Improving the content of administrative data by linking different register-based data sources

Daniela Hochfellner and Axel Voigt*

Research Data Centre (FDZ) of the Federal Employment Agency (BA) at the Institute for Employment Research (IAB) in Nuremberg
Regensburger Str. 104
90478 Nuremberg, Germany
daniela.hochfellner@iab.de, axel.voigt@iab.de

Keywords: Data Linkage • Administrative Data • Data Cleansing • Process-Generated Data

1. Motivation

In the project "Biographical data of selected social insurance agencies in Germany" (BASiD), assisted by the Federal Ministry of Education and Research, German administrative data on individuals of two social security agencies, namely the Federal Employment Agency and the German Pension Insurance, are merged. The aim of the project is to generate a combined longitudinal biography dataset for the first time and to provide it to the scientific community as Scientific Use File as well as a weak anonymous dataset accessible by on-site use. The joint research project was started in the beginning of 2009 with a current time of three years. Due to the rising interest of administrative data in science, it is necessary to work constantly on projects that deal with the improvement of the quality of register based data. The BASiD project is such a kind of project. The richness of information on individuals will be increased, through filling up gaps in the single data sources by using the information of the other data source. The intension is to achieve a two-way accumulation of information in the existing data that allows researchers to accomplish more differentiated analysis in various research topics.

Normally nobody will assume problems with the linking procedure because the used data sources are built on the equal legal basis namely the Code of Social Law. Additionally they are based on the same notification process. Thus a unique identifier is available. Anyway the information is not the same in the different sources because the responded social security agencies only collect data on employment histories of individuals which are relevant for their own field of activity. There are inconsistencies which result either out of the fact of different editing procedures at the agencies or possible mistakes during the data collection which lead to an existence of multiple states in the data. The cleansing procedures for the first mentioned inconsistencies can be easily executed with mathematical conversion algorithms, whereas the second mentioned inconsistencies are much harder to deal with. To cope with the multiple states we analyse sequences of all possible states to identify misfits in the combined data. Finally these identified misfits are corrected with heuristics based on certain assumptions regarding the Code of Social Law.

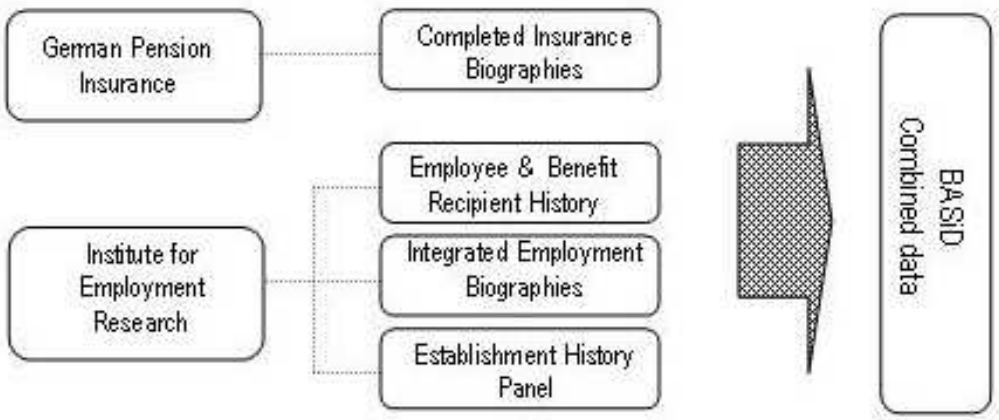
The paper informs about the contents respectively the data sources of the new developed dataset, the cleansing procedures and heuristics we use to generate the combined BASiD data.

*This paper was written in the project BASiD. The financial support is given from the Federal Ministry of Education and Research. We would like to thank Patrycja Scioch of the Institute for Employment Research for supporting us in implementing the data splitting for our project.

2. Description of the used data sources

The project has the objective to create a dataset that reproduces employment histories as complete as possible. This requires a multitude of information. The project combines four different data sources. These data sources are the Completed Insurance Biographies of the German Pension Insurance and the Employee and Benefit Recipient History, the Integrated Employment Histories and the Establishment History Panel of the IAB. The connection between the different data sources is reviewed in Figure 1.

Figure 1: Overview of the data sources



2.1 The Completed Insurance Biographies

The Completed Insurance Biographies (VSKT) is a running special survey of the German Pension Insurance. It contains every insured person of the pension insurance and provides information about the state of their entitlement to the German Pension Insurance with all (stored) pension relevant facts. Pension relevant facts include all times and contributions that are stored on a person`s pension account (see Himmelreicher and Stegmann 2008). The information is available since 1938.

2.2 The Employment and Benefit Recipient History

The Employment History and Benefit Recipient History (BLH) is an individual data set that includes information of two different data sources. The first data source displays the Employee History. It covers the time span since 1975 and contains every social security notification of a single person. Since 1st April 1999 notifications about marginal part-time employment are recorded additionally. The second component of the BLH is the Benefit Recipient History. The data source contains any deregistration from the receipt of unemployment benefit, unemployment assistance or maintenance benefit since 1975. With the so-called "Hartz" reforms which were introduced on 1st January 2005 the content of the data has changed. The receipt of unemployment assistance and maintenance benefit was pooled together and is now called unemployment benefit II. This information is stored separately from that time on. Hence it is no more content of the Benefit Recipient History.

2.3 The Integrated Employment Biographies

The Integrated Employment Biographies (IEB) consist of three additional data sources. The Unemployment Benefit II Recipient History fills the gap that appears in the Benefit Recipient History after the 1st January 2005 and contains information about the receipt of unemployment benefit II. The participants-in-measures data displays times of participation in programmes of active labour market policy. Finally the third data informs about the job-search status. However one has to take into consideration that information from the IEB is initially available since 1st January 2000 (see Jacobebbinghaus and Seth 2007).

2.4 The Establishment History Panel

The Establishment History Panel (BHP) contains every establishment in Germany that employs at least one person liable to social security at 30th June for every year. Since the 1st January 1999 this is also true for establishments with at least one marginal part-time employee. The BHP is constructed by yearly cross-sections since 1975 in the case of West-German establishments and since 1992 for East-German establishments. From this data source the BASiD data set will be provided with firm information like the firm size or the industrial sector (see Spengler 2009).

3. Data cleansing and application of heuristics

The linking and cleansing process of the different data sources is done in successively arranged steps. To sum up, first of all, the different sources are merged. In a second step the problems which turn up with regard to the data merging have to be solved. Therefore we check for every spell whether the information about a person is identical in both data sources. If this is the case we call the spell a “twin-spell”. After having identified all of the “twin-spells” the information transmission can be executed. In this chapter each step of the linkage, cleansing routine and the developed heuristics for finding statistical twins are described in detail.

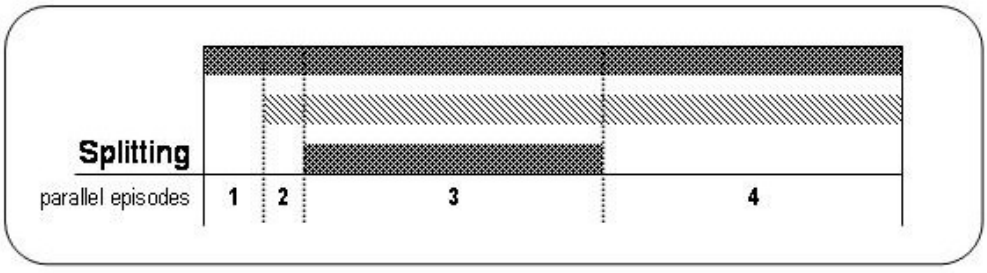
3.1 Merge of the different data sources

The data of the Federal Employment Agency as well as the data of the Pension Insurance have a uniform identifier available: the social security number. Beyond that, the data of both institutions have the same basic structure. The linkage of both data sources was done via the identifier, begin and end date of the episode, the actual state in the employment history of the person and the daily wage. Although one will assume that there will be no problems with the data merging, there are inconsistencies turning up which have to be taken care of. These inconsistencies come from different editing procedures of the two agencies or the characteristics of the used data. There are observations which do not essentially must have a match, for example the ones which can only be found in one of the two sources. Due to this matter of facts it is not remarkable that only ten percent of the observations match perfectly when doing the data merging. A further task is to develop strategies and heuristics to find more identical information in the data, which are displayed in the following.

3.2 Data splitting

The explanation for not finding all the identical information in the first run is the notification process which is responsible for the data collection. Due to different notification procedures in both institutions, identical information in the various data sources is held differently. One of the disparities can be explained through different observation periods. Therefore not all of the identical information is recognized as "twin-spell" when linking the data. To identify these "twin-spells", an episode splitting, with regard to construct identical observation periods, was executed. Figure 2 illustrates the splitting routine.

Figure 2: Illustration of the splitting



Exemplary the original episode can be split in several single episodes, which - except the different begin and end date - contents identical information. The longer episodes were separated in shorter episodes in a way, that in all data sources identical observation periods are generated in reference to the particular state of the observed employment history.

3.3 Application of mathematical algorithms

One also has to consider, that besides the “twin-spells” that are found directly through the merge procedure, there also can exist hidden “twin-spells”. These are observations of a given person that are identical in both data sources relating to the respective state in the employment histories, but differ in the daily wage. It may be misleading to consider these observations as misfits. Three reasons apply why the daily wage differs between the two data sources and make mathematical algorithms necessary to adjust the information. First the calculation of the daily wage is based on working days until 1999 and after that in calendar days in the data of the Federal Employment Agency, while the German Pension Insurance uses calendar-days continuously. Second the data of the German Pension Insurance do not display the wage that is really earned by a person, but the wage that is relevant for the calculation of pensions. Because of that, the daily wages employees in Eastern Germany earn have to be converted corresponding to annex 10 SGB VI § 256. Finally the German Pension Insurance starts to convert from DM into Euro in the year 2001, while the Federal Employment Agency already has done this since 1999. So the currency too, has to be adjusted. These mathematical conversions help to identify about another seven percent of the overall number of spells as “twin-spells”. Moreover, as a consequence of the conversions, rounding errors in the daily wage variable arise. It is obvious that, if the difference in the daily wage between the two data sources is small, we are dealing with a “twin-spell”. So we assume a “twin-spell” if the daily wage in the two data sources only differs in the decimal place. This way further eleven percent of the overall number of spells can be recognized as “twin-spells”. Relating to different daily wages there is also the possibility that a “twin-spell” is overlooked because the daily wage is displayed with a missing value in one of the data sources. We decided to use as much information as possible. That is why we replaced the missing value with the available information of the well-stocked data source.

3.4 Account declaration of the insurance accounts

The data of the Pension Insurance has one advantage, namely the account declaration, which can be used for the data cleansing. Account declaration means, that the Pension Insurance proofs the reported notifications. From the age of 30 on, employees which are subject to social security, get a regularly information writing, which contains the employment times that are relevant for the annuity computation. This way originated mistakes are recognized and corrected. The data of the Federal Employment Agency do not have this correction loop. Out of this fact we assume that the information out of the notification of the pension insurance is the correct one, if there are deviations between the different sources in the case of an account declaration. The information on the observation of the German Employment Agency is corrected therewith. After this correction most of the twins can be found. At this point in the project about 80% of the observations are proofed.

3.5 How to deal with Deviations in the different data sources

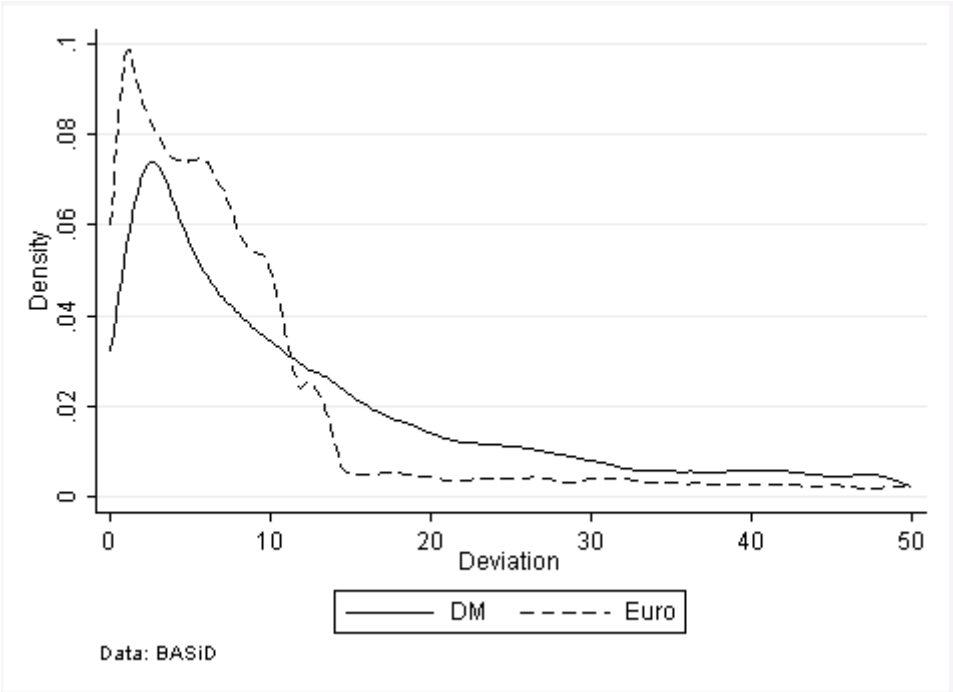
Finally there remain spells without clarified accounts. We assume a "twin-spell" when the deviation in the daily wage between both data sources is not larger than one. Because we cannot decide, which data source is more reliable, the daily wage is set to the mean of the different wage indications.

$$x = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ if } \max(x) - \min(x) \leq 1$$

Whereas $i = 1,2,...,n$ is the number of simultaneous spells of different data sources for a given person and $x = salary$.

Figure 3 displays the Deviations of the daily wage of the Pension Insurance data and the data of the Employment Agency separated by currency.

Figure 3: Deviations of the data sources

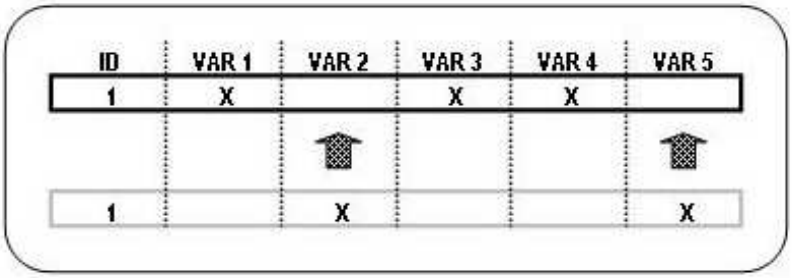


The distribution of the density shows that most of the discovered deviations are very small. That is why we decided to put our matching criterion to a deviation to a maximum of one. The monthly deviation is therefore at a maximum of 30 Euro or 60 DM. This heuristic declares around 10 percent of the number of overall spells as "twin-spells". When increasing the maximum deviation of the daily wage to two DM/Euro the number of additionally found "twin-spells" is disproportionately low. The graph also shows that the deviations are less for DM than Euro. If the deviation between the daily wages is greater than one, the daily wage is set to missing, because there is no other hint in the data, which helps to find the true information.

3.6 Information transmission loop

The aim of this step is to transfer information for the located “twin-spells” in the data. One “twin-spell” can be equated with the existence of two identical observations. For each of these the information is compressed onto one single observation. After the transmission the duplicate is deleted. With the cleansing procedures, the “twin-spells” have been already marked. So the transmission of the content of the single variables can be easily executed. Figure 4 displays how we implemented the value endorsement.

Figure 4: Illustration of the information transmission



In the developed routine the information on the observation of the data of the German Pension Insurance has been transferred onto the observation of the Federal Employment Agency data. As result the duplicate was dropped. The transmission loop has to be executed several times, because the existence of multiple parallel episodes is possible. Each loop creates one “twin-spell” episode out of the two single episodes. After the transmission routine about seven percent of the overall number of spells cannot be defined as “twin-spells”. How to proceed with these will be shown in the next step.

3.7 Code of Social Law

The fact that seven percent of spells cannot be found as a “twin-spells”, is because the indication about the state in the observed employment history differs in both data sources. For every possible combination of employment states that appear in the merged data set, it has to be proven, if this combination is possible. This means, we checked for every combination-sequence that was found in the merged data set if the coexistence is possible regarding the Code of Social Law. The checking was done by analysing sequences of different states that appear at the same time. The following example shall illustrate the procedure: In the merged dataset there may be a person that first receives unemployment benefit and second is a job-seeker in the data of Federal Employment Agency. At the same time this person is also a marginal part-time employee. While marginal part-time employment and job-seeking is compatible with both the receipt of unemployment benefit and unemployment assistance, it is not continuously in the case of simultaneous receipt of unemployment benefit and unemployment assistance.
Most of the analysed sequences can be seen as conform regarding the Code of Social Law. Only one percent of the overall observations are left for further cleansing procedures.

4. Summary and future tasks

To sum up, the developed cleansing procedures and heuristics are very promising. At the end there are about one percent of all observations left, which cannot be cleared with the routines. For these cases we assume mistakes in the data collections which can neither be illustrated nor corrected. These episodes will be deleted of the BASiD data. In this case the person will have an information lack for the respective time. The project is currently work in progress, additional heuristics about how to deal with the misfits that remain after the application of the mentioned routines will be developed. The future agenda is to test the combined data and generate a Scientific Use File for the scientific community. The next step therefore is to make the data anonymous whereas the concept of factual anonymization has to be implemented.

5. Characteristics of the developed data

The combined BASiD data will differ in certain characteristics from the previous existing datasets. It contains a variety of characteristics, which allows the researchers to deal with research questions that could be answered for Germany only less precise in the present. Another benefit is the fact that the dataset displays complete employment biographies of individuals. The data contains all information gained from the social security notification process. This implies apprenticeship-, employment-, unemployment-, job-seeker-, training and payment-details, pension times, consideration times, allowance times, payment dates and birth dates for children of the individuals. Additionally the data contains establishment-information, like the establishment-size, classification of industries or regional information (see Hochfellner and Voigt 2010). For example analyses with regard to birth-rates and employment histories of women, the influence of military/ civil service on the employment histories, life-income and earnings points for the pension or influence of start-up-conditions on the career can be arranged.

REFERENCES

Himmelreicher, R. K. & Stegmann M. (2008): New Possibilities for Socio-Economic Research through Longitudinal Data from the Research Data Centre of the German Federal Pension Insurance (FDZ-RV). Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften, 4, 647-660.

Hochfellner, D., Voigt, A., Budzak, U., & Steppich, B. (2010): Das Projekt BASiD: Biographiedaten ausgewählter Sozialversicherungsträger in Deutschland. Projektinhalte, aktueller Stand der Arbeiten und Analysemöglichkeiten. DRV Schriften, 55/2009, 74-86.

Jacobebbinghaus, P., & Seth, S. (2007): The German integrated employment biographies sample IEBS. Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften, 2, 335-342.

Spengler, A. (2009): The Establishment History Panel. Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften, 3, 501-509.