RESEARCH DATA CENTRE (FDZ)
of the German Federal Employment Agency (BA)
at the Institute for Employment Research (IAB)

FDZ-METHODENREPORT

Methodological aspects of labour market data

01|2025 EN Creating cross-sectional data and biographical variables with the Sample of Integrated Labour Market Biographies 1975-2023 - Programming examples for Stata

Philipp vom Berge, Alexandra Schmucker

Bundesagentur für Arbeit

Creating cross-sectional data and biographical variables with the Sample of Integrated Labour Market Biographies 1975-2023 - Programming examples for Stata

Philipp vom Berge (IAB), Alexandra Schmucker (IAB)

Die FDZ-Methodenreporte befassen sich mit den methodischen Aspekten der Daten des FDZ und helfen somit Nutzerinnen und Nutzern bei der Analyse der Daten. Nutzerinnen und Nutzer können hierzu in dieser Reihe zitationsfähig publizieren und stellen sich der öffentlichen Diskussion.

FDZ-Methodenreporte (FDZ method reports) deal with methodical aspects of FDZ data and help users in the analysis of these data. In addition, users can publish their results in a citable manner and present them for public discussion.

Contents

List of the Appendix

Zusammenfassung

Der vorliegende FDZ-Methodenreport (einschließlich der Programmierbeispiele für Stata) beschreibt die Erstellung von Querschnittdaten auf Basis der Stichprobe der Integrierten Arbeitsmarktbiografien (Version 1975-2023) zu frei wählbaren Stichtagen. Zudem wird die Generierung von biografischen Merkmalen erläutert.

Abstract

This FDZ-Methodenreport (including Stata code examples) outlines an approach to construct cross-sectional data at freely selectable reference dates using the Sample of Integrated Labour Market Biographies (version 1975-2023). In addition, the generation of biographical variables is described.

Keywords

Sample of Integrated Labour Market Biographies (SIAB), data preparation, cross-sectional data, data management

Note

This FDZ-Methodenreport and the attached programs are updates of the FDZ-Methodenreport 03/2023 (vom Berge and Schmucker 2023) which was written for use with SIAB version 1975-2021 (DOI: 10.5164/IAB.SIAB7521.de.en.v1). This update describes the programs suitable for SIAB version 1975-2023 (DOI: 10.5164/IAB.SIAB7523.de.en.v1).

1 Introduction

In this report, which includes example programs for Stata, we demonstrate an approach to prepare SIAB data. The paper provides examples of data reorganization techniques that simplify the data structure. The main purpose is to facilitate the use of the SIAB dataset, especially for researchers who do not have a lot of experience in analyzing spell data. This FDZ-Methodenreport and the attached programs are updates of the FDZ-Methodenreport 03/2023 (vom Berge and Schmucker 2021) which was written for use with SIAB version 1975-2021 (Schmucker et al. 2023). This update describes the programs suitable for SIAB version 1975-2023 (Schmucker and vom Berge 2025). In the new version, the program regarding one-time payments is no longer included, as SIAB 7523 now already contains a corresponding variable.

The main goal of the data preparation shown here is to create cross-sectional data sets at freely selectable reference dates. Also, we simplify the SIAB data structure by keeping only one 'main' observation per person and date. (In the original SIAB data, in contrast, there may be concurrent information for a given period.) We mitigate the drawback of this procedure – the loss of information – by generating biographical variables such as the number of days in employment or the date of entry into the current job. Finally, we show how to create variables that preserve information on parallel periods before deleting them from the data set.

The Stata do-files provided are examples which can be adjusted to specific user needs. The files have been developed using the Sample of Integrated Labour Market Biographies 7523. Since individual micro data provided by the FDZ are standardized, the files can be used for other FDZ data products as well with minor modifications.

The steps presented in the following sections merely simplify the data. We do not consider any techniques to improve data quality or to impute missing values. For this we refer the reader to existing volumes in the series FDZ-Methodenreporte. A description of comprehensive preparing steps for SIAB version 7521 can also be found in Stüber et al. 2023.

2 Outline of the Stata do-files

Starting with a longitudinal data set, the accompanying Stata do-files generate several biographical variables, add one-time payments to the regular wages and create cross-sectional data sets at freely selectable points in time. They are structured as follows.

master.do: This do-file defines Stata macros for directories, file names, variables on individual and establishment identifiers and reference dates. Users will have to customize the macros accordingly. It calls the do-files that create the biographical variables and cross-sectional data. Temporary data sets are deleted at the end of the program run.

01_SIAB_bio.do: This do-file uses the longitudinal data to create the biographical variables (see chapter 3) from the longitudinal data. Durations are calculated based on the end date of each observation. A temporary data set (siab_7521_v1_bio.dta) is

saved at the end of the routine which comprises all information from the input data plus the biographical variables.

03_SIAB_quer.do: This program first creates variables indicating labor market statuses recorded at the same time as the identified main observation for each reference date (cf. chapter 4). Note that this do-file uses the data set which was created by 01_SIAB_bio.do. Second, all observations except the main observations are deleted so that there is only one observation per person and reference date. If episodes with one-off payments have not been deleted before, they could also appear as main observations. Next, the duration variables calculated in 01_SIAB_bio.do are cut off at the selected reference date. Finally, a data set is saved for each of the selected reference dates. If required, it would then be possible to construct a panel data set by linking the separate files by person ID (${pers_id}) and reference date (stichtag).

Disclaimer: The attached Stata do-files have been tested with SIAB 7522 v1 using Stata 17. Before using them with data products other than SIAB 7523, users should consult the respective FDZ data documentation to make sure that this is appropriate. This is especially important when dealing with "non-7523" data, that is, data not covering the years 1975-2023, because the meaning of underlying variables might differ.

The data version, as well as the data documentation can be obtained from the respective FDZ-Datenreport. The FDZ does not guarantee that the specifications chosen in the provided codes can be applied to all research interests. We strongly advise users to check if the specifications can be transferred to their research project before adopting the routines.

Users who are unfamiliar with processing longitudinal data of the IAB may consult the FDZ-Methodenreport 6/2007 (Drews et al. 2007, only available in German). For a general introduction to data analysis with Stata we recommend Kohler and Kreuter (2016 and 2012).

3 Generated biographical variables

3.1 First day in employment (ein_erw)

Category	Description
Variable label	First day in employment
Variable name	ein_erw
Category	Generated biographical variables
Origin	Generated from BeH
Data type	Date
Hierarchy	None
Detailed description	This variable specifies the start date of the first employment subject to social security or the first marginal employment. Training periods are disregarded (employment statuses 102, 121, 122, 141). Second jobs during periods of vocational training are considered though. Persons who pass a training period but do not have any employment covered by the social security system are assigned a missing value throughout. Episodes prior to the first employment subject to social security or marginal employment are also set to missing. Episodes with one-time payments (reason of notification = 154) are not considered.

Category	Description
	The start date of first employment (ein_erw) might occur a long time after the first day in establishment (ein_bet) and the first day in job (ein_job) because in the latter cases training periods are included.
Notes on quality	For West Germany the variable is left censored on 1.1.1975. For East Germany the censoring is not that unambiguous. Entries are definitely censored on 1.1.1990, but entries on 1.1.1991 and 1.1.1992 may also be affected because many employment notifications for 1990 and 1991 are missing.

3.2 Number of days in employment (tage_erw)

Category	Description
Variable label	Number of days in employment
Variable name	tage_erw
Category	Generated biographical variables
Origin	Generated from BeH
Data type	Date
Hierarchy	None
Detailed description	This variable sums up the number of days a person has been employed up to the end date of the current observation. For the cross-sections, the duration is cut off at the respective reference date. Training periods (employment statuses 102, 121, 122, 141) are excluded. If an individual was in training throughout, the variable has a value of 0. Episodes with one-time payments (reason of notification = 154) are not considered.
Notes on quality	For West Germany the variable is left censored on 1.1.1975. For East Germany the censoring is not that unambiguous. Entries are definitely censored on 1.1.1990, but entries on 1.1.1991 and 1.1.1992 may also be affected because many employment notifications for 1990 and 1991 are missing.

3.3 First day in establishment (ein_bet)

Category	Description
Variable label	First day in establishment
Variable name	ein_bet
Category	Generated biographical variables
Origin	Generated from BeH
Datatype	Date
Hierarchy	none
Detailed description	This variable indicates the start date of the first employment in the current establishment. This might also be a training period but not an episode with one-time payments. An interruption of the employment in the establishment does not change the start date, i. e. it is constant for each combination of person number and establishment number. In the case of a missing or invalid establishment number, the variable contains a missing value. The start date of first employment (ein_erw) can occur a long time after the first day in establishment (ein_bet) and the first day in job (ein_job) because in the latter cases training periods are included.
Notes on quality	For West Germany the variable is left censored on 1.1.1975. For East Germany the censoring is not that unambiguous. Entries are definitely censored on 1.1.1990, but entries on 1.1.1991 and 1.1.1992 may also be affected because many employment notifications for 1990 and 1991 are missing.

3.4 Number of days in establishment (tage_bet)

Category	Description
Variable label	Number of days in establishment
Variable name	tage_bet
Category	Generated biographical variables
Origin	Generated from BeH
Datatype	Numerical
Hierarchy	none
Detailed description	The variable indicates the number of days a person has been working in the establishment until the end date of the episode. For the cross-sections, the duration is cut off at the respective reference date. Training periods in the establishment are included. Employment gaps are not included, but all periods of employment in the respective establishment are added up. Episodes with one-time payments (reason of notification = 154) are not considered. If the number of days in the establishment was alternatively calculated as the interval between the first day in the establishment (ein_bet) and the end date of the episode (or the cutoff date), the values obtained might be larger than tage_bet because tage_bet does not include employment interruptions.
Notes on quality	For West Germany the variable is left censored on 1.1.1975. For East Germany the censoring is not that unambiguous. Entries are definitely censored on 1.1.1990, but entries on 1.1.1991 and 1.1.1992 may also be affected because many employment notifications for 1990 and 1991 are missing.

3.5 First day in job (ein_job)

Category	Description
Variable label	First day in job
Variable name	ein_job
Category	Generated biographical variables
Origin	Generated from BeH
Datatype	Numerical
Hierarchy	none
Detailed description	This variable indicates the start date of the first employment notification in the current job. Training periods (employment statuses 102, 121, 122, 141) in the same establishment are classified as separate jobs, even if they follow directly or are followed directly by a job in the same establishment. An employment in the same establishment after a gap is considered a new job if the reason for notification of the previous employment record indicates the termination of this job (reasons for notification 30, 34, 40, 49) and the gap is longer than 92 days or the reason for notification of the previous employment record does not indicate the termination of this job, but the gap is longer than 366 days. Episodes with one-time payments (reason of notification = 154) are not considered but the start date of the job is assigned to them. The first day in new job (ein_job) cannot be earlier than the first day in establishment (ein_bet). It can however be earlier than the first day in employment (ein_erw), because the latter does not include training periods.
Notes on quality	For West Germany the variable is left censored on 1.1.1975. For East Germany the censoring is not that unambiguous. Entries are definitely censored on 1.1.1990, but entries on 1.1.1991 and 1.1.1992 may also be affected because many employment notifications for 1990 and 1991 are missing.

3.6 Numbers of days in job (tage_job)

Category	Description
Variable label	Numbers of days in job
Variable name	tage_job
Category	Generated biographical variables
Origin	Generated from BeH
Datatype	Numerical
Hierarchy	none
Detailed description	The variable indicates the number of days a person has been working in the current job until the end date of the episode. For the cross-sections, the duration is cut off at the respective reference date. Training periods (employment statuses 102, 121, 122, 141) in the same establishment are treated as separate jobs, even if they follow directly or are followed directly by a job in the same establishment. An employment in the same establishment after a gap is considered a new job if the reason for notification of the previous employment record indicates the termination of that job (reasons for notification 30, 34, 40, 49) and the gap is longer than 92 days or the reason for notification of the previous employment does not indicate the end of the last job, but the gap is longer than 366 days. Durations of episodes with one-time payments (reason of notification = 154) are not considered but the durations of parallel employment episodes in the same establishment are assigned to them. If the number of days in the current job was alternatively calculated with the first day in job variable (ein_job), the values obtained might be larger than tage_job because tage_job does not include employment interruptions.
Notes on quality	For West Germany the variable is left censored on 1.1.1975. For East Germany the censoring is not that unambiguous. Entries are definitely censored on 1.1.1990, but entries on 1.1.1991 and 1.1.1992 may also be affected because many employment notifications for 1990 and 1991 are missing.

3.7 Number of benefit receipts (anz_lst)

Category	Description
Variable label	Number of benefit receipts
Variable name	anz_lst
Category	Generated biographical variables
Origin	Generated from LEH/LHG
Datatype	Numerical
Hierarchy	none
Detailed description	The variable gives the number of episodes a person has been in benefit receipt up to the end date of the current observation. The variable includes both Social Code II and Social Code III benefits. Hence, the meaning of the variable changes in 2005 with the inclusion of Social Code II benefits. The variable is not incremented if a benefit receipt spell is interrupted by a period of less than 10 days or if the type of benefit changes.
Notes on quality	For West Germany the variable is left censored on 1.1.1975. For East Germany the censoring is not that unambiguous. Entries are definitely censored on 1.1.1990, but entries on 1.1.1991 and 1.1.1992 may also be affected because many employment notifications for 1990 and 1991 are missing.

3.8 Number of days of benefit receipts (tage_lst)

Category	Description
Variable label	Number of days of benefit receipt
Variable name	tage_lst
Category	Generated biographical variables
Origin	Generated from LEH/LHG
Datatype	Numerical
Hierarchy	none
Detailed description	The variable gives the number of days a person has been in benefit receipt up to the end date of the current observation. For the cross-sections, the duration is cut off at the respective reference date. The variable includes both Social Code II and Social Code III benefits are treated the same. Hence, the meaning of the variable changes in 2005 with the inclusion of Social Code II benefits. It is possible that a person is employed (employment subject to social security or marginal part-time employment) and receives benefits at the same time. In this case, tage_lst still counts the benefit receipts spells.
Notes on quality	For West Germany the variable is left censored on 1.1.1975. For East Germany the censoring is not that unambiguous. Entries are definitely censored on 1.1.1990, but entries on 1.1.1991 and 1.1.1992 may also be affected because many employment notifications for 1990 and 1991 are missing.

4 Generated variables for parallel statuses at the reference date

4.1 Type of second job (nb)

Category	Description
Variable label	Type of second job
Variable name	nb
Category	Generated biographical variables
Origin	Generated from BeH
Datatype	Numerical
Hierarchy	None
Detailed description	The variable indicates whether there is a secondary employment at the reference data and specifies the type of that employment. Only one secondary employment is taken into account. Information on any further parallel employment relationships are discarded. The variable distinguishes between full-time and part-time employment. Marginal part-time employment has been recorded since 1999. Any secondary employment relationships that show no valid information regarding the variables „employment status“ or „full-time / part-time employment“ and therefore cannot be classified as either full-time, part-time or marginal part-time employees are coded as „secondary employment not specified“. The variable is missing for persons that did not have a second employment relationship. Values and Labels: 1 full-time job 2 part-time job 3 marginal part-time job 4 not specified second job

Category	Description
Notes on quality	There is a considerable increase in the number of missing values in the variable “full-time / part-time employment” in 2011 due to the change in the reporting procedure. In order to reduce this problem, the working hours were imputed at the IAB for the period in question. Further information about the procedure can be found in Ludsteck/Thomsen (2016).

4.2 Secondary employment in the same establishment as main occupation (nb_betr)

Category	Description
Variable label	Secondary employment in the same establishment as main occupation
Variable name	nb_betr
Category	Generated biographical variables
Origin	Generated from BeH
Datatype	Numerical
Hierarchy	None
Detailed description	This variable indicates if the secondary employment at the reference date (see variable nb) is in the same establishment as the main occupation. If there is no valid establishment number for the primary or secondary occupation, the variable is set to missing. Values and Labels: 0 other establishment 1 same establishment
Notes on quality	-

4.3 Parallel observation: LeH (leh)

Category	Description
Variable label	Parallel observation: LeH
Variable name	leh
Category	Generated biographical variables
Origin	Generated from LeH
Datatype	Numerical
Hierarchy	None
Detailed description	This variable indicates if there is a parallel observation from the LeH at the reference date.
Notes on quality	-

4.4 Parallel observation: (X)ASU (asu)

Category	Description
Variable label	Parallel observation: (X)ASU
Variable name	asu
Category	Generated biographical variables
Origin	Generated from ASU/XASU
Datatype	Numerical
Hierarchy	None
Detailed description	The variable indicates if in addition to the main observation an observation from the Job-Search History File (ASU) or the Job-Search History File by XSozial (XASU) is present at the respective reference date.
Notes on quality	-

4.5 Parallel observation: LHG (lhg)

Category	Description
Variable label	Parallel observation: LHG
Variable name	lhg
Category	Generated biographical variables
Origin	Generated from LHG
Datatype	Numerical
Hierarchy	None
Detailed description	The variable indicates if there is a parallel observation from the Unemployment Benefit II Recipient History at the reference date.
Notes on quality	-

4.6 Parallel observation: MTH (mth)

Category	Description
Variable label	Parallel observation: MTH
Variable name	mth
Category	Generated biographical variables
Origin	Generated from MTH
Datatype	Numerical
Hierarchy	None
Detailed description	The variable indicates if there is a parallel observation from the Participants-in-Measures History File (MTH) at the respective reference date.
Notes on quality	-

4.7 Total income, all sources (gtentgelt)

Category	Description
Variable label	Total income, all sources
Variable name	gtentgelt
Category	Generated biographical variables
Origin	Generated from BeH/LeH
Datatype	Numerical
Hierarchy	None
Detailed description	This variable contains the sum of all income from employment notifications and benefit receipt observations at the respective reference date.
Notes on quality	-

4.8 Cutoff date of the cross section (stichtag)

Category	Description
Variable label	Cutoff date of the cross section
Variable name	stichtag
Category	Generated technical variables
Origin	Generated
Datatype	Date
Hierarchy	None
Detailed description	This variable gives the date of the respective reference date for which the cross-sectional data was created.
Notes on quality	-

References

Drews, Nils; Groll, Dominik; Jacobebbinghaus, Peter (2007): Programmierbeispiele zur Aufbereitung von FDZ Personendaten in STATA. FDZ-Methodenreport, 06/2007

Drews, Nils (2006): Qualitätsverbesserung der Bildungsvariable in der IAB-Beschäftigtenstichprobe 1975-2001. FDZ-Methodenreport, 05/2006

Fitzenberger, Bernd; Osikominu, Aderonke; Völter, Robert (2005): Imputation rules to improve the education variable in the IAB employment subsample. FDZ-Methodenreport, 03/2005

Gartner, Hermann (2005): The imputation of wages above the contribution limit with the German IAB employment sample. FDZ-Methodenreport, 02/2005

Kohler, Ulrich; Kreuter, Frauke (2016): Datenanalyse mit Stata: allgemeine Konzepte der Datenanalyse und ihre praktische Anwendung. 5. Auflage. Berlin/Boston: De Gruyter Oldenbourg

Kohler, Ulrich; Kreuter, Frauke (2012): Data Analysis Using Stata. Third Edition. Stata Press

Kruppe, Thomas; Müller, Eva; Wichert, Laura; Wilke, Ralf A. (2007): On the definition of unemployment and its implementation in register data * the case of Germany. FDZ-Methodenreport, 03/2007

Ludsteck, Johannes; Thomsen, Ulrich (2016): Imputation of the Working Time Information for the Employment Register Data. FDZ Methodenreport, 01/2016 (en)

Schmucker, Alexandra; Seth, Stefan; vom Berge, Philipp (2023): Sample of Integrated Labour Market Biographies (SIAB) 1975-2021. FDZ-Datenreport 02/2023 (en). DOI: 10.5164/IAB.FDZD.2302.en.v1

Schmucker, Alexandra; vom Berge, Philipp (2025): Sample of Integrated Labour Market Biographies (SIAB) 1975-2023. FDZ-Datenreport 02/2025 (en). DOI: 10.5164/IAB.FDZD.2502.en.v1

Stüber, Heiko; Dauth, Wolfgang; Eppelsheimer, Johann (2023): A guide to preparing the sample of integrated labour market biographies (SIAB, version 7519 v1) for scientific analysis. Journal for Labour Market Research 57, 7. https://doi.org/10.1186/s12651-023-00335-w

vom Berge, Philipp; Schmucker, Alexandra (2021): Creating cross-sectional data and biographical variables with the Sample of Integrated Labour Market Biographies 1975-2019 - Programming examples for Stata. FDZ Methodenreport 05/2021 (en). DOI: 10.5164/IAB.FDZM.2105.en.v1

Appendix

Appendix 1: Download of the Stata do-files

https://doku.iab.de/fdz/reporte/2025/MR_01-25_EN_programs.zip

Imprint

FDZ-Methodenreport 01|2025 EN

Date of publication
31 March 2025

Publisher
Research Data Centre (FDZ)
of the Federal Employment Agency (BA)
in the Institute for Employment Research (IAB)
Regensburger Str. 104
D-90478 Nuremberg

Rights of use
This publication is published under the following Creative Commons licence:
Attribution – ShareAlike 4.0 International (CC BY-SA 4.0)
https://creativecommons.org/licenses/by-sa/4.0/deed.de

Download
https://doku.iab.de/fdz/reporte/2025/MR_01-25_EN.pdf

Documentation version
DOI: 10.5164/IAB.FDZM.2501.en.v1

All publications in the series "FDZ-Methodenreport" can be downloaded from
https://fdz.iab.de/en/research/publications/fdz-methodenreport-series/

Website
https://fdz.iab.de

Corresponding author
Alexandra Schmucker
Phone: +49 911 179-1752
Email alexandra.schmucker@iab.de