ALWA-Befragungsdaten verknüpft mit administrativen Daten des IAB (ALWA-ADIAB)

Erster Workshop des German RLC

Nürnberg

5. April 2012

Manfred Antoni

Outline

- Data

- Process of record linkage

- Data access and outlook

Work and Learning in a Changing World (ALWA) (I)

- Retrospective interviews (CATI) with > 10,000 German residents
- Birth cohorts 1956–1988 (aged 18–52 at time of interview in 2007/08)
- Monthly longitudinal information on, e.g.
    - educational,
    - employment,
    - residential,
    - partnership, child-rearing and parental leave histories,
    - military and alternative services, etc.
- Aided recall techniques reduce recall error during interviews. (Drasch and Matthes, forthcoming)

Work and Learning in a Changing World (ALWA) (II)

- Detailed cross-sectional information on, e.g.
  - place and date of birth, immigrant background, language skills,
  - family background,
  - cultural capital,
  - religiousness,
  - importance of different domains of life,
  - self-reported measures of cognitive skills and personality traits etc.
- Literacy and numeracy tests (PAPI, 30 min. per domain) for a sub-sample
- Data access: Scientific Use File provided by the FDZ

  `(http://fdz.iab.de/en/FDZ_Individual_Data/ALWA.aspx)`

Administrative data of the German Federal Employment Agency

- Sources of data of the German Federal Employment Agency (BA):
  - Social security notifications by employers
  - Data from internal processes of the local employment agencies and the BA
- Used to compute social security contributions and benefit entitlements
- High accuracy of income information
- Recall error or sample selectivity impossible
- Consistent Social Security Number $\Rightarrow$ once a survey respondent is identified in the register data, the complete employment history is available.
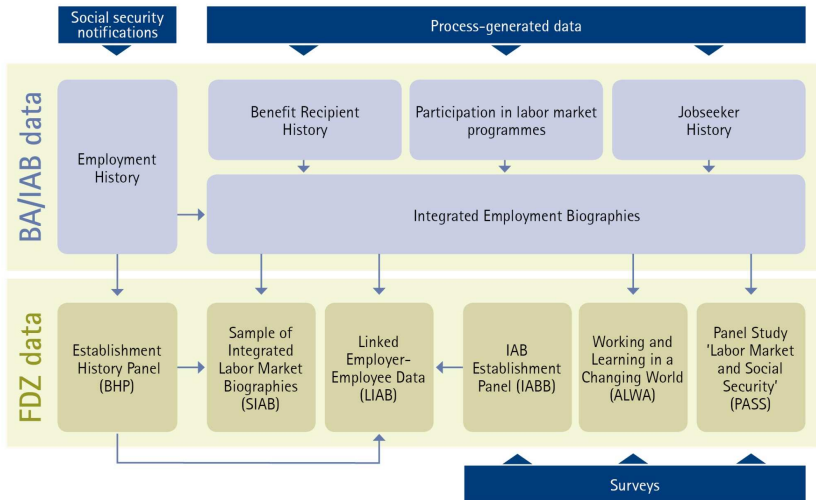
Administrative labor market data on individuals

- Day-to-day information on histories of dependent employment and registered unemployment since 1975 (over 80% of the German labor force)
- Longitudinal information on earnings and benefit receipt
- Since 2000: detailed information on the participation in active labor market policy measures and job seeking
- Detailed social security benefit information since 2005 (household-level)
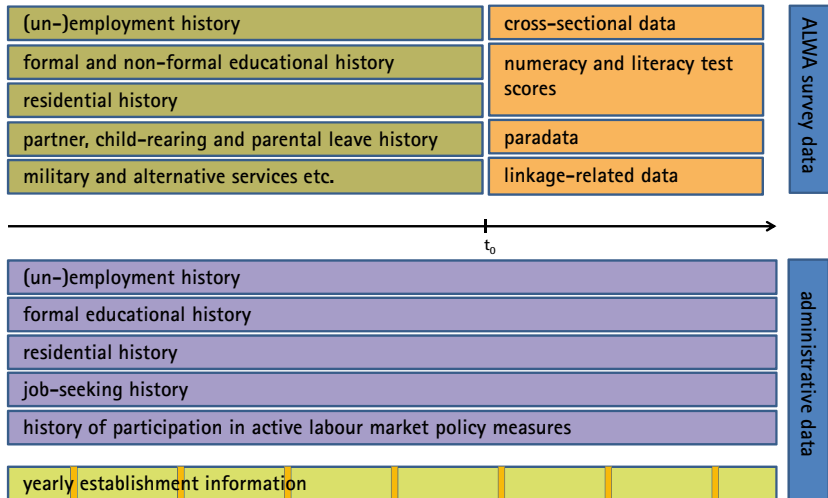
Administrative labor market data on establishments

- Daily establishment-level employment and wage bill from 1975-present
- Establishment History Panel (BHP) includes yearly measures on:
    - economic sector
    - qualification and age structure
    - wage distribution inside the firm
    - worker flows for different subgroups of employees
    - founding and closing of firms
- Variables computed for every firm based on 30th of June of each year

Overview of IAB data

Social security
notifications

Process-generated data

BA/IAB data

Employment
History

Benefit Recipient
History

Participation in labor market
programmes

Jobseeker
History

Integrated Employment Biographies

FDZ data

Establishment
History Panel
(BHP)

Sample of
Integrated
Labor Market
Biographies
(SIAB)

Linked
Employer-
Employee Data
(LIAB)

IAB
Establishment
Panel (IABB)

Working and
Learning in a
Changing World
(ALWA)

Panel Study
'Labor Market
and Social
Security'
(PASS)

Surveys

Structure of ALWA–ADIAB

(un–)employment history	cross–sectional data
formal and non–formal educational history	numeracy and literacy test scores
residential history	
partner, child–rearing and parental leave history	paradata
military and alternative services etc.	linkage–related data

ALWA survey data

t_0

(un–)employment history
formal educational history
residential history
job–seeking history
history of participation in active labour market policy measures

yearly establishment information

administrative data

Consent to record linkage in Germany

- Survey institute or interviewers are legally bound to inform respondents about the nature of information that is going to be matched.
- The interviewer has to ask for approval to that procedure explicitly.
- The respondent has to be able to make an informed decision.
- During the ALWA survey, consent was given by 9,531 (92%) of the respondents.

Question of consent

Um das Interview im Folgenden möglichst kurz zu halten, würden wir gerne bei der Auswertung der Befragung Auszüge aus Daten einbeziehen, die beim Institut für Arbeitsmarkt- und Berufsforschung der Bundesagentur für Arbeit in Nürnberg vorliegen.

Dabei handelt es sich zum Beispiel um Informationen zu vorausgegangenen Zeiten der Beschäftigung und der Arbeitslosigkeit und zur Teilnahme an Maßnahmen während einer Arbeitslosigkeit.

Zum Zweck der Zuspielung dieser Daten an die Interviewdaten möchte ich Sie herzlich um Ihr Einverständnis bitten.

Werden diese Informationen ausgewertet, so ist absolut sichergestellt, dass alle datenschutzrechtlichen Bestimmungen strengstens eingehalten werden.

Ihr Einverständnis ist selbstverständlich freiwillig. Sie können es auch jederzeit wieder zurückziehen.

Ist das in Ordnung?

Identifiers of respondents

- ALWA lacks unique identifiers for a direct link to administrative records.
- Identifiers for matching:
  - first and last name
  - gender
  - day, month and year of birth
  - postal code, place name, street name and house number
- Sources of identifiers:
  - Field information (infas Institute for Applied Social Sciences)
  - Personal register data (IAB department IT Services and Information Management)
- Reference year for both sources: 2007
- Extensive standardization of identifiers

Standardization

- Extensive standardization of identifiers before records are compared:
    - minimizing variation in spelling of names, places and street names,
    - filling in missing information in postal codes or place names,
    - deleting blanks and special characters,
    - standardizing or deleting abbreviations,
    - deleting suffixes of house numbers.
- All steps are applied consistently to both address files.

Subsequent steps of the linkage process

- Exact matching
  - Records from both data sources are directly matched based on all available fields.
  - Even smallest variations in spelling lead to a rejection of a potential match.
- Probabilistic record linkage
  - Comparison with software Merge ToolBox (MTB, v0.7). (Schnell, Bachteler, and Reiher, 2005)
  - String similarity functions penalize deviations between identifiers.
  - Parameters for Jaro-Winkler metric according to prior experience with IAB data (Bachteler, 2008)
  - Blocking on postal codes (about 1100 blocks)
  - Classification of record pairs into links, possible links and non-links (Fellegi and Sunter, 1969)
- Manual matching
  - Independent manual review and classification of possible links by two staff members
  - Supervision and final decision in case of contradicting classifications by third person

Subsequent steps of the linkage process

- Exact matching
  - Records from both data sources are directly matched based on all available fields.
  - Even smallest variations in spelling lead to a rejection of a potential match.
- Probabilistic record linkage
  - Comparison with software Merge ToolBox (MTB, v0.7). (Schnell, Bachteler, and Reiher, 2005)
  - String similarity functions penalize deviations between identifiers.
  - Parameters for Jaro-Winkler metric according to prior experience with IAB data (Bachteler, 2008)
  - Blocking on postal codes (about 1100 blocks)
  - Classification of record pairs into links, possible links and non-links (Fellegi and Sunter, 1969)
- Manual matching
  - Independent manual review and classification of possible links by two staff members
  - Supervision and final decision in case of contradicting classifications by third person

Subsequent steps of the linkage process

- Exact matching
  - Records from both data sources are directly matched based on all available fields.
  - Even smallest variations in spelling lead to a rejection of a potential match.
- Probabilistic record linkage
  - Comparison with software Merge ToolBox (MTB, v0.7). (Schnell, Bachteler, and Reiher, 2005)
  - String similarity functions penalize deviations between identifiers.
  - Parameters for Jaro-Winkler metric according to prior experience with IAB data (Bachteler, 2008)
  - Blocking on postal codes (about 1100 blocks)
  - Classification of record pairs into links, possible links and non-links (Fellegi and Sunter, 1969)
- Manual matching
  - Independent manual review and classification of possible links by two staff members
  - Supervision and final decision in case of contradicting classifications by third person

Subsequent steps of the linkage process

- Exact matching
  - Records from both data sources are directly matched based on all available fields.
  - Even smallest variations in spelling lead to a rejection of a potential match.
- Probabilistic record linkage
  - Comparison with software Merge ToolBox (MTB, v0.7). (Schnell, Bachteler, and Reiher, 2005)
  - String similarity functions penalize deviations between identifiers.
  - Parameters for Jaro-Winkler metric according to prior experience with IAB data (Bachteler, 2008)
  - Blocking on postal codes (about 1100 blocks)
  - Classification of record pairs into links, possible links and non-links (Fellegi and Sunter, 1969)
- Manual matching
  - Independent manual review and classification of possible links by two staff members
  - Supervision and final decision in case of contradicting classifications by third person

Number of observations over all stages of linkage

	N	$\frac{N}{N_c}$	$\frac{N}{N_r}$
CATI respondents (N_r)	10404		100%
Consenting CATI respondents (N_c)	9531	100%	91.61%
Exact matches	5035	52.83%	48.39%
Exact and probabilistic matches (Jaro-Winkler)	7919	83.09%	76.11%
Exact, probabilistic and manual matches	8243	86.49%	79.23%
Matches with administrative spell data	8166	85.68%	78.49%

Data access

- Official data product of the FDZ: "ALWA survey data linked to administrative data of the IAB" (ALWA-ADIAB)
- Access via on-site use and subsequent remote execution at the FDZ (`http://fdz.iab.de/en/FDZ_Individual_Data/ALWA.aspx`)
- BTW: Time lag from completion of linkage to dissemination of data: < 12 months

Outlook

- Update of administrative data in ALWA–ADIAB in 2012
- Further efforts to improve the potential of the data:
    - Considering administrative address data from before and after 2007
        - ⇒ How much observations can be gained by the additional step?
        - ⇒ Does it affect the selectivity of the linked data set?
    - Respondents willing to participate in further panel waves were included in the National Educational Panel Study (NEPS, Allmendinger et al., 2011).
        - ⇒ Non-consenters were asked for consent again.
- Planned methodological analyses:
    - Evaluation of success of aided recall techniques in reducing recall error
    - Validation of educational histories in administrative data and evaluation of existing correction procedures
    - Analysis of social desirability bias in respondent behavior

Thank you for your attention

Manfred Antoni
Contact: Manfred.Antoni@iab.de

References

Jutta Allmendinger et al. (2011). "Adult education and lifelong learning". In: *Zeitschrift für Erziehungswissenschaft* 14.Supplement 2, pp. 283–299

Manfred Antoni, Peter Jacobebbinghaus, and Stefan Seth (2011). *ALWA-Befragungsdaten verknüpft mit administrativen Daten des IAB 1975-2009 (ALWA-ADIAB 7509).* FDZ Datenreport 05/2011

Manfred Antoni and Stefan Seth (2012). "ALWA-ADIAB - linked individual survey and administrative data for substantive and methodological research". In: *Schmollers Jahrbuch. Journal of Applied Social Science Studies* forthcoming

Manfred Antoni. *Linking survey data with administrative employment data: The case of the German ALWA survey,* forthcoming

Manfred Antoni (forthcoming). *Record linkage of the ALWA survey with administrative data of the Federal Employment Agency (ALWA-ADIAB): technical report.* FDZ Methodenreport, forthcoming

Tobias Bachteler (2008). *Dokumentation Record Linkage IEB-PASS.* unpublished document

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

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

Rainer Schnell, Tobias Bachteler, and Jörg Reiher (2005). "MTB: Ein Record-Linkage-Programm für die empirische Sozialforschung". In: *ZA-Information* 56, pp. 93–103