Identifying and explaining inconsistencies in linked administrative and survey data: the case of German employment biographies

Huber, Martina; Schmucker, Alexandra

Identifying and Explaining Inconsistencies in Linked Administrative and Survey Data: The Case of German Employment Biographies

Martina Huber & Alexandra Schmucker [*]

Abstract: *»Bestimmung und Erklärung von Inkonsistenzen in verknüpften administrativen und Befragungsdaten: Aufgezeigt am Beispiel deutscher Erwerbsbiografien«*. Surveys often cope with special problems: gaps in retrospection appear or respondents could not provide details. Sometimes these problems can be solved by using additional qualitative information. Another – so far disregarded – possibility is to use process-generated data to expand survey data. The focus of this article is on the potentials and problems of linking administrative and survey data. In particular this is shown by comparison of retrospective survey information on employment cycles and the according process-generated data.

Keywords: Longitudinal Analysis, Process-Generated Data, Social Bookkeeping Data, Public Administrational Data, Survey Data, Mixed Methods, Data Management Record Linkage, Data Fusion, Labour Market Data, Identifying Inconsistencies, Survey Data, Linking Data, Sequence Analysis.

1. Motivation

Most quantitative longitudinal research in social sciences is done with survey data. Surveys suffer from non-response in many ways, for example, coverage errors, unit and item-non-response or attrition. In particular answers to retrospective questions in surveys often imply gaps or incomplete details of remembered episodes. Furthermore retrospective survey data often do not cover a very long period. In order to correct for these errors administrative data can be linked to survey data (Baur 2004; Wallgren and Wallgren 2007). Though the

[*] Address all communications to: Martina Huber; Research Data Centre of the German Federal Employment Agency, Institute for Employment Research, Regensburger Str. 104, 90478 Nuremberg, Germany; e-mail: martina.huber@iab.de.
Alexandra Schmucker, Research Data Centre of the German Federal Employment Agency, Institute for Employment Research, Regensburger Str. 104, 90478 Nuremberg, Germany; e-mail: alexandra.schmucker@iab.de.
This article was written within the project 'Further Training as a Part of Lifelong Learning' (WeLL – Berufliche Weiterbildung als Bestandteil Lebenslangen Lernens). Financial support is provided by the Leibniz-Association (WGL). The authors would like to thank Rainer Schnell for his suggestion to use sequence analysis to explore memory errors (Pigeot-Kübler and Schnell 2006). The authors would like to thank Peter Jacobebbinghaus as well as Nina Baur for helpful comments.

administrative data have drawbacks too – e.g. small number of variables or time lag – they contain valid and exact information. By linking survey data with administrative data, the data quality can be improved by creating a dataset that balances the disadvantages of the administrative and survey data using the advantages of these two different types of data.

Having information from the administrative data for a long time period, the employment biographies from the survey can be completed before the surveyed period and persecuted after the surveyed period (Pyy-Martikainen and Rendtel 2003). This can solve the disadvantage of time restriction of surveys (e.g. interviews should not last longer than one to one and a half hour). Furthermore one can have a look at the overlapping period and reduce gaps. Missing data in the administrative data can be explained by reported information in the survey data. Vice versa recall errors (Becker 2001) or missing data in the survey data can be corrected by linking administrative data (Lane 2008). Furthermore you have additional variables (e.g. on school and university degree or about the household) in the survey data which are not provided by the administrative data. Hence we can learn much about the quality of each dataset and more detailed and reliable information can be used for research (Lane 2008).

One of the problems that can arise from data fusion are inconsistencies between survey and administrative data. Using the example of Germans' employment biographies, in this paper, we will illustrate the advantages of data fusion of survey and administrative data. We will suggest a procedure for identifying, classifying and explaining inconsistencies between these two data sources.

2. Employment Biography Data

2.1 Administrative Data

The administrative data used as an example in this paper are the 'Integrated Employment Biographies of the IAB' (Integrierte Erwerbsbiografien des IAB). These data contain complete employment histories on a daily basis since 1975. The population consists of all employees liable to social security in Germany and gives information about the employment status and the employing firm. This information is very reliable because of the notification scheme which requires employers to report data on their employees.

Furthermore the data contain unemployed persons who receive benefits, participate in measures of active labour market policy or search for a job. Additionally, personal characteristics for all these individuals are collected. Despite all these advantages, one big disadvantage of administrative data is that there is no information on persons who are not liable to social security (e.g. self-employed, maternity leave, military service or education, sick leave or civil-servants). Additionally, there is a time lag between the collection and the avail-

ability of administrative data for research (see Jacobebbinghaus and Seth 2007).

2.2 Survey Data

The survey data used for the following linkage procedure are a part of the results of the cooperation project 'Further Training as a Part of Lifelong Learning' which aims to analyze training decisions of employers and employees. For this purpose in a first step data on firms were collected. These establishments from the IAB Establishment Panel (IAB-Betriebspanel) (Fischer et al. 2008) were classified by industry sector, region and size. In the second step the employees of these establishments of the gross sample were surveyed. The employee data will be longitudinal data combining three waves conducted in the years 2007, 2008 and 2009. We use the first wave of the employee survey in this paper. The survey contains detailed information on individuals' training activities, expectations as well as socio-economic and household characteristics of 6404 individuals. Furthermore the complete employment biography of every individual in the period from January 2006 until the end of 2008 was collected. This includes information about labour market status, job characteristics and changes on a monthly base (see Bender et al. 2008).

2.3 Linking Administrative and Survey Data

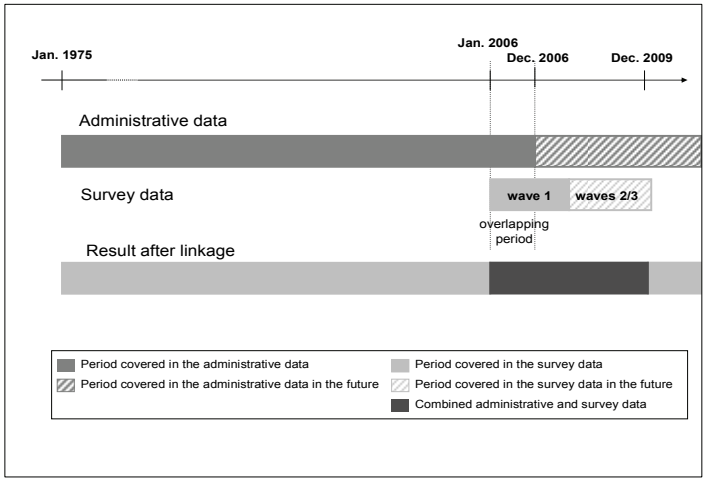
Figure 1 shows the observation periods of the two different data. Linking the administrative data which are available from January 1975 to December 2006 with the first wave of the survey which covers the period from January 2006 to October 2007 yields an overlapping period from one year. After the second and third wave of the survey there will be an overlapping period of altogether four years, from January 2006 to December 2009. Additionally, we will be able to persecute the employment biographies after the surveyed period. For the following analysis the overlapping period of one year between January and December 2006 will be considered.

Altogether we have 5819 individuals in the survey who allowed us to link their biographic data to administrative data. 585 respondents did not give their permission to link their data[1].

[1] Taking a look at some characteristics of the individuals who refused the linkage doing a probit regression, we arrive at the following result: The characteristics having a significant influence are age and the net income. Increasing age and income of the person increases the probability giving the permission to link the survey data with administrative data. Other characteristics used for this probit model were: sex, foreign and school education. There was not any significant influence of these variables to forbid the linkage of survey and administrative data.

Out of the 5819 individuals 5656 could be linked using the insurance policy number and considering age and gender. In 5349 cases the biography is consistent. 5309 of these persons were employed during the whole overlapping period without any gaps. For the remaining 307 individuals, survey and administrative data are inconsistent, as they provide different information on employment status at the same time or overlapping episodes.

Figure 1: Observation periods



Our previous analysis shows that the linkage seems to be correct for a large fraction of the respondents. But there remain important questions: Which kind of deviations can we find? Which characteristics influence the deviations? How well can data from two different sources been linked? These questions we will answer in the following chapters.

3. Using Sequence Analysis to Identify Types of Inconsistencies

This section gives a short introduction into sequence analysis and describes the data preparation which is necessary to apply this method.

Linking our survey and administrative data creates a new dataset with biographical information from both sources for the year 2006. In the following we want to analyse, if the information taken from different sources are consistent.

In the first step, we compare the two biographic sequences for each person and describe the deviations. This approach is very time consuming and almost not feasible for a large number of cases. However we apply this method on

some selected cases. Hence we can identify types of deviations and explain them by using additional information from both sources. Further we can learn about the specific quality of the datasets.

In the second step, we want to quantify the difference between the two sources for each individual so that we can estimate the determinants of the deviation. We use a sequence analysis to calculate a distance measure which specifies the extent of the difference between the two sequences. In contrast to time series or event history analysis the sequence analysis considers the whole employment cycle in a particular time period and not just individual events or durations (MacIndoe and Abbott 2004).

> A sequence is defined as an ordered list of elements (…). The positions of the elements are fixed and ordered by elapsed time or by another more or less natural order (…) (Brizinsky-Fay et al. 2006, 435).

An item is the smallest element of a sequence and can assume diverse values. In our case an item stands for a certain labour market status (e.g. employed, unemployed, schooling) in one month. Episodes are also parts of sequences. In these parts identical items appear in a consecutive order, e.g. an episode of employment with the duration of four months (see figure 2).

According to these definitions we can compare the sequences taken from both data. Further, we want to measure the difference between two sequences. We use optimal matching analysis[2] to generate a distance measure. First the so-called 'Levenshtein distance' is calculated by counting and weighting the steps needed to align two sequences. For this purpose two types of transformations can be used: one can insert or delete items ('indel') or one can substitute items. Every operation causes costs and we have to define the rates for every type of operation:

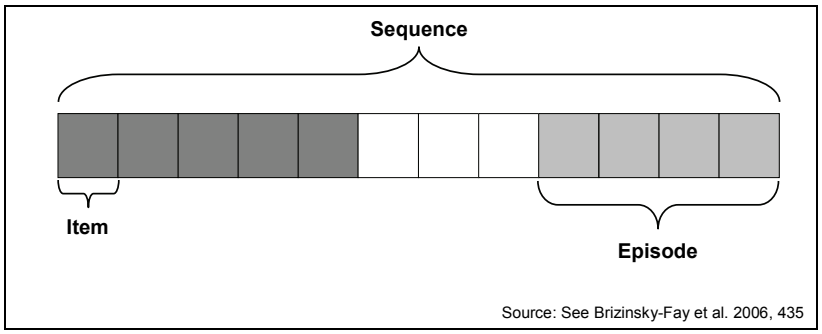
First we have to specify the costs of different substitutions. E.g. one can weight the substitution between the status unemployment and employment higher than the substitution between the unemployment and other statuses. In our case we could not find any plausible reason why diverse transitions should be weighted by different values (Brzinsky-Fay 2006 or Scherer 2001). Second, we have to define the relation between substitution and indel costs. We set up the substitution cost to equal double of the indel cost[3]. This means one substitution causes the same costs as one insertion and one deletion. As there is more than one possible alignment of the two sequences, the alignment with the minimum distance should be chosen (Needleman-Wunsch algorithm). Finally,

[2] MacIndoe and Abbott (2004) give a detailed description of the sequence analysis and optimal matching.

[3] In our case insert/delete is weighted with '1' and substitute with '2'.

the distance measure can be standardized. In our case it is divided by the number of items in the sequence[4].

Figure 2: Sequence, episode and item



Source: See Brizinsky-Fay et al. 2006, 435

Typically the distance measure is calculated for every pair of sequences or for the alignment from every sequence to one particular reference sequence. E.g. researches want to measure the deviation between the employment cycles of diverse individuals or the deviation to a standard biography. In our case optimal matching is applied different: We calculate a distance measure for every person. This measure represents the extent of the deviation between the two sequences taken from the two sources.

As the sequence analysis requires a certain data structure we have to prepare our datasets. Both of our original datasets have a longitudinal structure with sometimes two and more different states for one period (e.g. a person is employed and searches for a new job at the same time). In the first step we define three different states: 'employment', 'unemployment' and 'others'. The last status contains different original statuses which depend on the data source. In the administrative data 'others' can represent 'participation in active labour market programs' or gaps. In the survey data 'others' can stand for 'self-employment, 'education (school, university, apprenticeship)', 'maternity leaves', 'sick leaves', 'military service' or gaps. In the second step both datasets are transformed so that there is one status for each month in the year 2006. If more than one status in one month appears, we apply the rule: We prefer employment to unemployment and unemployment to other statuses.

[4] We conduct our analysis with the software package Stata. For the sequence analysis we use the SQ-Adofiles designed by Brizinsky-Fay et al. 2006. This is a set of tools for sequence analysis which can be implemented in Stata. The codes are available on the following web site: http://econpapers.repec.org/software/bocbocode/s456755.htm.

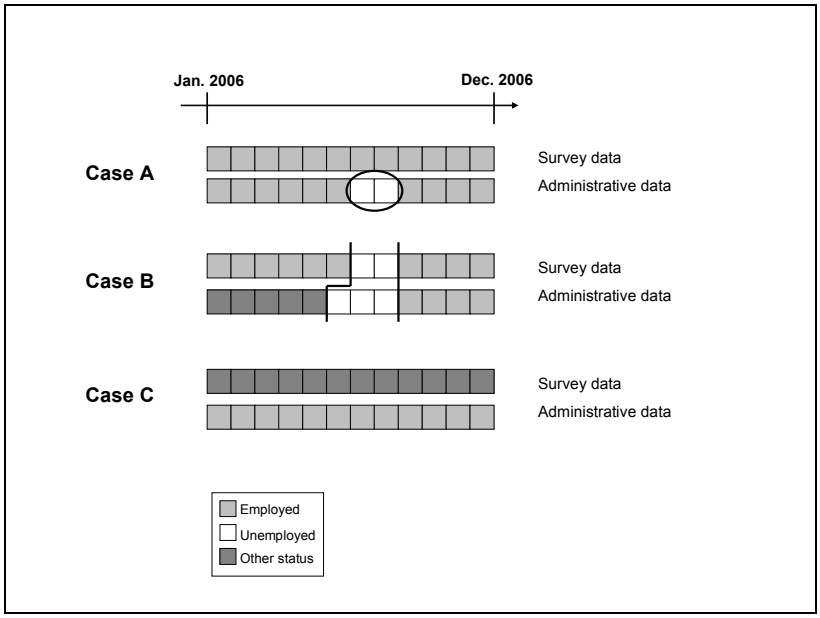
4. Empirical Results

The analysis consists of three parts: first some hypothetical examples of deviation are described. Second, a distance measure is calculated for each individual. Finally, the determinants of the distance measure are analysed using a Tobit model.

4.1 Hypothetical Examples of Deviation

Figure 3 shows three hypothetical examples of deviations. The upper line represents the sequence taken from the survey data, the lower line stands for the sequence taken from the administrative data. Each box represents a month in the year 2006. The grey boxes depict the items 'employed', the white boxes the items 'unemployed' and the dark grey boxes the items 'other status.'

Figure 3: Hypothetical examples of deviation



In the first example (case A) the person reports in the survey a continuous employment for all the year, but there is a short unemployment episode in July and August in the administrative data. Here it seems to be obvious that the person could not remember the short episode of unemployment and hence the administrative data are correct.

In case B, the person reports employment with a short episode of unemployment in between. The administrative data contain a corresponding em-

ployment spell from September until December and an almost consistent unemployment spell in the summer. But there is different information on the status from January until May/June. The administrative data do not provide any information on this period, but in the survey data one can find that the person employed abroad. As the administrative data only cover employment episodes in Germany this fact could not be found in this data. In this case the information taken from the survey data is supposed to be right.

Case C shows completely differing sequences. The individual reports 'other status' in the interview, but the administrative data provide a continuous employment spell. A detailed look in both datasets gives a plausible explanation for the deviation: in the survey data, we can find that the person was enrolled at a university. The administrative data show that the person worked as marginal part-time employee or student trainee. Here both data sets cover partial information and complete one another.

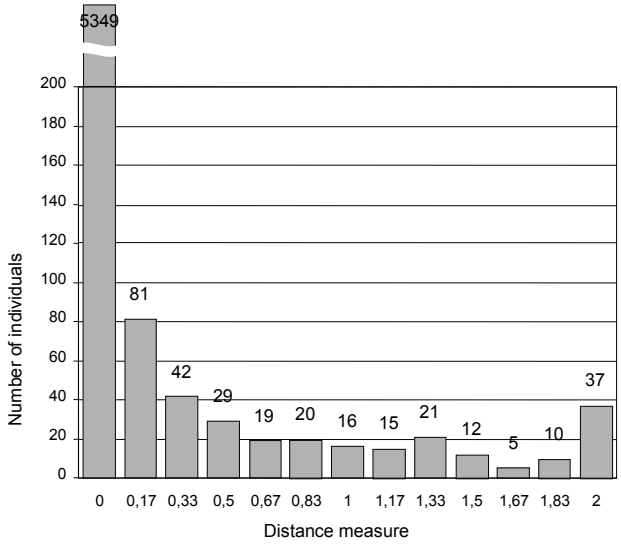
These practical examples show that we can often find plausible explanations for the deviations. Further, we could not conclude that the quality of one dataset is better than the other. These findings should be kept in mind, if one wants do define rules to adjust the combined data.

4.2 Descriptive Results

In the next step, we calculated a distance measure for every person. For this purpose we use optimal matching. We standardized the measure using a distance measure with a range from 0 to 2. '0' means that the both sequences for an individual are exactly the same. '2' means that every status in the 12 month differs in the two datasets.

Figure 4 shows the distribution of the calculated distance measure. The distribution is right skewed with a pike at the value '2' and 95% have no deviation. In other words: only 5% (307 from 5.656 individuals) show inconsistencies. A remarkable group (94 persons) of these individuals reports no change of labour status in the survey but we can find changes in the administrative data for them. Here we can identify a typical pattern (as described in case A): Most of the individuals were employed almost the whole year and have only a short interruption of unemployment (16 individuals) or other status (71 individuals). This is a specific characteristic of survey data, because respondents often do not mention short episodes. 'This can be seen as a result of respondents' tendency to simplify and conventionalise their careers.' (Reimer and Künster 2004: 17).

Figure 4: Distribution of the distance measure



Source: own calculations based on linked survey and administrative data (WeLL Employee Survey 2007 and Integrated Employment Biographies of the IAB)

4.3 Explaining Inconsistencies: Determinants of the Distance Measure

We now want to test which respondent characteristics influence the deviation between biographic sequences in administrative and survey data. According to the results of Reimer and Künster (2004), we assume that persons with more events in the considered time period have more problems to remember everything in particular short episodes. In order to test this hypothesis we apply a probit model (Greene 2000). The dependent variable takes the value 0 if no deviation can be found and 1 if at least one deviation appears. We include a variable that contains the number of status changes based on the administrative data. In our model we additionally control for sex, age, nationality, school degree, vocational education, participation in further training and net income. The base of these variables is the administrative data.

The estimation results show that older persons have a decreasing probability for deviations, until the age of 45 afterwards the probability increases. Persons with no school degree and persons with middle net income have less probability for deviations (see Table 1). Sex, nationality and further training have no significant influence. But the more important result concerns the variable

change. This variable has the biggest impact on the probability of deviations. The more changes occur in a sequence the higher is the probability to find deviations.

Table 1: Determinants of the distance measure

	distance measure	
	Coef.	t
Age	-0.178	(5.99)**
Age2	0.002	(5.55)**
School degree (reference group: no degree)		
Secondary degree	2.015	(2.34)*
Intermediate secondary degree	2.017	(2.34)*
Upper secondary degree	1.907	(2.20)*
Other degree	2.267	(2.48)*
Net income (reference group: < 500 Euro)		
500 up to 999 Euro	-0.297	(1.53)
1000 up to 1499 Euro	-0.685	(3.76)**
1500 up to 1999 Euro	-0.967	(4.89)**
2000 up to 2499 Euro	-1.118	(4.81)**
2500 up to 2999 Euro	-1.014	(3.28)**
3000 up to 3999 Euro	-1.032	(3.50)**
4000 up to 4999 Euro	-0.859	(2.48)*
5000 Euro and more	-0.364	(1.02)
Number of status changes (reference group: no changes)		
1 change	2.877	(21.03)**
2 or more changes	3.899	(20.74)**
Constant	0.254	(0.25)
Observations		5473

Absolute value of t statistics in parentheses, * significant at 5%; ** significant at 1%

5. Conclusion

We link the WeLL survey data to administrative data and show that this is fruitful in many aspects, especially with regard to analyse and check the infor-

mation on the employment biographies. We find in our analysis: Only 5% of the individuals show deviations in the observed time period of one year. Additionally, determinants were identified which have an influence on the probability of deviations. More changes in the sequences of the administrative records cause a higher probability of deviations. Further, we can say that the survey information is quite accurate, at least if the surveyed retrospective biographic episodes are not too long and are in the recent past.

As the work we presented in this paper is just the first step to link our survey and administrative data there is much work left to be done. Our next challenge will be to extend the analysis to the data of the 2nd and 3rd wave of WeLL. As soon as the data are available we moreover can link the administrative employment biography after the survey.

As we know that the linkage of administrative and survey data bears no bigger problems questionnaires could be reduced on biographic data and only episodes and status have to be surveyed which are not covered by the administrative data.

References

Baur, Nina. (2004): Wo liegen die Grenzen quantitativer Längsschnittanalysen? In: Bamberger Beiträge zur empirischen Sozialforschung 23.

Becker, Rolf. (2001): Reliabilität von retrospektiven Berufsverlaufsdaten. In: ZUMA-Nachrichten 49. 29-56.

Bender, Stefan / Fertig, Michael / Görlitz, Katja / Huber, Martina / Schmucker, Alexandra (2008): WeLL – Unique Linked Employer-Employee data on Further Training in Germany. Ruhr Economic Papers 67. Essen.

Brzinsky-Fay, Christian (2006): Lost in Transition: Labour Market Entry Sequences of School Leavers in Europe. Discussion Paper, Wissenschaftszentrum für Sozialforschung. Berlin.

Brzinsky-Fay, Christian/Kohler, Ulrich/Luniak, Magdalena (2006): Sequence analysis with Stata. In: The Stata Journal 6(4). 435-460.

Fischer, Gabriele / Janik, Florian / Müller, Dana / Schmucker, Alexandra (2009): The IAB Establishment Panel things users should know. In: Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften, Jg. 129, H. 1. 133-148.

Greene, William. H. (2000): Econometric Analysis. New Jersey.

Jacobebbinghaus, Peter / Seth, Stefan (2007): The German Integrated Employment Biographies Sample IEBS. In: Schmollers Jahrbuch 127(2). 335-342.

Lane, Julia (2008): Linking Administrative and Survey Data. Forthcoming

MacIndoe, Heather / Abbott, Andrew (2004): Sequence Analysis and Optimal Matching Techniques for Social Science Data. In: Handbook of Data Analysis**.** 387-406.

Pigeot-Kübler, Iris / Schnell, Rainer (2006): Errors in autobiographical memory and their effects in time-to-event analysis. Proposal for a research grant to the DFG. New application within the program „Survey Methodology“ (unpublished manuscript).

Pyy-Martikainen, M. and U. Rendtel (2003): The Effects of Panel Attrition on the Analysis of Unemployment Spells. CHINTEX Working Paper 10.
Reimer, Maike / Künster, Ralf (2004): Linking Job Episodes from Retrospective Surveys and Social Security Data: Specific Challenges, Feasibility and Quality of Outcome. Berlin: Max-Planck-Institut für Bildungsforschung.
Scherer, Stefani (2001): Early career Patterns: A Comparison of Great Britain and West Germany. In: European Sociological Review 17(2). 119-144.
Wallgren, Anders / Wallgren, Britt (2007): Register-based Statistics: Administrative Data for Statistical Purposes. England.