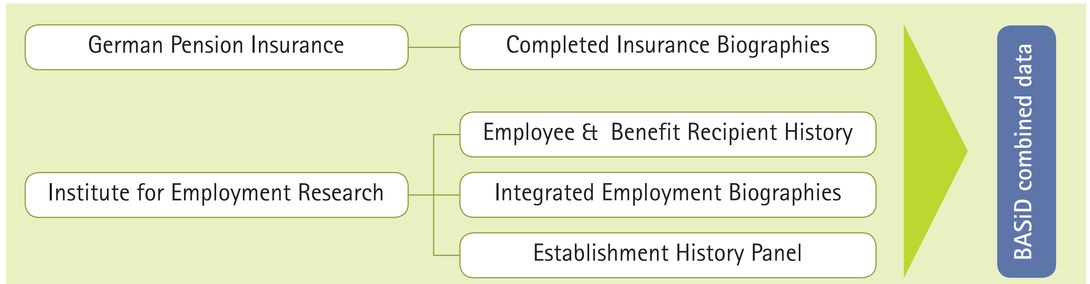
Research Data Centre (FDZ)
of the German Federal
Employment Agency (BA)
at the Institute for
Employment Research (IAB)

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

Motivation & data sources

The project BASiD: „Biographical data of selected social insurance agencies in Germany" is about to merge German administrative data of two social security agencies, the Federal Employment Agency and the Pension Insurance. There is a weak spot recognizable when you work separately on the two datasets, namely the existence of gaps in the single data sources. These come along with a loss of information for the employment histories of the observed individuals. We link these different sources to fill up the gaps.

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. The information is not the same in both 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 have to be taken care of.



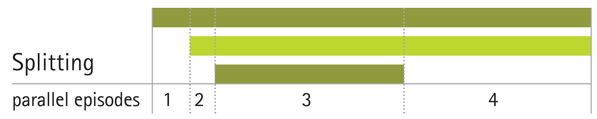
Linkage process & data cleansing

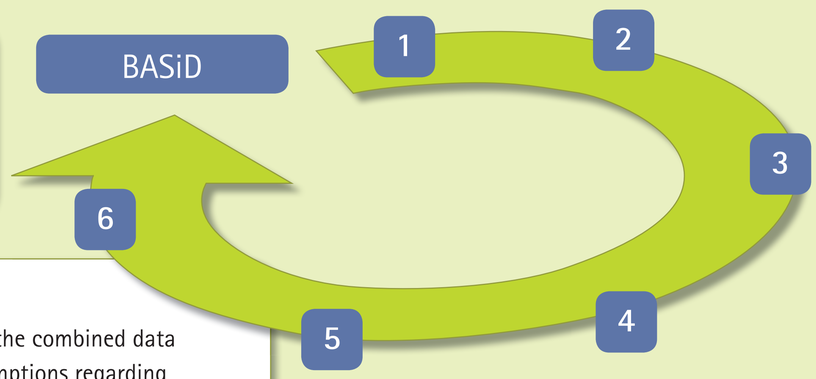
1 Merge

- Uniform identifier is used for the linkage
- Anyway there are inconsistencies:
  - Different editing procedures at the agencies
  - Possible mistakes during the data collection
- Problem: Only 10% of the observations match perfectly
- Solution: Find the perfect match by running cleansing procedures

2 Splitting

- Construct identical observation periods through application of data splitting routine



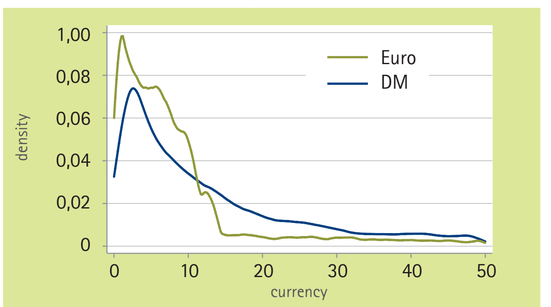


3 Mathematical algorithms

- Information in different data sources have to be adjusted to the new observation period by mathematical conversion algorithms

Let $i = 1, 2, \ldots, n$ be the number of simultaneous spells from different data sources for a given person and $x = salary$ then x is set to $\frac{1}{n}\sum_{i=1}^{n} x_i$ if $\max(x) - \min(x) \leq 1$

- Deviation of the daily wage



Data: BASiD

4 Information transmission loop

- Transfer information of the different data sources of one located match onto one single observation
- Delete duplicate

ID	VAR 1	VAR 2	VAR 3	VAR 4	VAR 5
1	X		X	X	
		↑			↑
1		X			X

5 Code of Social Law

- Analyse sequences of all possible states to identify misfits in the combined data
- Identified misfits are corrected with heuristics based on assumptions regarding the Code of Social Law
- Example: Different states for a given person within the same time period in both data sources

BA/IAB	RV
unemployment benefit	marginal part-time employee
training programme for the unemployed	unemployment assistance
job-seeker	

- Every combination of state is possible since the 1st January of 2005 when Hartz IV reforms substituted unemployment assistance for unemployment benefit II
- Before the simultaneous receipt of unemployment benefit and unemployment assistance was not allowed

6 Final testing

- After the cleansing procedures and the application of heuristics there are less than 1% of the observations in the data which do not match

Information in the final BASiD data (from1940 to 2007)

- Employment and benefit history
- Information on education (military and civil service)
- Times of illness
- Job seeking, training measures participation
- Information on occupation
- Job payments
- Pension insurance payments
- Information on motherhood, number of children
- Regional information
- Establishment information
- Sociodemographic characteristics

The developed dataset is expected to be offered in 2011 as Scientific Use File (SUF) and in a weakly anonymous version for one-site-use.

Authors' Contact:

- Daniela Hochfellner
  E-mail: daniela.hochfellner@iab.de
- Axel Voigt
  E-mail: axel.voigt@iab.de

www.iab.de