WP. 5
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Tarragona, Spain, 26-28 October 2011)

Topic (i): Disclosure risk assessment

Why should official statistics care about data integration? First experiences in linking microdata in Germany

Prepared by Christopher Gürke, Federal Statistical Office, Research Data Centre, Germany

Why should official statistics care about data integration? First experiences in linking microdata in Germany

Christopher Gürke[*]

[*] Federal Statistical Office of Germany, Research Data Centre, e-mail: christopher.guerke@destatis.de

Abstract: In many European countries efforts are being made by official statistics to integrate data from different sources. Sometimes such efforts aim as a combination of register data and survey data. In other cases the objective might be a linkage of survey data and process generated data or an integration of different registers. Three major reasons can be given for such activities: First of all, the matching of existing data sets occasionally constitutes a faster and cheaper way to gather certain information than the collection by means of a survey. In the second place, it can also be a possibility to reduce the reporting duties of respondents. Finally, the integration of data from different sources often enables novel and more comprehensive economic analyses. The paper will focuses on the technical, methodological and legal challenges that arise in using data integration methods and focuses on the perspective for data producers and researchers.

.

1 Introduction

For many economic analyses longitudinal microdata on the level of enterprises is needed. In order to answer certain research questions it is also often necessary to integrate enterprise-level data from different sources. In Germany, the access of the scientific community to such integrated longitudinal micro-data has improved considerably during the last ten years, due to the work of the research data centres of the Federal Statistical Office and the statistical offices of the Länder, the research data centre of the Federal Employment Agency and the research centre of the Deutsche Bundesbank. One important step that has not been taken yet is an integration of enterprise-level data across the borders of different data producers. There are usually some methodological problems to solve though: unique identifiers – for instance – may not be available or relevant units may not be included in all data sources.

A German project directly located in this context is the project "Combined Firm Data for Germany" (KombiFiD – Kombinierte Firmendaten für Deutschland). It is carried out by the Federal Statistical Office, the Institute for Employment Research of the Federal Employment Agency, the Deutsche Bundesbank and the Leuphana University of Lüneburg. The project is funded by the Federal Ministry of Education and Research (BMBF). One of the features of KombiFiD is the fact that official business data are collected independently by several institutions (Statistical Offices, Federal Employment Agency, German Federal Bank). The project aims to find solutions in reducing the response burden for enterprises by merging the cross-institutional micro data and looking for overlapping information.

However due to the current legal situation cross-institutional merging of micro data cannot be carried out without enterprises giving their consent. Therefore the second important feature is to change the legal situation in Germany to allow cross-institutional record linkages.

2 Legal situation and the consequences

In the past business micro data were offered only as longitudinal micro data on the level of enterprises for scientific analyses. Due to the growing number of enquiries for temporal cross-sectional data, the research data centre of the Statistical Offices of the Federal States launches the project „Amtliche Firmendaten für Deutschland“ (AFiD). The function of the project AFiD is to merge official business micro data over time and different fields of official statistical business micro data. The underlying idea of the panel data is to advance the analytical potential and to offer information about enterprises and their local units with respect to different official statistics, different times and different thematically issues. Due to the German Statistical Law (§13a) different business datasets can only be matched if the data producer is the same.

This legal issue limits the possibilities with regard to the integration of enterprise data, collected by different public institutions like the statistical offices, the Federal Employment Agency, and the Deutsche Bundesbank. All enterprises for which a cross-border-merging of data were carried out had to give their approval to this procedure. Therefore 54.000 enterprises were asked in a survey to give their approval for linking their data of the years 2003 until 2006.

The result of the KombiFiD-Survey was that nearly 57% of the addressed enterprises answered the question and nearly 31% gave their approval for linking their data across the borders of data producers.

Due to the legal situation the consequence for the dataset is a drop-out process that results from the fact that some enterprises had not answer or did not want to give their consent and the fact that enterprises which do not exist any more (at least not in their former legal shape) would have to assent. This – for obvious reasons – cannot be achieved. To simplify matters, both cases will in the following be referred to as cases of nonresponse. There will be also a third drop-out process – also referred to as a case of nonresponse. This process will occur in the context of record linkage and will result from the fact that there will almost certainly be at least some cases where the data of an enterprise cannot be merged, due to uncertainty issues.

All this nonresponse is not a problem as long as it does not occur in a systematic way. If there is no connection between the properties of enterprises that increase the drop-out-probability and those variables that are of interest in research context, there is nothing to worry about (Provided, the overall amount of non-response does not exceed a certain level). Whether or not this is the case, is an empirical question. Currently, it is only possible to present some preliminary results concerning those enterprises, which cannot be asked for approval any more (due to closure, split off or similar reasons).

According to analyses conducted in the Research Data Centre (FDZ) of the German Federal Statistical Office, 8,77% of the enterprises in manufacturing existing in 2003 closed until 2008. It is important to point out that there is no (significant) correlation between size class and mortality of enterprises in construction industry between 2003 and 2008: According to our analysis, mortality only varies slightly between 8,35% for enterprises occupying 20-49 employees and 9,19% in case of major enterprises with 500 and more employees. In contrast, particular sectors of economic activity within manufacturing are characterized by a higher rate of companies closed between 2003 and 2008 than other branches of industry: 14,51% of the companies in paper manufacturing were shut down in this timeframe, enterprises in production of data and sound carriers (13,42%) and coking plants (11,64%) are also affected by a relatively high percentage of mortality. Sectors with a more steady number of enterprises are, among others, located in production of beverages (5,26%) and clothing industry (5,03%).

All in all, study of the Cost Structure Survey in Manufacturing, Mining and Quarrying suggests a correlation between the economic sector of an enterprise and the likelihood of closure, whereas the size of enterprise has only a very low effect on their mortality. Analyses based on the Annual Survey in Wholesale reveal that there is also a low correlation between size of the enterprise and probability of shut-down (or split off or takeover) in this economic sector, varying from 6,30% (20-49 employees) to 8,52% (50-99 employees). Contrary to these results major enterprises in the construction industry had to close more often than small firms in this sector of economic activity. Especially enterprises with more than 500 employees had to shut down in the period from 2003 to 2008 (17,33%), in comparison to 12,88% of the companies with 20-49 employees.

In summary, the aforementioned preliminary results indicate that, ahead of the start of the KombiFiD-survey, a certain level of unit-nonresponse due to the closure of enterprises had to be taken into account. In addition to the nonresponse resulting from the mortality of enterprises the fact that enterprises asked in the postal survey were assumed to reject the merging of their respective data which led to the second cause of nonresponse in the KombiFiD-Survey.

Finally, a third drop-out process considered in the project is that an enterprise can be found in the Cost Structure Survey in Manufacturing, but can not be detected in the Foreign Direct Investment Stock Statistics of the Deutsche Bundesbank. In this case two explanations come into question: Either enterprise A has not invested abroad and therefore has no information in the corresponding statistic, or enterprise A did in fact realize foreign direct investments, but the appropriate data can not be matched, for example because of errors in the labelling of the company’s name. This situation would lead to item-nonresponse regarding variables for enterprises in the Foreign Direct Investment Stock Statistics.

These consequences arising with the legal situation intends to change the legal situation in Germany to allow cross-institutional record linkages without carrying out a survey. Therefore the project commissioned a feasibility study that discusses possibilities in this context.

The following sections describe the technical challenges arising in the context of record data linkage.

3 The combined dataset

The produced dataset within the project KombiFiD consists of different datasets collected by different data producers.

The Federal Statistical Office of Germany and the Statistical Offices of the States offer 12 different micro data sets which can be grouped in three categories. The first one is the German Business Register which can be seen as a master file during the linkage process. The German Business Register contains information of the whole population of enterprises in Germany. This implies also unique identifiers like the business register ID and the unique establishment number assigned by the Federal Employment Agency that allow matching data directly of the Statistical Offices of Germany and the Länder and the Federal Employment Agency (BA) and the Institute for employment research (IAB). 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. This information is useful for aggregating business surveys which contain information on a disaggregated report level e.g. local unit level.

The second group of datasets can be subsumed as business surveys on industrial sector including the mining industry, wholesale and retail trade and service. All micro datasets contain the business register ID which makes it easy to link the different survey with each other. Nevertheless some of the surveys reports on the level of local units and therefore have to been conditioned such that the report level is the same in all linked micro datasets. The 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 tax statistics can be covered as the third group of datasets. The tax numbers identify enterprises such that a one to one linkage is easy to perform.

The Federal Employment Agency (BA) and the Institute for employment research (IAB) offer two micro datasets, the Establishment-History-Panel (BHP) and the IAB Establishment Panel. The BHP is a process generated dataset that come from the notification procedure of the social security system. The BHP is a full sample of all establishments with employees subject to social security contribution. Using the unique establishment number the data can be linked one to one with the corresponded record in the German business register.

The IAB Establishment Panel was adopted in 1993 and offer information about labour market's demand side. It also contains a unique establishment number. Therefore the linking procedure is also easy to perform.

The datasets offered by the Deutsche Bundesbank are process generated micro data. The Microdatabase Direct Investment covers annual enterprise data on foreign direct investments in Germany and direct investments abroad.

The second dataset, the corporate balance sheet statistic, contains information on balance sheet statistic of non-financial enterprises from the economic sectors manufacturing, construction, wholesale and retail trade.

Due to the absence of unique identifiers like the business register ID and the unique establishment number the record linkage process of the data offered by the Deutsche Bundesbank is more difficult. The challenges are described in section 3.

All datasets described above are provided by the Research Data Centres of the different data producers.

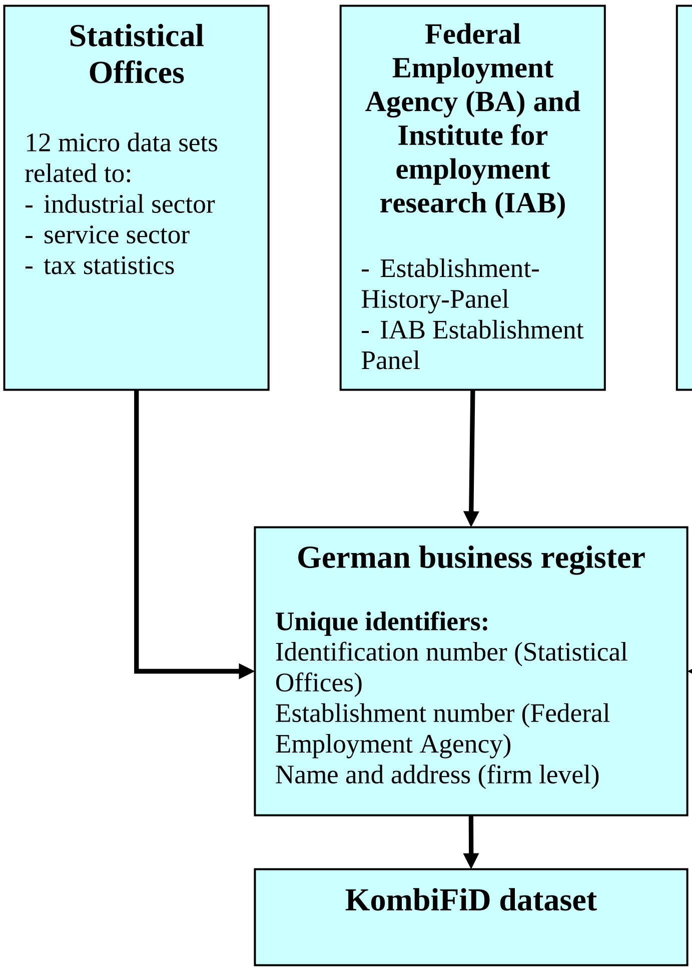
Statistical Offices
12 micro data sets related to:
- industrial sector
- service sector
- tax statistics
Federal Employment Agency (BA) and Institute for employment research (IAB)
- Establishment-History-Panel
- IAB Establishment Panel
Deutsche Bundesbank
- Microdatabase Direct Investment
- Corporate balance sheet statistics
German business register
Unique identifiers:
Identification number (Statistical Offices)
Establishment number (Federal Employment Agency)
Name and address (firm level)
KombiFiD dataset

Fig 1 Linked datasets grouped by data producers

Figure 1 highlights the role of the German business register in the context of linking micro data in the project.

4 Challenges when using record linkage

Linking enterprise data of the Federal Statistical Office and the statistical offices of the Länder, the Institute for Employment Research (IAB) and the Deutsche Bundesbank can only be realised if a unique identifier exists in all data sets. Otherwise linking has to be done by using address information of each enterprise. Such a record linkage is complex and time consuming. The effort that has to be put into this endeavour is reduced by the fact though, that data matching without unique identifiers has been an area of intensive research over the past decades. A lot of theoretical work with respect to statistical matching has been carried out. Therefore, knowledge about typical problems, necessary decisions and good or best solutions is available for many cases.

The project KombiFiD benefits from the situation that a sufficiently broad range of overlapping variables exists. Useable are at least the name of the company, the place where the company headquarter is located and detailed information about the economic branch. In some cases it will also be possible to make use of data about the number of employees and the legal form. Nevertheless, some problems have to be fixed before using probabilistic record linkage methods. The main challenges are that the time reference has not necessarily to be identical and that there is often not only one fixed form of the full name of a company. The last one is an important issue in the context of matching different data sources. The main question is: Can one consider that two records constitute a correct match even so the concordance of the two name strings is at least on a medium level. Another important decision has to been made in this context is about the string comparison function used for linking the records. What function will perform best under the specific circumstances in the KombiFiD project is not easy to predict. A best solution solely on the basis of theoretical considerations is even impossible. Therefore a process of implementation, evaluation and re-evaluation is likely to be necessary. It is certain that decision rules that are based on exact matching would therefore produce a high number of type two errors (false non-matches). Therefore the chosen string comparison function has to be calibrated such that the trade off between correct matches and false non-matches is optimized.

In this context the KombiFiD project has also to deal with the problem of skewness of the linked business data. One or two big companies with unsuccessful record linkage can substantially reduce the value of the final data. It is therefore necessary to subject the top level size class of companies to special scrutiny.

The results of the record linkage process can be found in Hethey and Spengler 2009.

5 Data access

Since May 2011 the first version of the KombiFiD project file is available. The data set contains the described statistics of the statistical offices and the Federal Employment Agency (BA) and Institute for employment research (IAB). Access is provided via safe-centre and remote execution. The data can be used by domestic and foreign researcher.

6 Conclusion

The project KombiFiD shows the possibilities for linking micro data across the borders of data producers. Despite the challenges due to the legal and technical restrictions the project demonstrates that record linkage is feasible. The combination of different data sources offers new possibilities in research activities and therefore constitutes an important step in the development of data infrastructure.

From the data producer perspective the matching of data sets occasionally constitutes a faster and cheaper way to gather certain information than the collection by means of a survey. Nevertheless harmonising the assignment of classifications within the production process and awareness of the quality of classifications have to be taken into account. If the harmonisation can be realised the respondent benefits also in such way that the response burden can be reduced.

Further information about the project, the offered statistics and the data access can be found on the project's website www.kombifid.de.

References

Bender, S.; Wagner, J.; Zwick, M. (2007) *KombiFiD - Kombinierte Firmendaten für Deutschland*, Research Data Centres of the Federal Statistical Office and the statistical offices of the Länder, Working Paper No. 21 (also published as: University of Lüneburg, Working Paper in Economics Nr. 60 and Methodenreport 05/2007 of the Research Data Centre of the Federal Employment Service)

Braakmann, Nils (2010): New data for the analysis of fundamental change: Combined firm data for Germany. Will be published in: Tagungsband 6. CREPS Konferenz.

Spengler, Anja (2010): Verknüpfung und Abgleiche von Unternehmensregisterdaten des Statistischen Bundesamtes mit Betriebsdaten des Instituts für Arbeitsmarkt- und Berufsforschung. In FDZ Methodenreport Nr. 1/2010 des FDZ der Bundesagentur für Arbeit im IAB

Gruhl, Anja; Guerke, Christopher; Hethey-Maier, Tanja; Oberschachtsiek, Dirk; Seitz, Julia (2011): Combined firm data for Germany - (KombiFiD) * Handbook Version 1.0. In: FDZ Datenreport, 02/2011, Nürnberg, 68 S. (only in german)

Gürke, Christopher; Gruhl, Anja; Hethey-Maier, Tanja (2011): Matching enterprise data across different data producers – The project combined firm data for Germany“. In Wirtschaft und Statistik, Feb. 2011, 91-97. (only in german)

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, 204-214.