The Pros- and Cons of the IAB individual data. What do we know?

Ralf A. Wilke, University of Nottingham

Motivation

- Since more than 15 years IAB individual data is used for academic research.
  - Increasing number of data products and samples (IABS, IEB, SIAB,...).
  - Increasing number of researchers
  - A vast number of research articles

- In 2004 the Research Data Centre (FDZ) of the IAB was established to
  - Maintain and improve existing data
  - Produce and release new data
  - Make data accessible, ensure data security
  - Consult data users
  - Conduct research on data quality
  - Ease communication with and between data users
  - See **fdz.iab.de**

- General beliefs about the advantages of administrative data:
  - Administrative data are generally considered as more precise than survey data.
  - They contain policy related variables.
  - The size of administrative data makes random sampling errors less relevant and allows detailed analysis for small groups.

- General beliefs about the disadvantages of administrative data:
  - Only limited variable set about the household background and the motivation.
    - Omitted variable bias, lack of "interesting" variables.
  - Data structure and preparation are demanding.

- In this lecture I will discuss various aspects of data quality which shall exceed the common beliefs.
  - Expert audience. Any comments welcome.
  - Apologies if important contributions were overseen while working out this presentation.
- Survey of existing results from many different sources and authors.
- Identify research needs for
  - Data quality
  - Statistical methods

Remarks on the data generating process

- In contrast to many surveys, administrative data is primarily not collected for academic research and statistical analysis.
  - Generated by operations of the public administration.

- Administrative data consists of information:
  - Resulting from operations
  - Collected for statistical analysis but not required by the administration.

- Example: Employment Records (BeH)
  - Records submitted by employers to the public pension insurance.
  - Information is used to determine pension entitlements of employees.
  - Key variables for operations:
    - Employment periods (start and end date)
    - Daily wage
  - Other variables:
    - Household background (married, children, ..)
    - Individual characteristics (education, nationality,...)
    - Job related (full/part time, job title,...)
    - Firm characteristics (industry, location)

- BA data sources:
  - Product of operations. Additional variables added.
  - ASU (BewA)
    - Job seeking periods
  - LeH/LHG
    - Benefit claim periods, level and type of benefit.
  - MTH
    - Training periods, training type.
  - Introduction of various new IT systems lead to a number of inconsistencies and to new names of sources.

Measurement error in Variables

- Hypothesis:

  If information is relevant for or if it is resulting from operations it is more reliable than it was collected for statistical analysis. Information unrelated to operations may contain considerable measurement error.

- No formal proof but a number of papers provide empirical evidence.

- Examples:

- Bernhard/Dressel/Fitzenberger/Schnitzlein/Stephan(2006).
  - Analysis of overlapping spells between various sources.
  - The share of invalid overlaps between various data sources is small (<2.5%, often ~0.5%)
  - Start dates generally more reliable (correct end date).
  - Follow up work: End dates of training programmes sometimes unreliable, correction rules (Waller, 2008).

- Fitzenberger/Osikominu/Voelter (2005).
  - Many inconsistencies in the BeH education variable.
  - Suggest a number of imputation and correction rules.
  - Ready to use Stata code available from the FDZ.

Table 1: Cross tabulation of IP1 (imputed) vs. uncorrected education in the BeH, 20,960,096 spells.

	Education						
IP1	Missing	ND	VT	HS	HSVT	TC	UD
Missing	14.32	.02	.01	.06	.01	.01	.01
ND	24.79	75.12	.27	.68	.10	.05	.01
VT	50.06	23.01	94.51	.05	.03	.01	.01
HS	2.31	.69	.00	73.35	.01	.00	.01
HSVT	4.48	.84	3.46	21.50	87.96	.01	.00
TD	1.83	.18	1.08	1.50	5.77	90.09	.00
UD	2.21	.14	.67	2.86	6.12	9.83	99.96
Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Abbreviations: ND: no degree, VT: completed vocational training, HS: high school degree (Abitur), HSVT: high school degree and complted vocational training, TC: technical college degree, UD: university degree.

Source: Wichert/Wilke (2011)

- Various papers have already made use of the proposed correction rules for the education variable. Examples include:
  - Dustmann/Ludsteck/Schoenberg (2009), Application to wage structure
  - Biewen/Fitzenberger/Osikominu/Paul (2010), Evaluation of training
  - Extensions:
    - Dlugosz (2011) , EM Algorithm for estimation with missing values.
    - Wichert/Wilke (2011), misclassification analysis, side information
  - Studies confirm that imputation and correction rules have a large effect on empirical results.

- Wichert/Wilke (2011)
  - Compare information in BewA and BeH. Period 1999-2004.
  - Focus on education and nationality
  - Suggest a correction rule for nationality: migration background
  - Detect much less inconsistencies in BewA than in the BeH:
    - 6% vs. 20% in the case of the education variable
  - Pooling of educational categories reduces measurement error.

Table 2: Misclassification matrix for education (uncorrected).

Education	BewA						
BeH	Missing	ND	VT	HS	HSVT	TD	UD
Missing	48.65	43.67	30.80	43.43	29.97	25.97	26.26
ND	16.66	32.61	10.64	19.48	7.24	3.86	3.78
VT	28.50	22.76	56.70	18.25	39.37	23.36	17.64
HS	.30	.34	.25	10.19	3.14	2.49	2.97
HSVT	1.63	.35	.86	4.19	11.14	7.39	5.44
TD	2.24	.13	.55	1.83	4.76	22.52	6.24
UD	2.02	.14	.21	2.63	4.37	14.41	37.67
Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Source: Wichert/Wilke (2011)

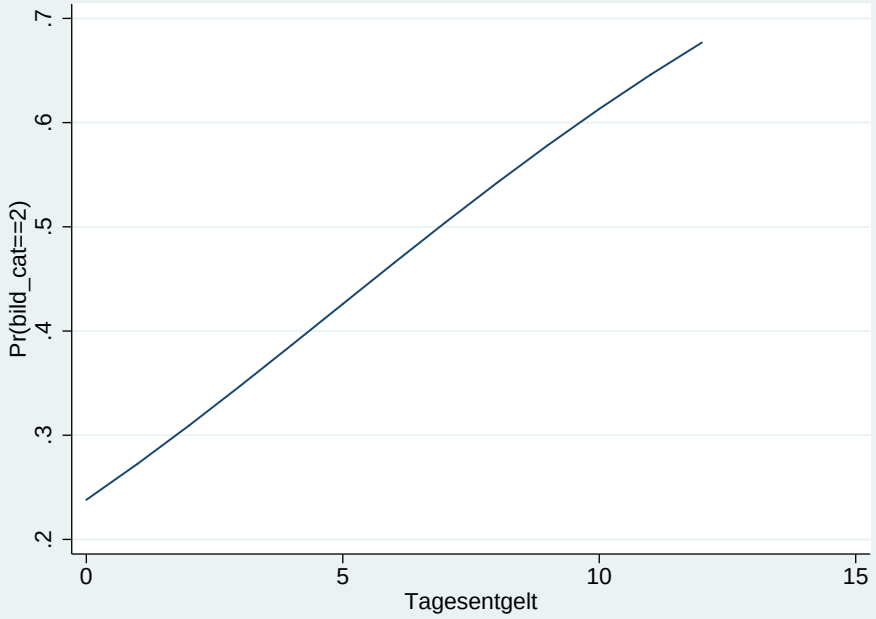
Table 3: Misclassification matrix for IP1 (imputed).

IP1	BewA						
BeH	Missing	ND	VT	HS	HSVT	TD	UD
Missing	9.96	5.01	1.62	7.73	1.94	2.99	3.30
ND	23.29	39.93	7.45	19.92	5.75	2.35	2.60
VT	54.13	51.79	84.80	27.53	38.67	14.88	11.52
HS	.17	.63	.14	14.87	3.68	2.45	3.08
HSVT	4.00	1.90	3.69	20.10	29.35	10.94	7.63
TD	4.13	.38	1.61	4.10	10.30	36.15	7.44
UD	4.33	.36	.68	5.75	10.30	30.25	64.43
Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00

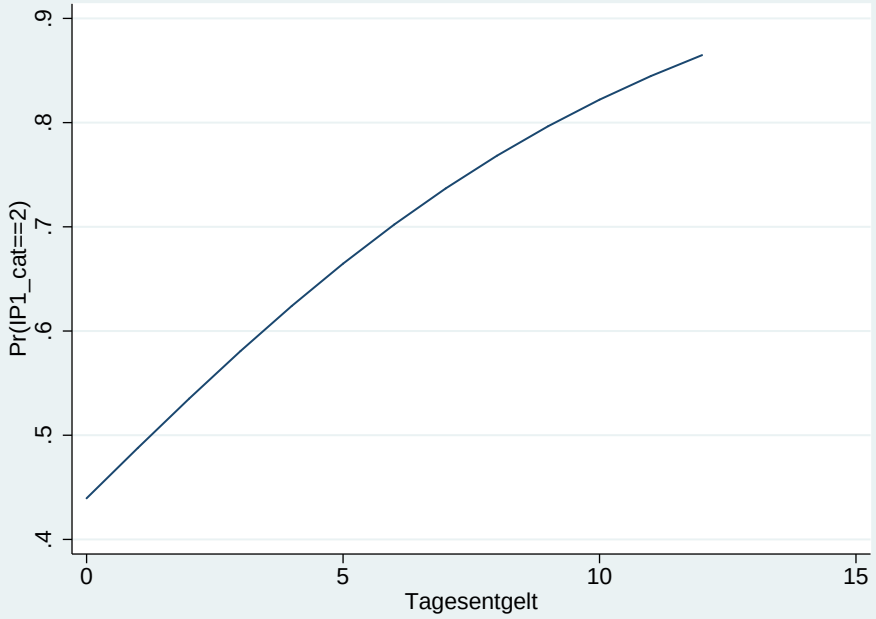
Source: Wichert/Wilke (2011)

- Probability of BeH educ = BewA edu varies strongly with other characteristics.

university degree, original



university degree, IP1



Source: Dlugosz/Mammen/Wilke (2011)

Table 6: Misclassification matrix for nation (uncorrected).

Nation	BewA		
BeH	Missing	German	non German
Missing	92.70	.03	.15
German	6.29	98.65	27.48
non German	1.01	1.31	72.37
Total	100.00	100.00	100.00

Source: Wichert/Wilke (2011)

- BewA information is less erroneous for these variables...
  - because they matter for the job search process...

  ...but not free of error.
  - How reliable is it?

- Recent action by the FDZ:
  - Scioch (2010b): use patent data to validate education information in IAB data.
    - Better validation source. Patent data free of error?
    - Imputation rules reduce the measurement error in the data.

- Required action on measurement errors:

- Availability of ready to use packages
  - A number of statistical approaches have been developed over the past years but
    - A lack of ready to use implementations.
    - Often too slow for practical analysis with large data.
- Analyse more IAB variables in terms of measurement error
  - Attempt to classify variables in the FDZ data.
- Ready to use data correction rules (not just for education and nationality)
  - Code provided or data cleaned by the FDZ.

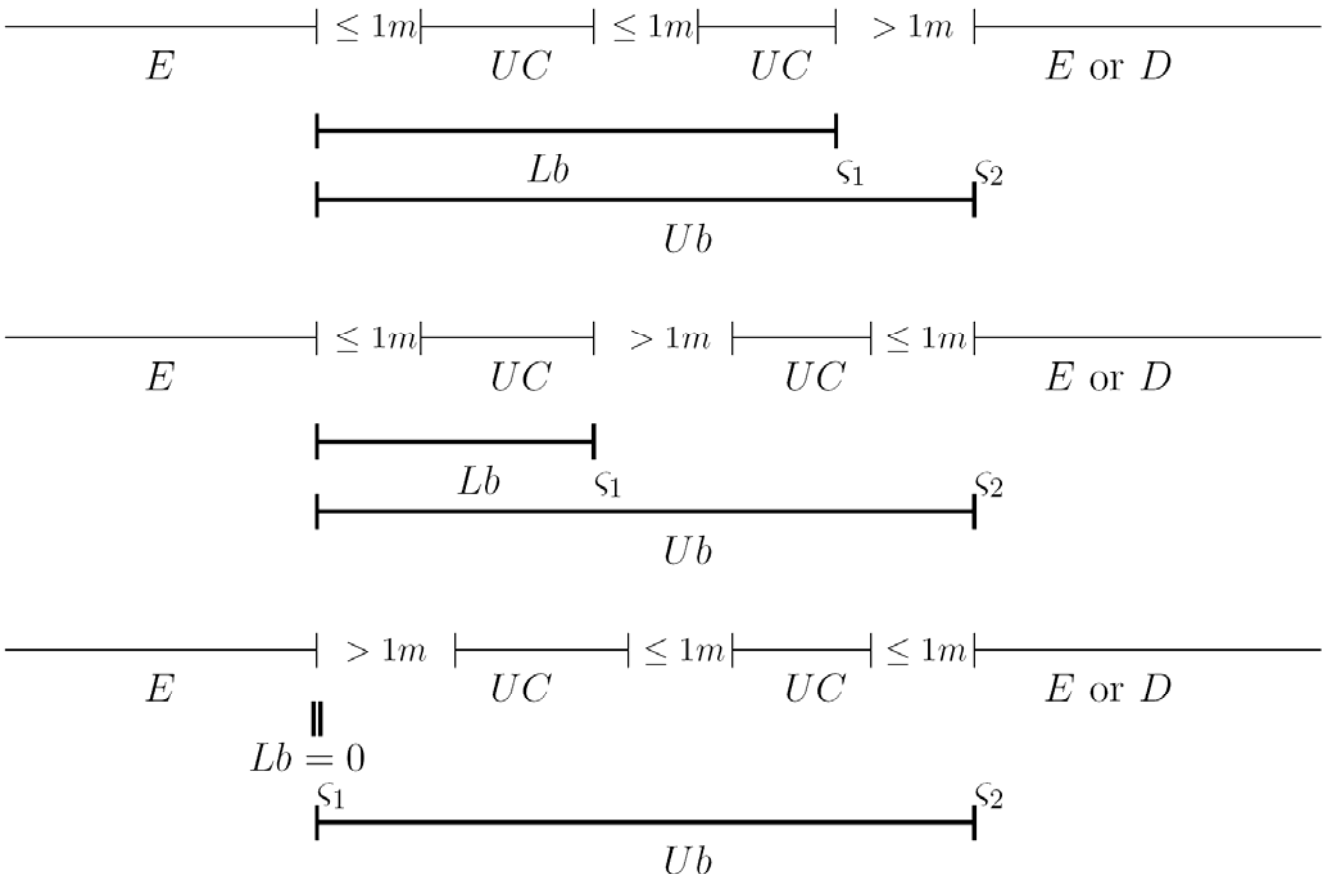
Partial identification of duration

- Start and end dates of spell information are often correct, but this does not make the complete duration of being in a labour market state, because...
  - Subsequent spells may need to be appended.
  - Overlapping spells
  - Due to competing risks structure (just the first termination reason is observed).
  - Due to unobserved periods in individual employment trajectories (can correspond to more than one labour market state),e.g. unemployment, out of the labour force.

- Examples:
  - Employment duration (Bookmann/Steffes, 2005)
    - Unobserved periods
    - LIAB
  - Training/ALMP durations (Waller, 2008)
    - Taking into account parallel spell information for the correction of end dates.
    - IEB

- Examples: Attempts to define unemployment duration
  - IABS:
    - Fitzenberger/Wilke (2010), Lee/Wilke (2009), Arntz/Lo/Wilke (2008)
    - No overlap between BeH and LeH.
  - IEB:
    - Mueller/Kruppe/Wichert/Wilke (2008), Kruppe (2009)
    - Lo/Stephan/Wilke (2011)
    - Mueller et al. (2008) develop 30 different implementation of unemployment duration. Ready to use Stata code available from the FDZ.
    - Large differences in unemployment duration and number of preriods.

Figure 2: Examples of Lb and Ub definitions of unobserved unemployment durations



Note: E: Spell of local employment; D: Spell of distant employment; UC: Spell receiving UC; m: Month
ς_1: Time at which unobserved period begins; ς_2: Time at which unobserved period ends

Source: Arntz/Lo/Wilke (2008)

Table 4: Compositions of unemployment durations with different destination states under *Lb* and *Ub* definitions, IAB data, final sample

	Control group		Treatment group	
	pre-1997	post-1997	pre-1997	post-1997
Lb unemployment spells				
ends with local employment	50.1%	50.8%	48.1%	49.4%
ends with distant employment	8.7%	10.4%	8.9%	10.2%
ends with unknown state	41.2%	38.8%	43.0%	40.4%
right-censored	-	-	-	-
total	100.0%	100.0%	100.0%	100.0%
Ub unemployment spells				
ends with local employment	75.2%	71.1%	72.5%	69.6%
ends with distant employment	14.1%	15.2%	14.2%	14.7%
ends with unknown state	-	-	-	-
right-censored	10.7%	13.7%	13.3%	15.7%
total	100.0%	100.0%	100.0%	100.0%
Total spells	104,069	94,309	39,434	36,104

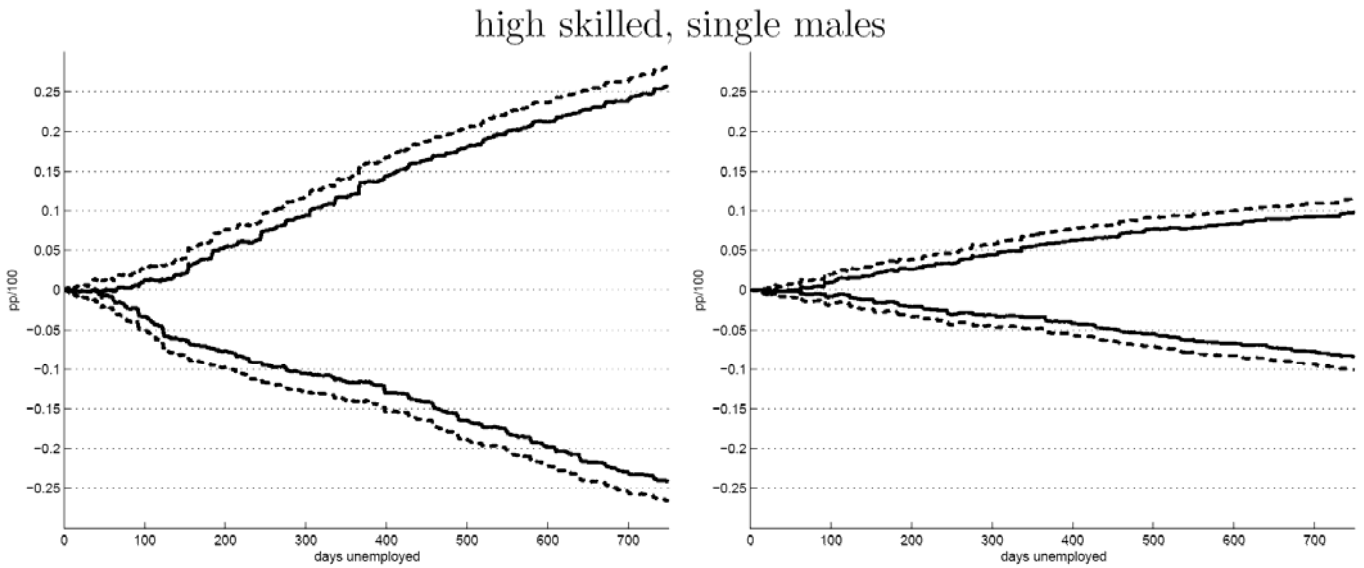
Source: Arntz/Lo/Wilke (2008)

Table 5: Median unemployment duration (days) by sub-sample and definition of unemployment, IAB data, final sample

	Control group		Treatment group	
	pre-1997	post-1997	pre-1997	post-1997
Lb unemployment spells				
high-skilled singles	274	178	364	214
less-skilled singles	243	183	303	214
high-skilled married men	184	123	233	165
less-skilled married men	159	134	183	154
overall:	185	152	214	172
Ub unemployment spells				
high-skilled singles	488	278	648	369
less-skilled singles	336	243	449	291
high-skilled married men	367	246	434	305
less-skilled married men	211	177	245	200
overall:	258	204	307	232

Source: Arntz/Lo/Wilke (2008)

Figure 3: Lower and upper bound of the DiD changes of the cumulative incidence of local (left) and distant (right) employment among selected groups.



Source: Arntz/Lo/Wilke (2008)

- Random sampling much less relevant problem.

- Required action on bounds for durations:
  - Provide ready to use code for
    - Different data structures
    - Different labour market states
    - Competing risks.
  - Ready to use implementations for estimation of bounds.
  - Link with survey data to identify unobserved periods or derive probabilities.
    - PASS?

Remarks on omitted variables

- Data does not contain a lot of information on
  - Household background and social environment.
  - Motivation and ability.
- Unavailability of information may render analysis impossible.
- Reduces the set of variables in multivariate analysis.
  - This does not necessarily cause problems.
  - When does it affect our results?
    - If omitted variables are correlated with important observed variables and if they are important for the analysis.

- What is done:
  - Use proxy variables.
    - Motivation or the "type" of an individual may be proxied by information from the work history.
    - Examples: Previous unemployment experience, previous failure to report to the employment agency (Biewen/Fitzenberger/ Osikominu/Paul, 2010), previous wage dynamics.
  - Problem: Not much empirical evidence for the feasibility of the proxies.

- Required action on omitted variables:
  - Use survey data (linked or not linked) to examine the correlation structure between
    - variables which are in the IAB data and which are omitted.
    - the proxies and the omitted variables.
  - Use a subset of the IAB data which contains more variables to examine the correlation structure and the effect of the variables.
    - ALG II sources, PASS, IAB ALWA.
    - May not be a random sample.

An attempt to identify major issues…

- Measurement error
  - Important in a subset of variables.
  - This set is still to be identified.
- Partial identification of durations
  - Important.
- Omitted variables
  - Importance unknown.
- Random sampling error
  - Often much less an issue than the other points.
  - Sometimes hard to obtain relevant statistics if data are large.

Thank you for your attention!

Appendix:

- A list of further known issues with IAB data
- References

A list of known issues…

- Missing Values
  - A large share of missing values in several variables such as education, marital status etc.
  - Solution: correction rules, imputation, EM-Algorithm
  - Dlugosz (2011), Buettner/Raessler (2008)

- Firm identifier
  - Branch not firm
  - May change due to merger
  - Affects determination of job changes, length of tenure, firm creation and destruction
  - Hethey/Schmieder (2010)

- Incomplete LeH during the 1970s
  - May be better to use data after 1980
  - Bender/Hilzendegen/Rohweder/Rudolph (1996)

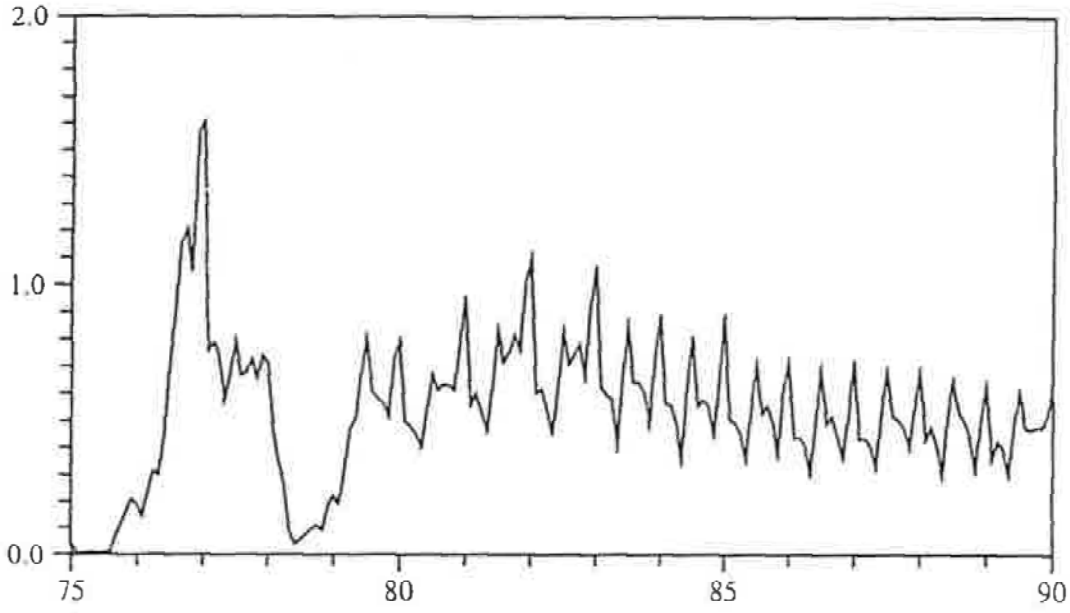


Abb. 8.4.4 Verteilung der Anfangsmonate der jeweils ersten (pro Person) Meldung über Leistungen aus der Arbeitslosenversicherung (N = 142,116). Ordinate: Prozentanteile.

Source: Bender et al. (1996)

- Cleansing procedures for overlapping spell information
  - Bernhard/Dressel/Fitzenberger/Schnitzlein/Stephan(2006) note that most overlaps in the data are feasible (no data problems).
  - Automatic decision rules to reduce information to one source.
  - Scioch (2010a), Wunsch/Lechner (2008)

- Wage variable:
  - Topcoding, right censoring (Buettner/Raessler, 2008)
  - Structural break in 1983/1984:
    - Before: permanent daily wage income
    - After: includes one off payment and bonuses
    - Bender/Hilzendegen/Rohweder/Rudolph (1996)
    - Correction rule suggested by Fitzenberger (1998) and sketched by Dustmann/Ludsteck/Schoenberg (2009).
  - Open questions:
    - How informative is the daily wage provided that working hours are unknown and part/full time variable may contain measurement error?
    - Many observations with zero or almost zero daily wage. What is the interpretation of this?

- Other problems include:
  - LeH information may not contain ex post changes which have been implemented in the employment offices. This includes for example sanction withdrawals if further evidence was provided by the unemployed.
    - No reference.
  - The MTH information on training periods is sometimes missing or wrong.
    - Reference?

References

- Arntz/Lo/Wilke (2008) Bounds analysis of competing risks : a nonparametric evaluation of the effect of unemployment benefits on migration in Germany, FDZ Methodenreport 06/008, Institut für Arbeitsmarkt- und Berufsforschung (IAB), Nürnberg.
- Bender/Hilzendegen/Rohweder/Rudolph (1996) Die IAB-Beschäftigtenstichprobe 1975-1990. (Beiträge zur Arbeitsmarkt- und Berufsforschung, 197), Nürnberg.
- Bernhard/Dressel/Fitzenberger/Schnitzlein/Stephan(2006) Überschneidungen in der IEBS: Deskriptive Auswertung und Interpretation (Overlapping spells in the IEBS: descriptives and interpretation), FDZ Methodenreport , 04/2006, Institut für Arbeitsmarkt- und Berufsforschung (IAB), Nürnberg.
- Biewen/Fitzenberger/Osikominu/Paul (2010) The Comparative Effectiveness of Public Sponsored Training Revisited: The Merits of Using Rich Administrative Data, previous version: IZA Discussion Paper No. 2885, 2007.
- Bookmann/Steffes (2005) Individual and Plant-level Determinants of Job Durations in Germany, ZEW Discussion Paper No. 05-89, Mannheim.
- Buettner/Raessler (2008) Büttner, T. and Rässler, S. (2008) Multiple imputation of right-censored wages in the German IAB Employment Sample considering heteroscedasticity. IAB Discussion Paper 44/2008, IAB, Nürnberg.
- Dlugosz (2011) Give Missings a Chance: Combined stochastic and rule-based approach to improve regression models with mismeasured monotonic covariates without side information, paper presented at the IAB-FDZ Nutertagung 2011, Nürnberg.
- Dlugosz/Mammen/Wilke (2011) ongoing work.

- Dustmann/Ludsteck/Schoenberg (2009) Revisiting the German Wage Structure, The Quarterly Journal of Economics, MIT Press, vol. 124(2), pages 843-881, May.
- Fitzenberger (1998) ,Wages and employment across skill groups: An analysis for West Germany, Habilitaion thesis, University of Konstanz.
- Fitzenberger/Osikominu/Voelter (2005) Imputation Rules to Improve the Education Variable in the IAB Employment Subsample, Journal of Applied Social Science Studies (Schmollers Jahrbuch) 126 (3), 2006, 405-436.
- Fitzenberger/Wilke (2010) Unemployment Durations in West-Germany before and after the reform of the Unemployment Compensation System during the 1980s. German Economic Review, 11(3), 336-366.
- Hethey/Schmieder (2010) Using worker flows in the analysis of establishment turnover : evidence from German administrative data, FDZ Methodenreport 06-2010, Institut für Arbeitsmarkt- und Berufsforschung (IAB), Nürnberg.
- Kruppe (2009) Empirical consequences of definitions , the case of unemployment in German register data, Historical Social Research Vol. 34, No. 3.
- Lee/Wilke (2009) Reform of Unemployment Compensation in Germany: A Nonparametric Bounds Analysis using Register Data. Journal of Business and Economic Statistics, 27(2), 176-192.
- Lo/Stephan/Wilke (2011) ongoing work.
- Mueller/Kruppe/Wichert/Willke (2008) On the Definition of Unemployment and its Implementation in Register Data - The Case of Germany. Journal of Applied Social Science Studies, 128(3), 461-488.

- Scioch (2010a) The impact of cleansing procedures for overlaps on estimation results evidence for German administrative data. FDZ Methodenreport, 04/2010, Institut für Arbeitsmarkt- und Berufsforschung (IAB), Nürnberg.
- Scioch (2010b) Quality and Quantity – Using Administrative Data for Scientific Purposes in Labor Market Research, poster presentation, German Statistical Week, September 2010, Nürnberg.
- Waller (2008) On the Importance of Correcting Reported End Dates of Labor Market Programs, Schmollers Jahrbuch (Journal of Applied Social Science Studies) 128, 213-236.
- Wichert/Wilke (2011) Which Factors safeguard Employment? An Analysis with Misclassified German Register Data. Journal of the Royal Statistical Society: Series A, (forthcoming).
- Wunsch/Lechner (2008) What Did All the Money Do? On the General Ineffectiveness of Recent West German Labour Market Programmes, Kyklos, 61(1),134-174.