

Matching Business Data from Different Sources: The Case of the KombiFiD-Project in Germany

MICHAEL KONOLD, SANDRO L'ASSAINATO

RESEARCH DATA CENTRE OF THE GERMAN FEDERAL STATISTICAL OFFICE

Michael.Konold@destatis.de

Sandro.Lassainato@destatis.de

Abstract

This paper deals with two methodological challenges occurring in a German project, which performs the task of merging enterprise data across the borders of different data producers. The first challenge arises from the fact that a part of the data integration has to be carried out without unique identifiers. Therefore, a suitable form of probabilistic record linkage has to be implemented. The second challenge relates to the fact that certain forms of nonresponse occur. The paper highlights the questions the project faces and discusses possible solutions to some of the issues.

Keywords: integration of enterprise data, record linkage, nonresponse

1. Introduction

For many economic analyses longitudinal micro-data 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. This is where the project “Integrated Enterprise Data for Germany” (KombiFiD) comes into play: The purpose of this project is to merge enterprise data of the German Federal Statistical Office, the Land Statistical Offices, the Institute for Employment Research (IAB) and the Deutsche Bundesbank. Besides, a modification of the present legal requirements is aspired in the long run, in order to facilitate the combination of data sets in the future. Thus, the reporting duties for enterprises due to official surveys and statistics could be reduced. KombiFiD is a joint project of the aforementioned institutions, the Institute of Economics of the Leuphana University Lüneburg and the University of Applied Sciences Mainz.

Through the process of data integration carried out in the KombiFiD project new possibilities for economic research will be opened up and it will be possible to analyse economic processes in a more detailed and comprehensive way. However, there will also be two challenges: The first one relates to the process of record linkage in the data integration process. The second one arises from the fact that due to legal circumstances, the project will also have to deal with certain cases of nonresponse. In the following, both issues will be highlighted and discussed.

2. Record linkage: questions and issues

When the enterprise data of the Federal Statistical Office and the statistical offices of the Länder, the Institute for Employment Research (IAB) and the Deutsche Bundesbank will be matched, a unique identifier will be available in only one case. Part of the data link-up has therefore to be realised on the basis of names, address information and variables which are available in all data sets, like the economic branch. 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 has been carried out.¹ One can also draw conclusions from the results of numerous empirical matching projects. Therefore, knowledge about typical problems, necessary decisions and good or best solutions is available for many cases. A very useful overview on the state of the art can be found in the reports of the ESSnet Statistical Methodology Project on Integration of Survey and Administrative Data (CEBEX-ISAD 2008a, CENEX-ISAD 2008b) and in a recent book by Thomas Herzog, Fritz Scheuren and William Winkler (Herzog et al. 2007). There are also several very useful computer programmes that have been developed and which can be used for a wide range of purposes. The Link Plus software of the CDC (U.S. Centers of Disease Control and Prevention), the Big-Match software of the U.S. Bureau of the Census, an application named RELAIS by IStat Italy and the Merge Toolbox by Rainer Schnell and others are four examples.² The following discussion will not highlight the features of different computer programmes, nor will it introduce the basic concepts and methods of record linkage. Instead, the focus will be on the basic conditions of the matching project in hand, the questions which have to be answered in this context and approaches that are likely to prove successful thereby. Only the setting in which there are no unique identifiers will be examined.

The initial situation in the KombiFiD project is favourable insofar as there is a sufficiently broad range of overlapping variables. What's more, the variables are of high quality in all data sets. 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 (Rechtsform).

As for the mapping of units, the only method that comes into consideration is a form of probabilistic record linkage. The reason for that is that variables like the economic branch offer some leeway in the process of categorisation, 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. Decision rules that are based on exact matching would therefore produce a high number of type two errors (false non-matches). The challenge in the KombiFiD project will be to optimize the threshold values, above which a possible match will be assumed and to optimize the threshold values above which it will be assumed that a pair definitely constitutes a correct match. A question is going to be, how to weight false non-matches in comparison to false matches and under which circumstances it is better to accept a certain amount of cases of the one or the other. Currently, considerations go into the direction of a "conservative" approach, that is to say, to rather match not in case of doubt.

Another important point is the significance of the company's name. The name is of great importance because one would not want to consider two records to constitute a match if there

¹ Seminal articles have been published early on. One classic proposal dates back to the 1960s (Fellegi & Sunter 1969).

² The popular SAS application „The Link King“(www.the-link-king.com), developed by Kevin Campbell, is explicitly designed for the matching of person level data. For the KombiFiD project it is therefore no option.

was not at least a medium level of concordance between the two name strings. As it is also true that string comparisons entail some challenges, substantial efforts have to be made with regard to this task. What kind of questions come up in this context shall be highlighted by means of two examples:

- Names of companies often contain some general terms. At least in Germany, one often finds words like “Gesellschaft” (society), “Unternehmen” (enterprise, company), “Handel” (trade) or abbreviations for the legal form like “GmbH & Co. KG”. These words can be longer than the name of the company in a more narrow sense and less prone to different forms of spelling. Solutions for this matter have to be evaluated. The easiest approach would be to delete such terms before string comparisons are carried out. However, this might not be the best way due to the structure of company names.
- Sooner or later, a decision with regard to the string comparison function has to be made. The number of potentially suitable functions is double digit. There is also a broad literature from several disciplines about what works best (cf. as a starting point Herzog et al. 2007). The reason for this is the fact that string comparisons are not only relevant in the context of record linkage but also in information retrieval, in all kinds of search processes respectively. Some string comparison functions which have proven to be powerful are the Jaro-Winkler-Similarity-Function (Jaro 1989, Winkler 2003), the Monge-Elkan-Distance (Monge & Elkan 1996), and the so called cosine similarity. What function will perform best under the specific circumstances in the KombiFiD project is not easy to tell.³ To predict a best solution solely on the basis of theoretical considerations maybe even impossible. A process of implementation, evaluation and re-evaluation is likely to be necessary.

Two issues will finally be discussed: The first one relates to the fact that it is not always known whether or not an enterprise is contained in another data set. In some cases it may be very likely, but not sure. For a record linkage project, this is not an unusual situation. It can even be considered as the standard case. Comparing different scenarios, this is the one that is more difficult. KombiFiD is faced with it mainly for two reasons: Firstly, because only a part of the enterprises exhibit certain properties (e.g. stocks of foreign direct investment above a certain threshold). Therefore, not all enterprises report to all statistics. Secondly, because non-response introduces some uncertainty.

The second issue is the fact that the relevance of false non-matches depends on the size of the enterprise. 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 delineated points sum up the tasks that will have to be tackled in the KombiFiD project with regard to data matching. When the first results are available, it will be interesting to compare these with results from similar projects that have been carried out in other countries.

3. Evaluation of the challenges related to nonresponse

The present legal situation in Germany 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 should be carried out have to give their approval to this procedure.

³ An empirical study done by Cohen et al. (Cohen et al. 2003) comes to the conclusion that the best results – at least in more complicated cases – are achieved with an approach that implements a combination of Jaro-Winkler and a scaled version of the Levenstein distance.

The consequence is a drop-out process that results from the fact that some enterprises will not answer or may 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.⁴ 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 employees⁵ and 9,19% in case of major enterprises with 500 and more employees.⁶ 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, has to be taken into account. In addition to the nonresponse resulting from the mortality of enter-

⁴ For essential information concerning nonresponse see Groves et al. (2002) and Schnell (1997).

⁵ Within the project KombiFiD, a cut-off threshold of 20 employees is utilized in case of Cost structure survey in manufacturing, mining and quarrying, whereas in the annual survey in wholesale the cut-off limit is set at 10 employees.

⁶ Analyses based on Cost structure survey in manufacturing, mining and quarrying and Cost structure survey in the construction industry, survey years 2003 to 2008.

prises the fact that enterprises asked in the postal survey can reject the merging of their respective data leads to the second cause of nonresponse in the KombiFiD-Survey. Data can only be integrated for enterprises which have approved this procedure in a postal survey. Within this survey a sample of about 60.000 enterprises will be asked whether they agree with the combination of their data. It is not allowed to merge data of enterprises not approving this procedure, but obviously information and data regarding the level of this kind of nonresponse can only be given following the end of the survey, which comprises the postal survey in April 2009 as well as a first and second reminder with an interval of six weeks in each case.

Finally, a third drop-out process has to be considered: For example, enterprise A 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 enterprise A in the Foreign Direct Investment Stock Statistics. One potential solution to this kind of methodological problem offers the imputation⁷ of the missing values (vgl. Braakmann 2008, S. 8). First results regarding the question whether or not procedure is reasonable and efficient to deal with the aforementioned item-nonresponse can be expected at the end of 2009, when first results are available.

4. Conclusion

The elaborations on the methodological issues which arise in the context of the KombiFiD project show that there are indeed some challenges. The discussion however, should also have made clear, that the issues can be tackled with appropriate solutions. As soon as results of empirical analyses become available, the quality and the benefit of the data will be evaluated in detail.

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, N. (2008). *Ein Konzept für non-response Analysen zum Projekt KombiFiD*. Lüneburg (Unpublished paper)
- CENEX-ISAD (2008a) Report of WP1. *State of the art on statistical methodologies for integration of surveys and administrative data*
- CENEX-ISAD (2008b) Report of WP2. *Recommendations on the use of methodologies for the integration of surveys and administrative data*
- Cohen, W.; Ravikumar, P.; Fienberg, S. (2003) A Comparison of String Metrics for Matching Names and Records. *Proceedings of the KDD-2003 Workshop on Data Cleaning, Record Linkage, and Object Consolidation*, 73-78

⁷ Regarding possibilities of imputation see Drechsler et al. (2008) and Schafer/Olsen (1998) among others.

- Drechsler, J.; Dundler, A.; Bender, S.; Rässler, S.; Zwick, T. (2008) A new approach for disclosure control in the IAB establishment panel - Multiple imputation for a better data access, in: *Advances in Statistical Analysis*, 92, 439-458
- Fellegi, I.; Sunter, A. (1969) A Theory for Record Linkage, in: *Journal of the American Statistical Association*, 64, 1183-1210
- Groves, R. et al. (Eds.) (2002) *Survey Nonresponse*. New York: Wiley
- Herzog, T.; Scheuren, F.; Winkler, W. (2007) *Data Quality and Record Linkage Techniques*. New York, Berlin: Springer
- Jaro, M. (1989) Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida, in: *Journal of the American Statistical Association*, 84, 414-420
- Monge, A.; Elkan, C. (1996) The field-matching problem: algorithm and applications, in: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 267-270
- Schnell, R. (1997) *Nonresponse in Bevölkerungsumfragen*. Opladen: Leske und Budrich
- Schafer, J. L; Olsen, M. K. (1998) Multiple imputation for multivariate missing-data problems: A data analyst's perspective; in: *Multivariate Behavioral Research*, 33, 545-571
- Winkler, W. (1999) *The state of record linkage and current research problems*. U.S. Bureau of the Census. Statistical Research Division (PDF – R99-04)