Linking survey data with administrative employment data: The case of the IAB-ALWA survey

Manfred Antoni[*]

January 25, 2011

JEL classification: C81
Keywords: record linkage

1 Motivation

For various research questions and methods of inference in sociology or economics, rich data sets are required. Since survey and administrative data sets have their respective comparative advantages, a combination of both data sources enhances the information at hand. This study describes the process of linking the newly available German survey "Work and Learning in a Changing World" (Arbeiten und Lernen im Wandel, ALWA) with administrative data from the Federal Employment Agency.

The range of research questions that can be tackled with a combination of these data sets is manifold. For instance, analyses that are conducted on the basis of survey data often lack accurate information on levels and the development of wages. Administrative data provide such information on wages, but they only contain very little information on educational attainment and completely lack characteristics of the social background.

Moreover, analyses of the reliability and accuracy of survey data can be conducted. The combined data set is supposed to include the employment histories both from the view of the respondent and their representation in administrative records. Comparisons of these employment histories make it possible to identify deviations in terms of the dating and duration of events.

[*] Institute for Employment Research, Regensburger Strasse 104, 90478 Nuremberg, Germany, email: manfred.antoni@iab.de

This might facilitate methodological improvements regarding the gathering of longitudinal information.

The goal of the study is threefold. First, the potential for research is demonstrated by giving an account of the information available in both data sets. Second, the study demonstrates how the consent of respondents to this procedure varies by individual characteristics. This is done my multivariate analyses based on the survey data. Third, the actual process of record linkage for the data at hand is described. It is shown how the success of the linking process depends on different strategies to find matches in both data sets (deterministic vs. probabilistic record linkage).

2 Data sets

The IAB ALWA survey (cf. Antoni et al., 2010) includes more than 10,000 retrospective interviews with people born between 1956 and 1988. Longitudinal information was gathered on residential, educational, employment and partnership histories as well as on children and times of parental leave. All these events are measured detailed to the month. This data is complemented by a rich set of cross-sectional variables. The interview covers topics such as place and date of birth, immigrant background, religiousness, language skills, family background, importance of different domains of life as well as informal learning and cultural activities. Aided recall techniques were used during the interviews (cf. Drasch and Matthes, 2011) to reduce recall error.

Administrative data from the IAB contain daily information on employment histories. They include information on transfer payments and wages, which are measured with high accuracy as they are related to social security contributions. From the year 2000 on, information on the participation in different active labour market policy measures is included. Employment spells can be supplemented by firm data on the economic sector, the qualification and age structure of employees as well as the wage distribution.

3 Determinants of consent

For any attempt to match survey data with other data sources, it is crucial whether data protection regulations make the consent of respondents necessary. If that is the case, interviewers are legally bound to inform respondents about the nature of information that is going to be matched and ask for consent to that procedure explicitly. This is the case in Germany. During the IAB ALWA survey, consent for record linkage was given by 9,531 (92%) of the respondents.

The consent of respondents is influenced by a wealth of factors. Aspects of the study in general are important, such as the topic of the survey, the nature and amount of the information that is going to be matched from other data sources, the context of the question of consent inside the interview, the purported usage of the combined data or the client of the survey institute. This enumeration is far from conclusive, but it shows that findings of other studies with similar topics might not be applicable to different survey contexts.

During a given interview situation, the elapsed duration of the interview or characteristics of the interviewer might influence the likelihood of consent. Individual characteristics of the respondent such as age, qualification, economic situation as well as trust in general or towards surveys might also play an important role. The influence of these factors on matching consent is examined with binary outcome models, which are based on information given by the respondents and on paradata of the survey.

4 The procedure of record linkage

Among the aspects that influence the success of the actual process of record linkage is the existence of good identifiers between records of the different data sets. As the population of interest in the ALWA survey consists of individuals living in Germany regardless of their labour market status or nationality, the sample was drawn from registers of the residents' registration offices of 250 German municipalities. The result is a sample of addresses without a unique identifier related to the administrative records of the Federal Employment Agency. For the ALWA survey, record linkage is therefore performed on the basis of the following identifiers:

- first and last name
- gender
- day, month and year of birth
- postal code
- place name
- street name
- house number

Before the records of both data sets are compared, extensive preprocessing was conducted to clean up typographical errors, to minimize the amount of different spellings of names, places and street names as well as to fill in missing information in postal codes or place names. These steps of standardization are done consistently for both the administrative records and the survey data. This leads to exact matches on all variables for about 70% of those survey respondents who gave their consent for record linkage.

To increase the amount of successful matches, probabilistic record linkage is applied with different string comparator metrics (cf. Herzog et al., 2007). The model based on the work of Fellegi and Sunter (1969) facilitates the classification of record pairs into links, possible links and non-links after comparing both files. Pairs that are classified as possible links are subsequently coded as either links or non-links by hand. The study documents the success of all different stages of the record linkage process and draws conclusions for similar projects based on the administrative records of the Federal Employment Agency.

References

Antoni, Manfred, Katrin Drasch, Corinna Kleinert, Britta Matthes, Michael Ruland, and Annette Trahms (2010). 'Arbeiten und Lernen im Wandel. Teil I: Überblick über die Studie'. In: *FDZ-Methodenreport* 05/2010. URL: http://doku.iab.de/fdz/reporte/2010/MR_05-10.pdf.

Drasch, Katrin and Britta Matthes (2011). 'Improving retrospective life course data by combining modularized self-reports and event history calendars. Experiences from a large scale survey'. In: *Quality & Quantity* forthcoming.

Fellegi, Ivan P. and Alan B. Sunter (1969). 'A Theory for Record Linkage'. In: *Journal of the American Statistical Association* 64.328, pp. 1183–1210.

Herzog, Thomas N., Fritz J. Scheuren, and William E. Winkler (2007). *Data quality and record linkage techniques*. New York, NY: Springer.