	locally

The Turnover tax statistic is a special case. The survey frame is a collection of data of the advance return for tax on sales/purchases *(Umsatzsteuervoranmeldung) (Vogel, Dittrich, 2009)*. These data are compiled by data centres of Land revenue authorities for every tax payer and afterwards reported to the Statistical Offices of the Länder and the Federal Statistical Office.

The *German Business register system* (URS) contains information about firm names, addresses and the unique business register IDs, all corresponding establishment numbers and tax numbers for all firms *(Statistisches Bundesamt, 2011)*. We use the business register as a master file to aggregate and match all of the different datasets by unique firm identifiers.

For the KombiFiD project we have selected the *Establishment-History-Panel (BHP)* which originates from the Institute for Employment Research (IAB) of the Federal Employment Agency (BA). This dataset is the total population of all establishments with employees subject to social security contributions. In Germany, every employer has to provide an annual notification of all employees liable to social insurance. Beside personal information those notifications contain the identification code of the working-place (establishment number). By

[2] Inside the Statistical Offices of the Länder and the FSO different departments compile the datasets. Within further research we have to investigate more about the data preparation that happens in these departments.

using the establishment number this individual data are aggregated on the establishment level. Using the identifiers included in the URS we aggregated the establishments on the firm level *(Hethey, Spengler, 2009)*.[3]

Table 2 presents the KombiFiD datasets and the corresponding number of firms that report to these datasets.[4]

Table 2: Number of firms included in the KombiFiD dataset, ordered by the original data sources

Dataset	Number of firms	Percentage
BHP	54,510	83.6 %
Turnover tax statistics	50,020	76.7 %
Monthly report incl. survey of orders received for local units in manufacturing, mining and quarrying	23,019	35.3 %
Cost structure survey in manufacturing, mining and quarrying	22,796	35.0 %
Annual report in manufacturing, mining and quarrying	22,680	34.8 %
Survey of investment in manufacturing, mining and quarrying	22,543	34.6 %
Structural survey in services sector	19,255	29.5 %
Annual survey in wholesale and retail trade	15,917	24.4 %
Annual survey incl. survey of investments in the building industry and in the finishing trade	4,780	7.3 %
Cost structure survey in building industry	3,436	5.3 %
Structure of earnings survey	2,816	4.3 %
Number of firms in the KombiFiD dataset	**65,231**	**100 %**

2. Selection of data for the analysis

Within this paper we focus on firms with just one establishment included to have a "simple" structure for our comparisons. Due to the fact that *one-establishment-firms (OEF)* represent the majority of all firms in the KombiFiD dataset, we have 49,613 observable units for the years 2003 - 2006 in a pooled dataset for our analysis. The vast majority – about 12,000 OEFs – can be observed over the four years of interest.

Not all firms are included in every single dataset integrated in the KombiFiD dataset. Some original data contain significantly more units than others do. This is a result of the above mentioned different methods of data collection that coincide with partly different underlying sampling frames within the original data. To give a sense of the numbers of observation *table 3* presents the datasets and the corresponding number of OEFs that report to these KombiFiD datasets for the period 2003-2006.

Table 3: Number of OEFs included in the KombiFiD dataset, ordered by the original data sources

Dataset	Number of OEF	Percentage
BHP	48,515	97.8%
Turnover tax statistics	39,000	78.6%
Cost structure survey in manufacturing, mining and quarrying	17.558	35.4%
Monthly report incl. survey of orders received for local units in manufacturing, mining and quarrying	17,460	35.2%

[3] The Microdatabase Direct Investment (Mikrodatenbank Direktinvestitionen) and the Corporate balance sheet statistics (Unternehmensbilanzen) are originally generated by the Deutsche Bundesbank. We integrate these datasets into the KombiFiD dataset later.

[4] The table contains only firms that gave their agreement to the linkage. This applies to all of the following tables and analysis.

Annual report in manufacturing, mining and quarrying	17,378	35.0%
Survey of investment in manufacturing, mining and quarrying	17,269	34.8%
Structural survey in services sector	15,590	31.4%
Annual survey in wholesale and retail trade	10,787	21.8%
Annual survey incl. survey of investments in the building industry and in the finishing trade	4,032	8.1%
Cost structure survey in building industry	2,776	5.6%
Structure of earnings survey	1,625	3.3%
Number of all OEF in the KombiFiD dataset	**49.613**	**100 %**

The combination of firms in the KombiFiD dataset may vary during the period of observation. The basic cause of this variation is the thematic focus of the single datasets on the one hand and the sampling frame on the other hand. Information on the BHP, for instance, is only available for a year when a firm holds employees subject to social security contribution.
In addition, some of the original data are strictly related to one specific industry, others report cross-industry information. Moreover, the BA and the FSO use different definitions of the establishments' industry. The Federal Employment Agency definition of industry is based on the number of employees. In contrast, the Federal Statistical Office uses the main business activities of the firm to define the branch of industry (EU definition). *Table 4* presents the two most frequently observed dataset combinations over the observation period as well as the corresponding number of units.

Table 4: Most frequently observed dataset combinations within the KombiFiD project

	Dataset combination	**Number of units**
1	BHP/ Structural survey in services sector/ Turnover tax statistics/	12,254
2	BHP/ Monthly report incl. survey of orders received for local units in manufacturing, mining, quarrying/ Turnover tax statistics/ Cost structure survey in manufacturing, mining, quarrying/ Annual report on enterprises in manufacturing, mining, quarrying/ Survey of investments in manufacturing, mining, quarrying	11,468

While the first row represents OEF in the service sector, the second row includes OEFs in the production sector. We therefore call the different dataset combinations the *"service sample"* or the *"production sample"*. The combinations include the most frequent original data within the KombiFiD project. The samples are by definition disjunctive.

3 Method and Findings

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 this paper we focus on the latter which implies two dimensions. First, it needs to be examined whether the correct units have been linked. If not, considerable mistakes may be the consequence for the analysis of the data. Secondly – this is our main objective – we have a look at variables that originally appeared in more than one of the linked datasets but seem to have the same content (German federal state (Bundesland) and industry). For our comparison of the industry we use the Classification of Economic Activities 2003, 2-digit code *(Statistisches Bundesamt, 2003)*.

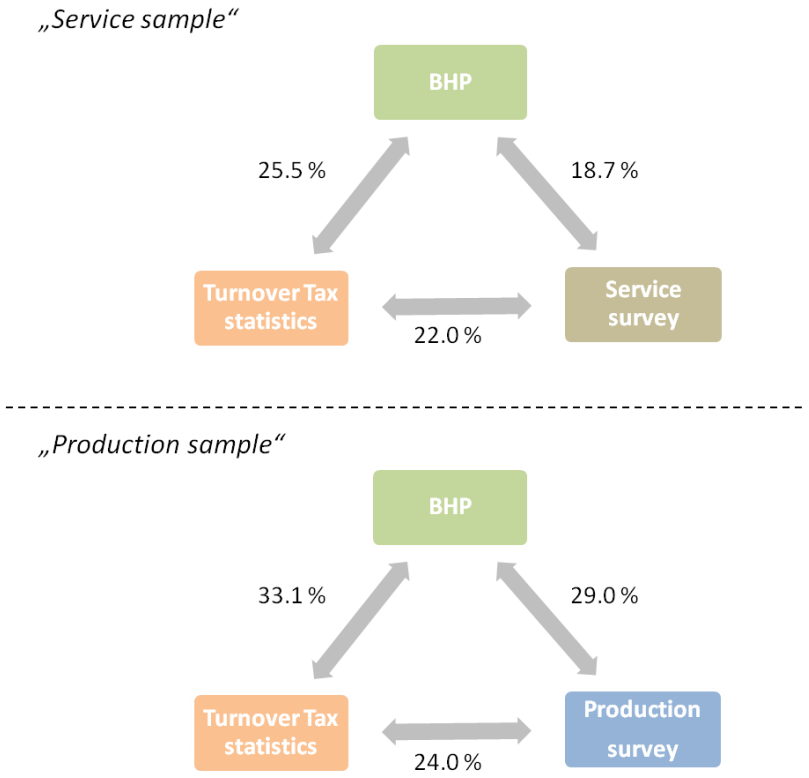
The *"production sample"* includes the BHP and the Turnover tax statistics and four surveys of the German Statistical Institutes of the Länder. Within these four surveys we have found nearly no deviations of the location of the OEF, measured by the German federal state

(0.07 % deviation) or industry (0.00 %-0.17 % deviation).[5] Thus for simplification reasons, we decided to integrate only one out of the four surveys into our comparison. We selected the "cost structure survey in manufacturing, mining and quarrying" because of the highest number of OEFs.

Exploring the variable location (German federal state), we have found a very high consistency in the two samples (98 % to 99 %). The result indicates that we have linked the identical OEFs. Moreover, the definitions of this variable seem to be consistent amongst the different original datasets.

In contrast we have explored remarkable deviations while comparing industry. *Figure 1* outlines the deviations of industry concerning the service sample and the production sample.

Figure 1: Deviations of industry in the two KombiFiD samples



Both samples show its maximum deviation between BHP and Turnover tax statistics (figure 1). Furthermore the differences of industry between the datasets of the production sample exceed the corresponding deviation in the service sample. As stated above different definitions of industry in the datasets of the BA and the FSO exist. Therefore little deviations would not be surprising. Nevertheless, the deviations we found are pronounced and thus not only explainable by the different definitions of industry. Particularly the using of the 2-digit industry classification should result in smaller deviations. Moreover the "trend" of the deviations is changing between the samples. While we explored the minimum deviation in the service sample between BHP and service survey, the production sample shows its minimum deviation of industry between Turnover tax statistics and production survey. The reason for that has to be subject of further research projects.

[5] We suppose no data editing or adjustments between the departments of the FSO and the Statistical Offices of the Länder exist. Nevertheless, we have to investigate this within further research.

3 Conclusions

To sum up, the KombiFiD project shows that it is technically possible to link firm data of different data producers with the help of the identifiers integrated in the *German Business register system*. In our analyses we have found a huge number of firms being included in more than one dataset. We have explored differences and similarities in the location of the firm (German Federal State) and industry of *one-establishment-firms* between those linked datasets. As expected, we have found the best consistency of data given by the combination of one data producer, identical methods of data collection and equal definitions of variables.

We have explored the location of the firm (German federal state) and industry within two samples, one in the service sector and the other in the production sector. We found a very high consistency concerning the location of the firm. In contrast, we have found remarkable deviations of the industry. The reasons for these deviations are manifold. On the one hand different data generation processes have to be taken into consideration. On the other hand – depending on the data producers - different definitions of industry are included in the datasets. As a consequence of that the same OEFs may be allocated to different industries in the datasets of the KombiFiD project.

Within this paper we have analyzed the simplest organizational form of a firm with just one establishment per firm. So, we expect higher deviations for firms with more than one workplace.

Furthermore, we need to know more about the data generation processes. Especially the data preparation and editing in the departments of the Federal Statistical Offices are of special interest because at the moment we do not know enough about data cleansing processes and the possible effects on the deviations of variables in different datasets.

From our point of view, one of the main challenges in the context of firm surveys is to find a way to deal with the data inconsistencies we have found. The standardization of variable definitions which are included in different datasets seems to be a promising approach. Additionally, close coordination of data producers and adjustments concerning methods of data collection can improve data consistency and therefore data quality.

A long-term objective could be to build up a central data collection, where a few relevant variables are surveyed and stored. These variables should be available for every survey conducted by researchers. As the result every survey on firms will include the same information concerning relevant variables, like industry or location. To generate these variables the relevant data producers, institutions and researchers should develop necessary standards.

Micro data of firms are subject to the German Federal data protection act concerning data collection, linking and handling. Within our paper we did not discuss these aspects. But data protection is of high relevance regarding implementation of our recommendations in daily practice. The advantages of centralization are obvious given the time and money savings to be expected as well as improvements on data quality. Moreover, it would reduce respondent burden for firms.

Literature

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

Hethey, Tanja/ Spengler, Anja (2009): Combined firm data for Germany (KombiFiD). Matching process generated data and survey data. In: Historical Social Research, Vol. 34, No. 3, p. 204-214.

Konold, Michael (2007): New Possibilities for Economic Research through Integration of Establishment-level Panel Data of German Official Statistics. In: Schmollers Jahrbuch/ Journal of Applied Social Science, 127 (2), p. 321-334.

Lenz, Rainer/ Zwick, Markus (2009): Business Microdata in Germany: Linkage and Anonymisation. In: Schmollers Jahrbuch/ Journal of Applied Social Science, 129 (4), p. 645-653.

Spengler, Anja (2010): Verknüpfung und Abgleiche von Unternehmensregisterdaten des Statistischen Bundesamtes mit Betriebsdaten des Instituts für Arbeitsmarkt- und Berufsforschung. FDZ-Methodenreport 1/2010, Nürnberg.

Statistisches Bundesamt (2003): German Classification of Economic Activities, Edition 2003 (WZ 2003).http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/DE/Content/Klassifikationen/GueterWirtschaftklassifikationen/klassifikationwz2003englisch,property=file.pdf, download 30.08.2011.

Statistisches Bundesamt (2011): Unternehmensregistersystem 95. Qualitätsbericht.Stand März 2011, http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/DE/Content/Publikationen/Qualitaetsberichte/UnternehmenGewerbeInsolvenzen/Unternehmensregister,property=file.pdf, download 10.08.2011.

Hethey-Meier, Tanja/ Seth, Stefan (2010): The Establishment History Panel (BHP) 1975-2008. FDZ-Datenreport 4/2010, Nürnberg.

Vogel, Alexander/ Dittrich, Stefan (2008): The German Turnover Tax Statistics Panel. In: Schmollers Jahrbuch/ Journal of Applied Social Science, 128 (4), p. 661-670.