Journal for Labour Market Research

ORIGINAL ARTICLE

Open Access

A guide to preparing the sample of integrated labour market biographies (SIAB, version 7519 v1) for scientific analysis

Heiko Stüber[1,2,3*], Wolfgang Dauth[1,3,4] and Johann Eppelsheimer[5]

Abstract

The *Sample of Integrated Labour Market Biographies* (SIAB) is the most frequently requested data set provided by the Research Data Centre (FDZ) of the Federal Employment Agency (BA) at the Institute for Employment Research (IAB). However, preparing the SIAB for scientific analysis is a complicated and error-prone task. This article explains steps necessary to prepare the latest version of the SIAB (7519 v1) and gives examples of how the preparation can be done. Among other things, it shows how to impute right-censored wages, deal with parallel employment episodes, and clean up the data set. The supplementary material to this article contains extensively annotated Stata do-files to replicate our data preparation.

Keywords SIAB 7519, Data preparation, Stata, IAB, FDZ

JEL classification C55, C81, J65

1 Introduction

The *Sample of Integrated Labour Market Biographies* (SIAB) includes data on all registrations with the German social insurance system for a 2-percent sample of all persons who have ever been registered with the social insurance system since 1975. This enables day-by-day analyses of the (complete) labor market biographies of many individuals. At the same time, the sample size allows aggregated analyses at the level of regions, industries, or occupations. The SIAB is an extremely versatile data set and one of the most sought-after data products made available to the scientific community by the Research Data Centre (FDZ) of the Federal Employment Agency (BA) at the Institute for Employment Research (IAB).[1] Over 43% of the more than 650 projects currently managed by the FDZ use the SIAB.[2]

Individual-level data sets—such as the SIAB—usually differ from macro data in one respect: they have a complicated structure. The SIAB is no exception: In principle, each line of the SIAB originates either from an employer's notification to the social security system or from a process in the unemployment insurance system. Unfortunately, individual biographies often do not follow a linear path: people change jobs, have several jobs at the same time, become unemployed, participate in active labor market policy measures, etc. A data set covering all these different biographies is, as a result, more complex than a sample data set familiar from econometric textbooks. Therefore, researchers must first invest a lot of time and

*Correspondence:
Heiko Stüber
Heiko.Stueber@iab.de
[1] Institute for Employment Research (IAB), Regensburger Str. 100, 90478 Nuremberg, Germany
[2] Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Nuremberg, Germany
[3] IZA, Bonn, Germany
[4] Otto-Friedrich-University of Bamberg, Bamberg, Germany
[5] urban analytica, Nuremberg, Germany

[1] Further details on the structure and origin of the SIAB can be found in the SIAB 7519 data report by Frodermann et al. (2021). SIAB users are strongly encouraged to read the Data Quality and Problems section (Section 4) of this data report.

[2] As of August 18, 2022.

effort in preparing the data before they can start the actual empirical analysis.

Because of its versatility, scholars analyze the SIAB in numerous contexts. In the first half of 2022 alone, several publications using the SIAB were reported to the FDZ. Examples are: "Why do some occupations offer more part-time work than others? Reciprocal dynamics in occupational gender segregation and occupational part-time work in West Germany" by Bächmann et al. (2022), "Changing selection into full-time work and its effect on wage inequality in Germany" by Fitzenberger and de Lazzer (2022), "The role of labor demand in the labor market effects of a pension reform" by Geyer et al. (2022), "Reservation wages and labor Supply" by Kesternich et al. (2022), "Task specialization and the native-foreign wage gap: Evidence from worker-level data" by Storm (2022), and "Early retirement of employees in demanding jobs: Evidence from a German pension reform" by Zwick et al. (2022). As these examples show, the SIAB enables high-quality research in numerous areas.

Because researchers can use the SIAB in so many different contexts, data preparation methods naturally vary. Practices that have proven useful in one research project may not be applicable in another. Ultimately, it is not possible to develop a linear guide to SIAB preparation that addresses all requirements for all potential research questions. However, we recognize that the preparation of the SIAB is time-consuming and error-prone, especially for researchers who do not have experience working with larger administrative data sets. In this article we describe the best practices that we have found useful in our own research. We also provide extensively annotated Stata do-files to replicate our data preparation. The goal of this article and the supplemental collection of do-files is to provide researchers with step-by-step instructions on how to prepare the SIAB for individual-level analyses. We point out the purpose of each step and explain intuitively how to perform it. All technical details can be found in the do-files attached to this article.

Although this article is the first complete guide to preparing the SIAB 7519 for scientific analysis, there are several papers that provide guidance on specific parts of the data preparation process for German social security data.[3] This article and the collection of do-files are based on the weakly anonymized version of the SIAB 7519 v1 (DOI: https://doi.org/10.5164/IAB.FDZ.2101.en.v1). Access to this data set is provided to the scientific community by the FDZ. The usual way to access the SIAB is to use it on-site at the FDZ in Nuremberg or at one of the other locations in Germany, France, Luxembourg, Poland, and Spain.[4] This allows users to develop their programs while working interactively with the data. Afterwards, users can continue their projects via remote data access through the JoSuA web interface. The FDZ requires users working in the secure on-site environment and through JoSuA to follow certain conventions when writing their programs and managing their files. Our guidelines follow these conventions, and our do-files have been tested in the JoSuA environment using Stata version 17.

We urge all users of the Stata do-file collection to check our code for bugs and to adapt it to the requirements of the project at hand. We provide this as a service to the scientific community, but do not accept responsibility for any problems arising from the use of our code. We also strongly recommend reading the SIAB 7519 data report (Frodermann et al. 2021) to familiarize yourself with the SIAB, as well as data quality and issues.[5]

This is a substantial update of the original guide to preparing the SIAB by Dauth and Eppelsheimer (2020) published in this journal. We decided to make this update for two reasons: First, the SIAB has been extended and updated.[6] Second, we received a variety of comments and suggestions on the original version regarding further enhancements and usability improvements that we wanted to share with the users of this guide. The following points have been revised, updated, or added compared to the original version:

- We tested the do-files in the JoSuA environment of the FDZ on the latest weakly anonymized version of the SIAB (SIAB 7519 v1) using Stata 17.
- We introduced global macros in `00_master_SIAB.do` to enable/disable certain steps of data preparation. The global macro `inspect` allows to enable/disable certain outputs and the macro `logfile` allows to enable/disable the creation of log files.
- We changed the order of some of the do-files to reduce runtime and improve quality of the wage imputation of top-coded wages.

[3] An overview of some of these guides is provided in Dauth and Eppelsheimer (2020). They provide the very first complete guide to preparing the SIAB (version 7517) for scientific analysis.

[4] For more information on data access, visit https://fdz.iab.de/en/data-access.

[5] The introduction of the new occupation code KldB 2010 in 2011 led to several serious problems. Therefore, users of the SIAB are strongly encouraged to read the Data Quality and Problems section (Section 4) of Frodermann et al. (2021). The introduction of the KldB 2010 let to gaps in some variables. While the variable `teilzeit` (part time) has been updated using an imputation procedure of Ludsteck and Thomsen (2016), no imputation was performed regarding the gaps in the other variables `beruf`/`beruf2010_3` (occupation according to KldB 1988/2010), `erwstat` (occupational status), `ausbildung` (vocational training), and `schule` (school leaving qualification).

[6] For an overview of changes as compared to the SIAB 7517, see Frodermann et al. (2021).

- We added a do-file to deal with spells that report a one-off payment (grund = 154). It adjusts daily wages of employment spells of individuals within establishments, when the individuals received one-off payments from the same establishment (see 02_grund154.do).
- We adjusted the definition of the dummy indicating whether an individual works in West or East Germany (see 06_wages_assessment_ceiling.do). Berlin is now treated as West German until 1991 and East German from 1992 onward.[7]
- We added the contribution assessment limit of the statutory pension insurance, the marginal earnings thresholds for part-time employees (see 06_wages_assessment_ceiling.do), and the consumer price index (see 07_wages_marginal.do) for all years through 2019.
- We added do-files to merge the following data to the SIAB:
  - Further (non-)sensitive Establishment History Panel (BHP) variables (see 11_merge_BHP.do),
  - the "worker flows" and "entry and exits" BHP extension files (see 11_merge_BHP.do), and
  - person and establishment fixed effects as proposed by Abowd et al. (1999) (AKM effects, see 12_merge_AKM.do).
- We made minor revisions to all do-files: more efficient commands, expanded/revised comments, etc.

The SIAB 7521 is scheduled for release in 2023. With the release, we will provide an updated and revised Stata do-file collection. The availability of the new do-file collection will be announced here in the journal and in the FDZ newsletter.

Please note that our files have not been tested with the anonymized SIAB version available to users on-site in the US, Canada, and the UK. In this SIAB version, inter alia, the variables erwstat and grund are aggregated from 56 to 23 and 205 to 45 categories, respectively.[8] Since the additional anonymization is also project specific, users need to customize our do-file to their data set. Our files have also not been tested with the factually anonymous version of the SIAB (SIAB-R 7519). In the SIAB-R 7519, many of the variables from the weakly anonymized version of the SIAB 7519 have been grouped (see Frodermann et al. 2021). Therefore, users need to customize our do-file to this data set.

[7] Up to and including 1991, the BeH only contains employment information from West Germany—thus only West Berlin is coded as Berlin. From 1992 onward, the BeH also contains employment information from East Germany—thus West and East Berlin are coded as Berlin.

[8] New variable names are erwstat_gr and grund_gr, respectively.

2 Data preparation procedure

In the following, we present guidance for preparing the weakly anonymized version of the SIAB 7519. Starting from the original SIAB provided by the FDZ, we deal with spells that contain one-off payments, generate and merge additional variables, impute right-censored wages, consider parallel episodes, and clean the data set. We also show how to convert the spell data format to an annual panel.

Our preparation of the SIAB follows a modular organization in which each step has its own do-file. All Stata do-files for this exercise are included in the online supplement to this article. The full set of do-files is run from 00_master_SIAB.do. Using the default settings, the runtime is about 2 1/4 h.[9] Table 1 in the Appendix gives an overview of all variables we generate and modify in this guide.

When working on-site at the FDZ or via remote data access, the working directory and the subfolders for do-files (**prog**), the original data (**orig**), prepared data (**data**), and eventual outputs (**log**) are preset. Global macros to address these folders are automatically set. All do-files must be uploaded via JoSuA. They are then accessible under the path $prog.

Please download the FDZ Template-Do-Files and use the template of master.do to create your own master.do (https://doku.iab.de/fdz/access/FDZtemplate.zip). Your master.do should be uploaded to JoSuA as well and it should be used to execute 00_master_SIAB.do.[10]

2.1 do-file: 00_master_SIAB.do

The main purpose of this do-file is to start all parts of the data preparation in the right order. Before executing the individual do-files to prepare the SIAB, it creates the variables jahr (*year*) and age.[11] In between, after running 03_SIAB_bio.do, it restricts the SIAB to selected years. After data preparation, 00_master_SIAB.do sets the label language (German or English) and saves the prepared data set as siab_clean.dta in the data folder.[12]

[9] If the data set is not restricted in 12_restrictions.do and all available data is merged with the SIAB, etc., the runtime increases to about 7 h.

[10] For users of the SIAB test data, we provide a commented out section of code in 00_master_SIAB.do that creates the necessary subfolders and sets the global macros. When working on-site at the FDZ or via remote data access, this section must remain commented out.

[11] Age is calculated using the year of birth, since the month of birth is considered a sensitive variable. This must be requested separately, with a justification that the month of birth is essential for the research project.

[12] With the default settings, the size of siab_clean.dta is about 2.9 GB. Without any restrictions and by merging all variables, it is about 14.8 GB.

Various global macros can be set in `00_master_SIAB.do`. These set the observation period for which the SIAB will be prepared (default setting: 1975–2019) and they determine whether certain preparation steps are carried out or not. If certain preparation steps are not necessary for the creation of the desired data set and these cannot be deactivated by a global macro, one can simply comment them out. In the do-file we give hints which do-files can be excluded for sake of parsimony and which are mandatory.

Also, the global macro `inspect` can be set (disabled by default). Setting this global macro to 1 will generate certain outputs (tables and figures) to inspect the data. In JoSuA presentation/publication (PP) mode, please leave the default setting to reduce the outputs that need to be reviewed by the FDZ as part of the data protection review.

In principle, FDZ users should only log outputs in JoSuA PP mode that they need for the publication or a presentation. However, it often happens that users log the complete data preparation. To allow users to conveniently decide whether they want to log the output of a particular do-file or not, we introduced the global macro `logfile`. It is always set before the execution of a do-file, the corresponding log-file will be opened only if the global macro is set to 1.

2.2 do-file: 01_split_episodes.do

In the SIAB, the spells are already split into episodes. This means that overlapping spells (e.g., multiple employment, job search during employment, etc.) are split so that parallel spells always have the same start (`begepi`) and end (`endepi`) dates.[13] The start and end dates of the original (non-split) spells are stored in `begorig` and `endorig`. By definition, employment episodes from the Employment History (Beschäftigten-Historik, BeH; `quelle`=1) break at the turn of the year. However, episodes from the Beneficiary History (Leistungsempfänger-Historik, LeH; `quelle`=2) and the Unemployment Benefit II Recipient History (Leistungshistorik Grundsicherung, LHG; `quelle`=3) can span multiple calendar years. For many applications, such as the creation of an annual panel data set, it is useful to have at least one observation per calendar year. Therefore, `01_split_episodes.do` divides periods that span more than one calendar year into multiple episodes and changes the start and end dates stored in the variables `begepi` and `endepi` accordingly. The original values are stored in the variables `begepi_orig` and `endepi_orig`. In addition, the do-file updates the variables `jahr` and `age`.

2.3 do-file: 02_grund154.do

Since 2013, the number of spells reporting one-off payments (deregistration reason 54; coded as `grund`=154) has increased sharply (cf. Frodermann et al. 2021, Sections 5.5.1 and 5.5.12). These spells always cover exactly an entire month. It is likely that one-off payments that were reported with annual spells before 2013 are now reported separately. It is therefore advisable, when analyzing wages over time, to add these one-time payments to wages for employment episodes within the same establishment in the corresponding year.

Therefore, `02_grund154.do` adjusts the daily wage (`tentgelt`). Within years, one-time payments of notifications with deregistration reason 154 are divided proportionally among all other periods of employment of a person in the same establishment. Subsequently, all periods with deregistration reason 154 are deleted.

2.4 do-file: 03_SIAB_bio.do

The SIAB covers the entire employment history of the individuals in the sample from 1975 onward. This information can be used to generate variables that summarize (previous) careers, such as length of employment or labor market experience. Vom Berge and Schmucker (2021) provide a do-file that generates several biographical variables, in particular `tage_erw` (days in employment) and `tage_bet` (length of service, resumption of count after interruptions).

To make the do-file by vom Berge and Schmucker (2021) compatible with `00_master_SIAB.do`, we have to comment out some lines. However, the code generating the biographical variables itself remains unchanged. We provide the modified do-file as `03_SIAB_bio.do`. An overview of all generated variables can be found in Table 1 in the Appendix, for the exact definitions of the generated variables please refer to the do-file itself.

After running `03_SIAB_bio.do`, `00_master_SIAB.do` restricts the SIAB to selected years.

2.5 do-file: 04_merge_basic_BHP.do

For many research projects, it is useful to have data on an individual's establishment, e.g., location and industry codes, number of employees, average wages, etc. The Establishment History Panel (Betriebs-Historik-Panel, BHP) consists of administrative data on the entire population of all employees covered by the German Social Security on June 30 of a given year, aggregated at the establishment level. The FDZ provides a subsample of the BHP, the Basic Establishment File, that includes only establishment-year combinations that appear in the SIAB.[14] Establishments from eastern Germany are

[13] For details on the episode split in the original SIAB, see Frodermann et al. (2021).

[14] For details on the BHP, see Ganzer et al. (2022).

included from 1992 onward. The Basic Establishment File is provided with the SIAB by default and merged to the SIAB in `04_merge_basic_BHP.do`. By default, only the variables `ao_bula` and `w93_3_gen` are merged to the SIAB. In the do-file you find instructions on how to merge further variables of the Basic Establishment File to the SIAB.

Please note that in `11_merge_BHP.do` we provide the possibility to merge additional variable blocks, sensitive BHP variables, and extension files to the SIAB. However, this data is not automatically provided with the SIAB, but must be requested when applying for the SIAB.

For an overview of the BHP variables contained in the Basic Establishment File, as well as an overview of the available additional variable blocks, sensitive variables and extension files, see https://doku.iab.de/fdz/access/BHP_Variablen_EN.pdf.

2.6 do-file: 05_educ_broad.do

The information on a person's highest educational attainment is often inconsistent in German administrative data. For example, in some periods individuals are registered with a university degree, while in subsequent periods their highest educational attainment is an apprenticeship. To correct for such implausible trends in educational attainment, the FDZ provides an imputed version of the variable `ausbildung`, which is stored as `ausbildung_imp`. The imputation procedure builds on Fitzenberger et al. (2006).

The variable `ausbildung_imp` specifies six levels of education. Since many researchers prefer broader education groups, we provide code in `05_educ_broad.do` to group the six education categories into three education categories. We distinguish between spells without vocational training (1), completed vocational training (2), and degrees from a university or university of applied science (3).

2.7 do-file: 06_wages_assessment_ceiling.do

Since the data originate from mandatory reports to the social security authorities, the wage data in the SIAB are generally very reliable. However, due to this administrative origin, wages are only reported until they reach the contribution assessment limit in the statutory pension insurance. If they exceed this limit, the wages are coded with this value. This contribution assessment limit varies by year and between East and West Germany. We collect the corresponding assessment limits in `06_wages_assessment_ceiling.do` and store them in the variable `limit_assess`.[15] In addition, the variable `east` is created in the do-file, which takes the value one if a person's workplace is located in the eastern part of Germany and the value zero if it is located in the western part of the country. Data from East German employers is included starting in 1992. Since the region code for Berlin does not distinguish between East and West Berlin, all of Berlin is assigned to West Germany until 1991 and to East Germany from 1992 onward.

[15] The IAB Research Data Center provides a detailed list of earnings limits: https://doku.iab.de/fdz/Bemessungsgrenzen_de_en.xls.

2.8 do-file: 07_wages_marginal.do

Another important threshold is the marginal earnings threshold for part-time employees. Jobs with wages below this threshold are either exempt from social security contributions (before 1999) or subject to a lump-sum contribution payable by the employer (1999 or later). Therefore, these jobs are only included in the SIAB from 1999 onward (cf. Frodermann et al. 2021). In `07_wages_marginal.do`, we create the variable `limit_marginal`, which stores the marginal part-time earnings threshold. In addition, the do-file marks observations with wages below the part-time income threshold using the dummy variable `marginal`.

2.9 do-file: 08_wages_deflation.do

To make wages comparable across years, we calculate real wages (`wage_defl`). For deflation, we divide nominal wages by the Federal Statistical Office's consumer price index (Statistisches Bundesamt 2022). In addition, we also deflate the income thresholds for contributions (`limit_assess_defl`) and the marginal income thresholds (`limit_marginal_defl`). We use a specific consumer price index for West Germany for the years 1975–1991 and a general consumer price index for the whole country for the years 1992–2019. The base year is 2015.

2.10 do-file: 09_restrictions.do

An important part of preparing the SIAB for scientific analysis is to limit the data to a meaningful subsample. For example, most projects that use daily wages limit the subsample to full-time workers. In addition, many projects are only interested in specific time periods, regions, employment status, or sociodemographic groups. To simplify data restriction, we collect typical sample restrictions in `09_restrictions.do`

Restricting the data to a meaningful subsample before imputing top-coded wages improves the quality of the imputation and reduces its runtime. However, restricting the data may change the "definition" of the aggregate variables that can be generated in `15_parallel_episodes.do` and `16_yearly_panel.do`. In `15_parallel_episodes.do`, we offer the option to generate potentially relevant information from parallel spells, e.g., the sum of (imputed) wages from all

parallel spells (parallel_wage, parallel_wage_imp). In 16_yearly_panel.do, we offer the option to aggregated (labor) outcomes for each person, e.g., the total number of employment days per calendar year (yearly_days_emp). These variables are calculated accurately only if only basic restrictions (e.g., on age or sex) are applied when forming the subsample. Constraints, e.g., on occupational status, occupations, full-time/part-time employment, or jobs, already cause these variables to be generated as a function of the constraint criteria. For example, if only employment in West Germany is considered, the total labor earnings per calendar year changes to the total labor earnings per calendar year from employment in West Germany. Therefore, we decided to include the global macros parallel_vars and yearly_vars in 00_master_SIAB.do, which enables/disables the generation of these aggregated variables (disabled by default).

If users are **interested** in these aggregated variables, we suggest the following: Implement only basic restrictions in 09_restrictions.do. All other restrictions should be implemented in the designated area in 15_parallel_episodes.do or 16_yearly_panel.do. This ensures that the aggregated variables are calculated accurately.

If users are **not interested** in these aggregated variables, all restrictions should be implemented in 09_restrictions.do. If they want to create an annual panel at the end, they should make sure that they only keep spells that exist on the cutoff date for creating the annual panel. This will significantly shorten the runtime of the calculation of right-censored wages. By default, 16_yearly_panel.do uses June 30 as the cutoff date because the BHP data is measured on the last day of June (cf. Ganzer et al. 2022).

Please note that our code contains the following restrictions in the default settings:

- Keep only spells that exist on the annual panel cutoff date (default setting: June 30).
- Only keep spells from people who are 18 to 65 years old.
- Delete all employment (BeH) spells with a...
  - missing establishment ID (betnr),
  - missing in the dummy variable for eastern Germany (east), or
  - daily wage (tentgelt) of zero Euros.

2.11 do-file: 10_wages_imputation.do

Since the SIAB is based on process data used to calculate retirement pensions and unemployment insurance benefits, the wage information is highly reliable in general. However, for these administrative purposes, the wage information is only relevant up to the social security contribution ceiling. Unfortunately, this means that the wage information in the process data is top-coded, and hence we only observe wages up to the social security contribution ceiling. While this feature only affects approximately 5.2% of all employment spells for workers between 1975 and 2019, the proportion of censored observations within certain subgroups is substantial. For instance, nearly 44% of the spells of regularly employed male workers with a degree from a university or university of applied science are affected by top-coding. The share of top-coded wages also increases over time. To prevent biased estimates in later empirical analysis, we impute top-coded wages. If censoring is moderate, imputed wages allow valid inference on the parameters generated by uncensored data. However, researchers should also be aware that imputed wages cannot compensate for the loss of information in subgroups with large shares of censored wages, such as high-skilled workers. Hence, analyses focusing on such subgroups should be carried out with great care.

Ahead of the actual imputation procedure, 10_wages_imputation.do creates the indicator variable cens that flags censored wages. To ensure that all censored wages are covered in the imputation procedure, we mark all observations with wages four Euros below the assessment ceiling. Furthermore, the do-file generates a new wage variable wage that is top-coded at four Euros below the real assessment ceiling (in 2015 Euros) for all observations.[16] We also log-transform wages.

To impute top coded wages, we use a two-step procedure similar to that in Dustmann et al. (2009) and Card et al. (2013).[17] By default, 10_wages_imputation.do first clusters observations by year, East and West Germany, and three education groups. One may also consider to further distinguish between women and men when the aim of the analysis is to study gender differences. However, one would have to weight this increase in validity with the drastic reduction of cluster sizes, when the data is split by a further dimension. For each of these clusters, we fit Tobit wage equations, controlling for worker characteristics X.[18]

[16] In some cases, reported wages in the original wage variable are above the assessment limit.

[17] In contrast to multiple imputation procedures, this two-step imputation procedure generates slightly biased standard errors but is easier to handle and more efficient in terms of computation times (cf. Gartner 2005).

[18] This is done by saving the data set in a temporary file and then running the Tobit regressions on smaller subsets of the data set, each containing only one cluster. After imputation, these subsets are appended to obtain the full data set with imputed wages. An alternative would be to keep the data set in the memory and use the if-condition within the 'intreg' command. This alternative consumes vastly more memory and computation time, as pointed out in https://twitter.com/marxmatt/status/1104570847648456704?s=20. In the present case, using the if-condition takes approximately 32 times longer compared to splitting the data into smaller subsets.

A naive estimator for the censored wages would be the simple expected value of the log wage, conditional on the observable characteristics $E[lnw|X] = X\beta$, where β are the regression coefficients. However, since this is a function only of X, it is more strongly correlated with the covariates than the true unobserved log wages. A way to mitigate this problem is to assume that wages are log-normally distributed and add a normally distributed random term to the fitted values. We hence overwrite censored log wages with $X\beta + \sigma\Phi^{-1}[k + u(1-k)]$, where σ is the standard deviation of the residual, Φ is the standard normal density function, u is a random draw from a uniform distribution ranging between zero and one, $k = \Phi[(c - X\beta)/\sigma]$ and c is the censoring point. For a detailed description of the underlying rationale, refer to Gartner (2005).

In an intermediate step, the do-file calculates the average log wage of each worker over time and of all workers within each plant in one year. Those averages are "leave-one-out means", which means that the averages are computed while not considering the respective observation. If there is only one worker or plant observation, we instead use the sample mean.

Next, we repeat the Tobit wage regressions from step one, including the computed average log wages as well as a variable that indicates whether the sample mean was used. Including those averages comes close to controlling for worker and plant fixed effects. Having predicted censored log wages, we transform the log wages into Euros (in 2015 Euros) and store imputed wages in the variable `wage_imp`. Although it is extremely unlikely, by chance, imputed wages could be exceedingly high in some cases. As a minor adjustment, we therefore limit imputed wages to ten times the 99th percentile of the wage distribution. Another very rare exception is that extraordinarily low plant leave-one-out means cause a numeric overflow when inverting the normal distribution in the second step of the imputation procedure. Consequently, the log wages of the affected observations cannot be predicted. In such uncommon cases, we use imputed wages from the first step.

Figure 1 shows the distribution of the censored daily wage (in 2015 Euros) and its imputed equivalent after the first and second steps. The raw wage has two distinct spikes at the social security ceilings in West and East Germany. After imputation, these spikes disappear, and more mass can be found in the right tail of the distribution. The difference between the first and second steps of imputation is very subtle.

Although `10_wages_imputation.do` could serve as a blueprint for wage imputations in various research projects, it is highly recommended to customize the program first. Most importantly, researchers should adjust the set of control variables. In most cases, the wage serves either as the outcome or as one of the explanatory variables in the econometric model of the main analysis. In principle, all variables from this model should also be included in the set of control variables of the imputation procedure. The reason is that omitting variables could lead to biased estimates.[19] The potential bias depends on the correlation between wage and the omitted covariates.[20] Additionally, it might also be reasonable to choose different subgroups from the ones we suggest in the do-file.

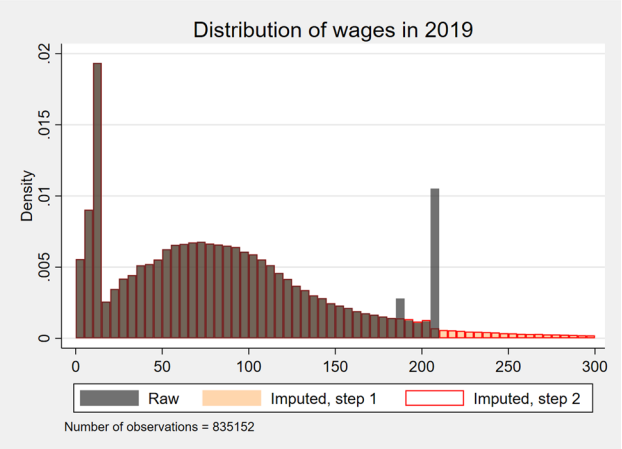

Fig. 1 Distribution of daily wages in 2019. Note: The figure reports the distribution of the censored daily wage (in 2015 Euros) and its imputed equivalent after the first and second imputation steps. Generate using the default settings of the provided do-files

In cases where control variables vary on an aggregate level (e.g., regional variables), it could be that most of their variation is already captured by the leave-one-out means. Such a multicollinearity issue could lead to unstable predictions of wages. Thus, when controlling for macro variables, it is particularly advisable to carefully inspect the Tobit estimates. Furthermore, under the suspicion of unstable estimates, it might be reasonable to omit problematic macro variables from the imputation.[21]

2.12 do-file: 11_merge_BHP.do

In this do-file, we provide the option to merge additional variable blocks, sensitive BHP variables, and extension files to the SIAB.[22] Please note that this data

[19] Compared to Dauth and Eppelsheimer (2020), we included the dummy variable part-time as a control variable in the default setting. Alternatively, users could add an additional loop and perform imputation separately for full-time and part-time employees. As in Dauth and Eppelsheimer (2020), marginally employed workers are not included in the imputation.

[20] If wage is the dependent variable, the coefficients of the variables omitted in the imputation are biased toward zero.

[21] One sign of unstable predictions could be implausibly large coefficients on macro variables.

[22] For details on the BHP, see Ganzer et al. (2022). For an overview of the available additional variable blocks, sensitive variables and extension files, see https://doku.iab.de/fdz/access/BHP_Variablen_EN.pdf.

is not automatically provided with the SIAB, it must be requested when applying for the SIAB.

To merge additional variable blocks, sensitive BHP variables, or the extension files "Worker flows" or "Entry and exit" with the SIAB, the corresponding global macros in `00_master_SIAB.do` must be set to 1.

2.13 do-file: 12_merge_AKM.do

The FDZ provides person and establishment fixed effects (AKM effects; cf. Abowd et al. 1999) for the following time-windows: 1985–1992, 1993–1999, 1998–2004, 2003–2010, and 2010–2017 (cf. Bellmann et al. 2020). The AKM effects can be requested when applying for the SIAB.[23]

In `12_merge_AKM.do`, we provide the option to merge the person and/or establishment fixed effects with the SIAB. To merge them with the SIAB, the corresponding global macros in `00_master_SIAB.do` must be set to 1.

2.14 do-file: 13_industries_1digit.do

The BHP contains several detailed industry classifications. However, studies are often interested in broader classifications. Therefore, we provide the option to translate the three-digit industry codes into two alternative one-digit aggregates.

For the assignment, we use the variable `w93_3_gen`, which contains a time-consistent version of the three-digit German equivalent of NACE Rev. 1 (Eberle et al. 2014). We map these industries to the classifications of the Federal Statistical Office (Statistisches Bundesamt 2002) and the classification scheme of the IAB Establishment Panel (see, e.g., Ellguth et al. 2014).

`13_industries_1digit.do` aggregates `w93_3_gen` to 15 industries stored in the variable `industry1_destatis` and nine 1-digit industries stored in the variable `industry1_estpanel`. The main difference between the two classifications is the level of detail within the primary sector and public services. To create one or both of the mappings, set the corresponding global macros in `00_master_SIAB.do` to 1.

2.15 do-file: 14_occ_blossfeld.do

The weakly anonymous version of the SIAB also contains very detailed occupational information.[24] For some applications this is too detailed, and information on broad categories would suffice. We therefore add the widely used occupational classification of Blossfeld (1987) to the data set (see also Schimpl-Neimanns 2003). The do-file creates the variable `occ_blo`, which contains 13 occupational groups. These 13 groups are created by recoding the 1988 three-digit occupation codes (*Klassifizierung der Berufe 1988, KldB 88*) in the variable `occupation`.[25] To generate the Blossfeld (1987) occupational classification, set the corresponding global macro in `00_master_SIAB.do` to 1.

2.16 do-file: 15_parallel_episodes.do

Many analyses require that each individual is observed only once at each point in time. In reality, however, biographies are nonlinear, and parallel spells are very common. Therefore, `15_parallel_episodes.do` restricts the data set to the main spell and provides the option to retain potentially relevant information from parallel episodes.

By default, `15_parallel_episodes.do` defines the parallel spell with the longest tenure as the main episode. However, one could also treat the spell with the highest wage as the main episode. Depending on the research question, alternative approaches might be reasonable. Only the main spells will be kept at the end of the do-file. Be aware of the fact that the data needs to be sorted unambiguously before dropping the other spells. Otherwise, the sample selection might differ each time the code is executed.[26]

The following potentially relevant information from parallel episodes can be generated: the number of non-parallel spells (`nspell`) and parallel jobs (`parallel_jobs`), sum up total (imputed) wages from all parallel spells (`parallel_wage`, `parallel_wage_imp`), and create an indicator variable for the receipt of unemployment benefits in parallel spells (`parallel_benefits`).[27] However, these aggregated variables might be not precisely calculated depending on the restrictions applied to the data in `09_restrictions.do`. We therefore add the global macro `parallel_vars` to

[23] To obtain the AKM effects for an existing project, please write an email to iab.fdz@iab.de.

[24] Please note that due to the introduction of the new occupational code KldB 2010 in 2011, the SIAB 7519 exhibits a significant structural break in the occupational information. Users are strongly advised to read the "Data quality and issues" section (Section 4) of Frodermann et al. (2021).

[25] It should be noted that we use the 1988 classification of occupations because it is easily transferable to Blossfeld's scheme. Alternatively, researchers could create broader occupational classes based on the variable `occupation2010_3`, which contains more recent occupational codes from 2010.

[26] For technical details, please refer to the note about the requirement for unique sorting in `15_parallel_episodes.do`.

[27] In rare cases, there are parallel spells of the same individual at the same employer. Often, the variable `grund` (reason of notification) indicates a one-off payment. With these spells we deal in `02_grund154.do` and delete them afterwards. The reason for the remaining cases (around 0.5% of all BeH-spells) is unclear. However, obvious cases of erroneously doubled spells (with identical values of the variables `persnr`, `betnr`, `begepi`, `endepi`, `tentgelt`, and `grund`) have already been eliminated during the construction of the underlying raw data. We leave it to the researcher to decide whether to keep only one observation or use our procedure to add up the wages in such cases.

00_master_SIAB.do, which enables/disables the generation of these variables (disabled by default).[28]

2.17 do-file: 16_yearly_panel.do

Since spell data is complicated to handle and many research projects do not depend on the exact duration of spells or the appearance of multiple spells per year, researchers often prefer to limit the data to one observation per year per individual. Most likely, the simplest procedure to do so is to restrict the data to observations that cover a predefined cutoff date.

Before we simplify the data set, we offer the option to retain some of the information of all spells from the same individual/year combination. Specifically, we compute the total number of days an individual was employed or received benefits from unemployment insurance during a calendar year (year_days_emp and year_days_benefits). Similarly, we also compute the total labor earnings of an individual during a calendar year (year_labor_earn). However, these variables might be not precisely calculated depending on the restrictions applied to the data in 09_restrictions.do. We therefore add the global macro yearly_vars to 00_master_SIAB.do, which enables/disables the generation of these variables (disabled by default).[29]

By default, 16_yearly_panel.do uses June 30 as the cutoff date because data from the BHP is measured on the last day of June (Schmucker et al. 2016). This cutoff date is set in 00_master_SIAB.do. Of course, there might be situations where a different cutoff date is more appropriate. Furthermore, researchers might be interested in panel data with a higher frequency, such as monthly or quarterly data. For instance, to create a monthly panel, users first have to count the months each spell contains and then *n-plicate* the spells accordingly using Stata's *expand* function. Next, users should adjust the start and end dates of spells to cover only one month. Finally, users could restrict the sample to one spell per month, e.g., by using multiple cutoff dates or by only keeping the main spell in each month. Compared to a yearly panel, a monthly panel allows us to more precisely model the timing of economic effects. However, studies should be aware that in Germany, employers are not obligated to register all status changes of their employees within the year. Therefore, the majority of employment spells come from compulsory notifications from the end of the year. Artificially expanding static data might result in overconfident estimates. Thus, researchers should study their data carefully before expanding it and judge whether such an operation is legitimate. Another alternative to a yearly panel is a data set that counts the days until/since a specific treatment date. For spells that do not cover the treatment date, users can simply count the days until/since the treatment. For episodes that include the treatment date, researchers would have to design a more elaborate procedure that is in line with their research questions. Generally, it is advisable to align the timing of the data set with the timing of the research question. For instance, when we are interested in the effect of work experience on earnings and are working with a yearly panel with June 30 as the reference date, it makes sense to also measure the accumulated work experience on June 30 every year.

2.18 do-file: 17_clean_up.do

Finally, 17_clean_up.do sorts the data, declares the panel structure (default setting: yearly panel), and compresses variables.[30] If variables are irrelevant for the data analysis, it might also be advisable to drop them using 17_clean_up.do. After running it, 00_master_SIAB.do sets the label language (default setting: English) and saves the final data set as siab_clean.dta in the data folder.

3 Concluding remarks

Since the first version of the SIAB was made available to the scientific community, researchers have asked the IAB to provide either a fully prepared version of the the SIAB or a field manual that explicitly describes the entire process of data preparation. While such a manual would certainly be very practical, it also presents a number of problems. It is extremely difficult to accommodate the idiosyncratic needs of any research project. There are several ways to deal with problems in the data, such as parallel spells or censored wages, and often there is no consensus on which method is the right one. Providing a manual for data preparation would risk consolidating a status quo where decisions should really be made by the researcher.

We are aware of these problems, but we also know that preparing the SIAB is a complicated task, especially for researchers who have little experience preparing administrative data. Even experienced researchers may not be aware of some of the problems in preparing the SIAB. Our goal is not to provide a comprehensive manual to preparing the SIAB. Rather, we want to highlight the steps necessary to prepare this data set and provide some examples of possible approaches to preparation.

[28] If users wish to generate these variables, please ensure that only basic restrictions are made in 09_restrictions.do. Additional restrictions on the data should be made in the designated area in 16_yearly_panel.do. Please refer to our explanations and suggestions in Section 2.10. (do-file: 09_restrictions.do).

[29] If users wish to generate these variables, please refer to footnote 28.

[30] When compressing the data, we exclude identifier variables. The reason is that in some rare cases, the compression of identifier variables could lead to a loss of information. For instance, this loss of information is problematic when merging other data products with the SIAB.

This collection of best practices is a combination of codes used by the authors and some of their colleagues. Running our do-files on the SIAB 7519 yields a processed version of this data set that can serve as a starting point for a variety of different research projects in applied micro labor economics or sociology. We have made a sincere effort to minimize the number of errors in our code. However, we cannot guarantee that there are no more bugs, nor that this collection is complete. We strongly recommend all users of this collection check our code for bugs and adapt it to the needs of their project. We are not responsible for any problems that may arise from using our code.

Appendix

List of generated or modified variables

Table 1 Generated and modified variables

Variable name	Short description	Do-file
age	Age (in years)	00_master_SIAB.do, 01_split_episodes.do
anz_lst	Number of benefit receipts; for exact definition, see do-file	03_SIAB_bio.do
azubi	Apprentice dummy; for exact definition, see do-file	03_SIAB_bio.do
begepi	Split version of begepi	01_split_episodes.do
begepi_orig	Original version of begepi	01_split_episodes.do
cens	1 if right-censored/imputed wage, 0 otherwise; (4 EUR below assessment ceiling)	10_wages_imputation.do
east	1 if workplace in East Germany (incl. Berlin from 1992 onward); 0 if West (incl. Berlin from until 1991)	06_wages_assessment_ceiling.do
educ	Education (university and uni. of applied science combined), imputed based on Fitzenberger et al. (2006)	05_educ_broad.do
ein_bet	First day in establishment; for exact definition, see do-file	03_SIAB_bio.do
ein_erw	Date of entry into first employment; for exact definition, see do-file	03_SIAB_bio.do
ein_job	First day in job; for exact definition, see do-file	03_SIAB_bio.do
endepi	Split version of endepi	01_split_episodes.do
endepi_orig	Original version of endepi	01_split_episodes.do
industry1_destatis	Industry; 1-digit; Statistisches Bundesamt; based on w93_3_gen	13_industries_1digit.do
industry1_estpanel	Industry; 1-digit; IAB establishment panel; based on w93_3_gen	13_industries_1digit.do
jahr	Year	00_master_SIAB.do, 01_split_episodes.do
limit_assess	Contribution assessment ceiling	06_wages_assessment_ceiling.do
limit_assess_defl	Contribution assessment ceiling, deflated (2015)	08_wages_deflation.do
limit_marginal	Marginal part-time income threshold	07_wages_marginal.do
limit_marginal_defl	Marginal part-time income threshold, deflated (2015)	08_wages_deflation.do
marginal	1 if marginal wage, 0 otherwise	07_wages_marginal.do
nspell	Nonparallel spell counter	15_parallel_episodes.do
occ_blo	Blossfeld occupations	14_occ_blossfeld.do
parallel_benefits	Indicator for receipt of UI benefits	15_parallel_episodes.do
parallel_jobs	Number of parallel jobs	15_parallel_episodes.do
parallel_wage	Total wage of all parallel employment spells	15_parallel_episodes.do
parallel_wage_imp	Total imputed wage of all parallel employment spells	15_parallel_episodes.do
tage_bet	Number of days in establishment; for exact definition, see do-file	03_SIAB_bio.do
tage_erw	Number of days in employment; for exact definition, see do-file	03_SIAB_bio.do, 16_yearly_panel.do
tage_job	Number of days in job; for exact definition, see do-file	03_SIAB_bio.do
tage_lst	Number of days with benefit receipt; for exact definition, see do-file	03_SIAB_bio.do, 16_yearly_panel.do
tentgelt	Daily wage, not imputed; corrected for one-off payments	02_grund154.do
wage	Daily wage, deflated (2015), not imputed, top-coded wages replaced by assessment ceiling (-4 EUR), deflated (2015)	10_wages_imputation.do
wage_defl	Daily wage, deflated (2015), not imputed	08_wages_deflation.do
wage_imp	Imputed daily wage, deflated (2015)	08_wages_imputation.do
year_days_benefits	Total days benefit receipt per calendar year	16_yearly_panel.do
year_labor_earn	Total labor earnings per calendar year	16_yearly_panel.do
year_days_emp	Total days employed per calendar year	16_yearly_panel.do

Acknowledgements
We thank Manfred Antoni, Melanie Arntz, Ann-Christin Bächmann, Johanna Eberle, Andreas Ganzer, Nina Gläser, Peter Haller, Michelle Hansch, Markus Janser, Oskar Jost, Markus Köhler, Max Kunaschk, Florian Lehmer, Johannes Ludsteck, Joachim Möller, Christoph Müller, Aderonke Osikominu, Alexander Patzina, Martin Popp, Alexandra Schmucker, Claus Schnabel, Alexandra Spitz-Oener, Philipp vom Berge, Florian Zimmermann, and anonymous referees for useful comments and suggestions and for sharing their code. All errors remain our own.
This is a revised, updated, and expanded version of (Dauth and Eppelsheimer 2020), which provides a guide for preparing SIAB 7517 along with an exemplary case study, which we have omitted in this version.

Author contributions
All authors read and approved the final manuscript.

Funding
This article was written when Heiko Stüber and Wolfgang Dauth were employees of the IAB and Johann Eppelsheimer was an employee of immowelt GmbH and co-founder of the startup "urban analytica" (https://urbananalytica.de/). The authors did not receive any special funding for this publication.

Availability of data and materials
The SIAB 7519 is not publicly accessible for data protection reasons (social security data). However, the Research Data Center (FDZ) of the Federal Employment Agency (BA) at the Institute for Employment Research (IAB) makes this data set available to the scientific community. All program codes can be found in Appendix.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 12 September 2022 Accepted: 12 January 2023
Published online: 13 February 2023

References

Abowd, J., Kramarz, F., Margolis, D.: High wage workers and high wage firms. Econometrica **76**(2), 251–333 (1999)
Bächmann, A.-C., Gatermann, D., Kleinert, C., Leuze, K.: Why do some occupations offer more part-time work than others? Reciprocal dynamics in occupational gender segregation and occupational part-time work in West Germany, 1976–2010. Soc Sci Res **104**, 102685 (2022)
Bellmann, L., Lochner, B., Seth, S., Wolter, S.: AKM effects for German labour market data. FDZ-Methodenreport 01/2020. (2020)
Blossfeld, H.-P.: Labor-market entry and the sexual segregation of careers in the Federal Republic of Germany. Am. J. Sociol. **93**(1), 89–118 (1987)
Card, D., Heining, J., Kline, P.: Workplace heterogeneity and the rise of West German wage inequality. Quart. J. Econ. **128**(3), 967–1015 (2013)
Dauth, W., Eppelsheimer, J.: Preparing the sample of integrated labour market biographies (SIAB) for scientific analysis: a guide. J. Labour Mark. Res. **54**, 10 (2020)
Dustmann, C., Ludsteck, J., Schönberg, U.: Revisiting the German wage structure. Quart. J. Econ. **124**(2), 843–881 (2009)
Eberle, J., Jacobebbinghaus, P., Ludsteck, J., Witter, J.: Generation of time-consistent industry codes in the face of classification changes. FDZ-Methodenreport **05**, 2011 (2014)
Ellguth, P., Kohaut, S., Möller, I.: The IAB Establishment Panel—methodological essentials and data quality. J Labour Mark Res **47**(1–2), 27–41 (2014)
Fitzenberger, B., de Lazzer, J.: Changing selection into full-time work and its effect on wage inequality in Germany. Empir. Econ. **62**, 247–277 (2022)
Fitzenberger, B., Osikominu, A., Völter, R.: Imputation rules to improve the education variable in the IAB Employment Subsample. Schmollers Jahrbuch. Z Wirt Soz. **126**(3), 405–436 (2006)
Frodermann, C., Ganzer, A., Schmucker, A., vom Berge, P.: Sample of Integrated Labour Market Biographies Regional File (SIAB-R) 1975–2019 (2021). FDZ-Datenreport 05/2021
Frodermann, C., Schmucker, A., Seth, S., vom Berge, P.: Sample of integrated labour market biographies (SIAB) 1975–2019 (2021). FDZ-Datenreport 01/2021
Ganzer, A., Schmucker, A., Stegmaier, J., Stüber, H.: Establishment history panel, 1975–2020 (2022). FDZ-Datenreport 03/2022
Gartner, H.: The imputation of wages above the contribution limit with the German IAB Employment Sample (2005). FDZ-Methodenreport 02/2005
Geyer, J., Haan, P., Lorenz, S., Zwick, T., Bruns, M.: The role of labor demand in the labor market effects of a pension reform. Ind. Relat. **61**(2), 152–192 (2022)
Kesternich, I., Schumacher, H., Siflinger, B., Valder, F.: Reservation wages and labor supply. J. Econ. Behav. Organ. **194**, 583–607 (2022)
Ludsteck, J., Thomsen, U.: Imputation of the working time information for the employment register data (2016). FDZ-Methodenreport 01/2016
Schimpl-Neimanns, B.: Mikrodaten-Tools: Umsetzung der Berufsklassifikation von Blossfeld auf die Mikrozensen 1973–1998 (2003). ZUMA-Methodenbericht 2003/10
Schmucker, A., Seth, S., Ludsteck, J., Eberle, J., Ganzer, A.: Establishment history panel, 1975–2014 (2016). FDZ-Datenreport 03/2016
Statistisches Bundesamt. Klassifikation der Wirtschaftszweige, Ausgabe 1993. Statistisches Bundesamt, Wiesbaden. https://www.destatis.de/DE/Methoden/Klassifikationen/Gueter-Wirtschaftsklassifikationen/klassifikation-wz-1993.html (2002)
Statistisches Bundesamt. Preise - Verbraucherpreisindizes für Deutschland (Lange Reihe ab 1948). Statistisches Bundesamt, Wiesbaden. https://www.destatis.de/DE/Themen/Wirtschaft/Preise/Verbraucherpreisindex/Publikationen/Downloads-Verbraucherpreise/verbraucherpreisindex-lange-reihen-pdf-5611103.html (2022)
Storm, E.: Task specialization and the native-foreign wage gap: evidence from worker-level data. Labour **36**(2), 167–195 (2022)
vom Berge, P., Schmucker, A.: Creating cross-sectional data and biographical variables with the Sample of Integrated Labour Market Biographies 1975–2019—programming examples for Stata (2021). FDZ-Methodenreport 05/2021
Zwick, T., Bruns, M., Geyer, J., Lorenz, S.: Early retirement of employees in demanding jobs: evidence from a German pension reform. J. Econ. Ageing **22**, 100387 (2022)

Publisher's Note