Dauth, Wolfgang; Eppelsheimer, Johann

Article

Preparing the sample of integrated labour market biographies (SIAB) for scientific analysis: a guide

Journal for Labour Market Research

Provided in Cooperation with:
Institute for Employment Research (IAB)

Suggested Citation: Dauth, Wolfgang; Eppelsheimer, Johann (2020) : Preparing the sample of integrated labour market biographies (SIAB) for scientific analysis: a guide, Journal for Labour Market Research, ISSN 2510-5027, Springer, Heidelberg, Vol. 54, Iss. 1, pp. 1-14, https://doi.org/10.1186/s12651-020-00275-9

Dauth and Eppelsheimer
J Labour Market Res (2020) 54:10
https://doi.org/10.1186/s12651-020-00275-9

Journal for Labour Market Research

ORIGINAL ARTICLE Open Access

Preparing the sample of integrated labour market biographies (SIAB) for scientific analysis: a guide

Wolfgang Dauth[1] and Johann Eppelsheimer[2*]

Abstract

Preparing the Sample of Integrated Labour Market Biographies (SIAB) for scientific analysis is a complicated and error-prone task. This paper elaborates on the steps necessary to prepare the SIAB and provides examples of how the preparation can be done. Among other topics covered, we show how to generate and merge additional variables, impute right-censored wages, deal with parallel employment episodes, and clean the dataset. Finally, we present a case study on the individual long-term effects of job loss from plant closure to demonstrate how our prepared version of the SIAB can be used to carry out an empirical analysis. The supplementary material of this paper contains extensively commented Stata do-files to replicate our data preparation and the subsequent analysis.

Keywords: IAB, SIAB, IEB, Data preparation, Stata, Event study

JEL Classification: C55, C81, J65

1 Introduction

The Sample of Integrated Labour Market Biographies (SIAB) offers data on basically *all* notifications to the social security system for a 2% random sample of all individuals who have ever been registered in the German social security system. This allows researchers to analyze the full labor market biographies of a large number of individuals with daily precision. At the same time, the sample size also permits aggregate analyses at the level of regions, industries, and occupations. The SIAB is an extremely versatile dataset that can be used to answer a virtually unlimited number of research questions. It is also one of the most important data products that the Institute for Employment Research (IAB) provides for the scientific community. Almost 60% of the over 500 projects currently supervised by the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the IAB are based either on the SIAB or the closely related Linked Employer-Employee Data of the IAB (LIAB).[1]

Scholars analyze the SIAB in numerous contexts. A prominent topic is employment polarization. In their influential paper, Goos et al. (2014) use the SIAB to explain employment polarization as a result of routine-based technological change and offshoring of tasks. Other examples are Antonczyk et al. (2018) and Rendall and Weiss (2016), who stress the importance of cohort effects and the apprenticeship system when investigating polarization in Germany. In another research field, Sanchez and Wellschmied (2020) used the SIAB to examine how nonpredictable earning shocks vary by age. They find that positive earning shocks are typical for young workers, whereas negative shocks are common among workers above 50. Furthermore, Riphahn and Schnitzlein (2016) study long-run wage mobility in Germany and identify structural shifts as the major explanation for declining wage mobility after 1990. The SIAB data are also used

*Correspondence: johann.eppelsheimer@iab.de
[2] Institute for Employment Research (IAB), Regensburger Str. 100, 90478 Nuremberg, Germany
Full list of author information is available at the end of the article

[1] More details on the structure and origin of the SIAB can be found in the data description by Antoni et al. (2019).

to analyze the effects of international trade on individuals. For instance, Dauth et al. (2014) show that the rise of China and Eastern Europe caused job losses for workers in import-competing industries, whereas exports had a stabilizing effect on employment relationships. Overall, the positive effects dominate, and Germany gained substantially in employment terms from the rise of the East. Further research based on the SIAB includes Kohlbrecher et al. (2016), who examine the consequences of vacancy free-entry conditions and idiosyncratic productivity shocks on labor market matching functions, and Gehrke and Weber (2018), who show that the short-term benefits of structural labor market reforms are weaker when initialized during recessions than during economic expansions. As these studies illustrate, the SIAB allows high-quality research in numerous areas.

Usually, individual-level datasets differ from macro data in one particular regard: They have a rather complicated structure. The SIAB is certainly no exception. In principle, each row in the SIAB stems from either an employer's notification to the social security system or a process in unemployment insurance. Individual biographies often do not follow a linear path: Individuals change jobs, hold several jobs at the same time, become unemployed, participate in active labor market policy measures, etc. A dataset that covers all those different biographies is necessarily more complicated than the example datasets familiar from econometric textbooks. This means that the researcher must first put in substantial effort to prepare the data before beginning the actual empirical analysis.

Due to its versatility, researchers use the SIAB in many different contexts and each context requires a different preparation of the raw data. Hence, practices that have proven useful in one project might not be applicable to another. At the end of the day, it is not feasible to develop a linear set of guidelines to prepare the SIAB that accommodates all possible needs for all potential projects.

However, we are aware that preparing the SIAB is costly and error-prone, especially for researchers who are inexperienced in working with large administrative datasets. In this paper, we therefore provide a detailed explanation of the best practices that we have found useful in preparing the SIAB for our own research. We additionally provide extensively commented Stata do-files to replicate our data preparation. The goal of this paper and the supplementary collection of do-files is to provide researchers with a step-by-step guideline on how the SIAB can be prepared for individual-level analyses. We point out the purpose of each step and provide an intuitive explanation of how it is implemented. All technical details can be found in the do-files that accompany this paper.

Although this paper provides the first complete set of guidelines on how to prepare the SIAB for scientific analysis, several other reports exist that give guidance on some specific parts of the data preparation process for German social security data. A popular example is Eberle and Schmucker (2017). In their report, the authors explain how to generate biographical variables such as work experience or the duration of unemployment benefit receipts. Another example is Jaenichen (2018), who demonstrates how to measure employment duration. There are also suggestions for identifying unemployment periods (Kruppe et al. 2008) or maternity leaves (Müller et al. 2017; Schönberg 2009). Additionally, several articles propose methods for improving the data quality of German social security data. There are different processes for improving information on educational attainment (Fitzenberger et al. 2006; Hutter et al. 2015; Thomsen et al. 2018), imputing top-coded wages (Dustmann et al. 2009; Card et al. 2013; Gartner 2005), generating time-consistent industry codes Eberle et al. (2011), and correcting part-time/full-time information provided before a structural break in the reporting procedure in 2011 (Fitzenberger and Seidlitz 2020; Ludsteck and Thomsen 2016). The procedures of Fitzenberger et al. (2006) to improve information on education, of Eberle et al. (2011) to generate time-consistent industry codes, and of Ludsteck and Thomsen (2016) to correct information on part-time work are already implemented in the latest version of the SIAB.

This paper and the collection of do-files are based on the weakly anonymous version of the SIAB (years 1975–2017, DOI:10.5164/IAB.SIAB7517.de.en.v1). Access to this dataset is provided to the scientific community by the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB). The usual way to access these data is via on-site use at the FDZ in Nuremberg, Germany, or one of several other locations in Germany, France, the USA, Canada, or the United Kingdom. This allows users to develop their programs while working with the data interactively. Subsequently, users can continue their projects via remote data access using the web interface JoSuA. The FDZ requires users working in the secure on-site environment and via JoSuA to follow certain conventions in writing their programs and managing their files. Our guidelines follow these conventions and our do-files have been tested to run in the JoSuA environment.[2]

[2] Alternatively, researchers can use the factually anonymous version of the SIAB (Regionalfile, Version 7517 v1, DOI: 10.5164/IAB.SIAB-R7517.de.en.v1), which is available as a scientific use file (SUF). However, the more comfortable handling of a SUF comes at the cost of more extensive anonymization. In particular, variables with a high risk of deanonymization, such as nationality, occupation, industry, or region, are aggregated in these files. The Appendix describes the necessary changes to adapt our code to working with the SUF.

The SIAB is a derivative of the Integrated Employment Biographies (IEB). Due to this relatedness our codes can, with some minor adaptations, also be used to prepare the IEB. In particular, it is necessary to rename variables and account for the fact that some variables have already been preprocessed in the SIAB (for example, while the SIAB already has an indicator variable `female`, the IEB has a variable `sex_id`, which takes the value 1 for men and 2 for women).

We strongly encourage all users of this collection to check our code for mistakes and adjust it to suit the requirements of each specific project. We provide this as a service to the scientific community, but we do not assume responsibility for any problems that result from using our code.

We complement our preparation of the SIAB with a case study on the long-term employment effects after an involuntary job loss due to a plant closure, using an event study approach. This kind of analysis is currently very popular in applied labor economics. The goal of this case study is to demonstrate how our prepared version of the SIAB can be used to carry out such an analysis. Our results corroborate the findings of earlier papers from the US and Germany that involuntary job loss leaves large and persistent scars in individual employment biographies.

2 Data preparation procedure

In the following, we present guidance for the preparation of the SIAB. Starting with the original SIAB, provided by the Research Data Center of the IAB, we generate and merge additional variables, impute right-censored wages, take care of parallel episodes, and clean the dataset. Additionally, we show how to transform the spell data format into a yearly panel.

Our preparation of the SIAB follows a modular organization, where each step has a dedicated individual do-file. All Stata do-files for this exercise are provided in the Additional file 1. The complete set of do-files is launched from the file `master.do` and takes around eight hours to run. Table 1 in the Appendix gives an overview of all variables we generate and modify within this guide.

When working on-site at the FDZ or via remote data access, all do-files must be uploaded via JoSuA. They will then be accessible in the path `$orig`. Otherwise, the whole replication folder can be copied into a new working directory, retaining the structure of the subfolders.

2.1 do-file: master.do

The main purpose of the file `master.do` is to configure the Stata environment and launch all parts of the data preparation in the correct order. Within the first lines, we give some standard commands to Stata, e.g., to specify the version under which the code was written, not to pause for –more– messages, and to declare the observation period.

For external users at the FDZ, the working directory and the folders for the original data, prepared data, and any output are preset. We only set the path for storing figures to be equal to the log-folder and create macros for an auxiliary do-file and the original data. Internal users at the IAB may wish to comment out this section of the code and instead use the lines below, where we define and create individual folders for data, graphs, and log-files.

Throughout the data preparation, `master.do` saves an intermediate version of the prepared SIAB as backup (`siab_intermediate.dta`) and a final version (`siab_clean.dta`). We also generate the variables `jahr` (*year*) and `age` at this point.

2.2 do-file: 01_split_episodes.do

Spells in the SIAB are already split into episodes. This means that overlapping spells (e.g., multiple jobs, job search while still employed, etc.) are split such that parallel spells always have the same start and end dates.[3] By definition, employment spells from the Employment History (BEH) always break at the turn of a year. However, episodes from the Benefit Recipient History (LeH) and Unemployment Benefit II Recipient History (LHG) sources might span multiple calendar years. For many applications, for example, when annual panel data are generated, it is useful to have at least one observation per calendar year. `01_split_episodes.do` splits those spells that span more than one calendar year into multiple episodes.

`01_split_episodes.do` also modifies the start and end dates stored in the variables `begepi` and `endepi` accordingly. The original values are stored in `begepi_orig` and `endepi_orig`, respectively. Furthermore, the Stata program updates the variables `jahr` and `age`.

2.3 do-file: 01_SIAB_bio_MODIFIED.do

The SIAB covers the entire employment biographies of the sampled individuals. The information from previous episodes can be used to generate variables that summarize the previous careers, such as job tenure or duration of labor market experience. Eberle and Schmucker (2017) provide do-files that generate several biographical variables, in particular, `tage_erw` (days in employment) and `tage_bet` (tenure at the current plant, restarting the count after breaks). Hence, following the link in the

[3] For details on episode splitting in the original SIAB, refer to Antoni et al. (2019).

according *FDZ-Methodenreport*, we download the file `01_SIAB_bio.do` and apply their code. Note that to make `01_SIAB_bio.do` compatible with our master file, we have to comment out lines 67–73 and 296–301 first and save the file as `01_SIAB_bio_MODIFIED.do`.

2.4 do-file: 02_merge_BHP.do

For many research projects, it is useful to have data on the individual's establishment, such as location and industry codes, number of employees, average wages, etc. The Establishment History Panel (BHP) consists of the administrative data of the full universe of all employees liable to social security on June 30 of a given year, aggregated to the level of establishments. While it would be possible to merge the full BHP data to the SIAB, a special version of the BHP, the *Basis Establishment File*, that only comprises establishment/year combinations that appear in the SIAB has been prepared.[4] In the file `02_merge_BHP.do`, we merge this version of the BHP to the SIAB.

2.5 do-file: 03a_industries_1digit_destatis.do and 03b_industries_1digit_iab.do

The BHP includes several detailed industry classifications. However, studies are often also interested in broader classifications. We therefore translate 3-digit industry codes into two alternative 1-digit aggregates.

For the mapping, we use the variable `w93_3_gen`, which holds a time-consistent version of the 3-digit German equivalent of the NACE Rev. 1 (Eberle et al. 2014). We map these industries to classifications from the Federal Statistical Office (Statistisches Bundesamt 2002) and the classification scheme from the IAB establishment panel. We provide two do-files for the mapping. The first file, `03a_industries_1digit_destatis.do`, generates 15 industries, stored in the variable `industry1_destatis`. The second file, `03b_industries_1digit_iab.do`, creates nine 1-digit industries, stored in `industry1_estpanel`. The main difference between the two classifications is the level of detail within the primary sector and public services.

2.6 do-file: 04_occ_blossfeld.do

The weakly anonymous version of the SIAB also includes very detailed occupational information. For some applications, this is too detailed, and information on broad categories would be sufficient. We thus add the widely used occupational classification of Blossfeld (1987) to the dataset (see also Schimpl-Neimanns 2003). The do-file `04_occ_blossfeld.do` creates the variable `occ_blo`, which contains 13 groups of occupations. These 13 groups are generated by recoding the 3-digit occupational codes (*Klassifizierung der Berufe*) from 1988 in the variable `beruf`. Note that we use the classification of occupations from 1988 because they are readily transferable into the scheme of Blossfeld. Alternatively, researchers could also create broader occupational classes based on the variable `beruf2010_3`, which, has included more recent occupational codes from 2010.

2.7 do-file: 05_educ_broad.do

Information on the highest level of education attained by individuals is often inconsistent in German administrative data. For example, individuals are registered with a degree from a university for some periods, and in subsequent periods, their highest educational degree is an apprenticeship. To correct such implausible developments in educational attainments, the FDZ provides an imputed version of `ausbildung`, stored as `ausbildung_imp`. The imputation procedure of the FDZ builds on Fitzenberger et al. (2006).

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

2.8 do-file: 06_wages_assessment_ceiling.do

Since the data stem from compulsory notifications to the social security system, wage information in the SIAB is highly reliable in general. However, because of this administrative origin, wages are only reported until they reach the upper earnings limit for statutory pension insurance. If they exceed this threshold, the wages are coded with this value. This assessment ceiling differs by year and location.[5] We gather the appropriate assessment ceilings in the do-file `06_wages_assessment_ceiling.do` and store them in the variable `limit_assess`. Furthermore, the do-file creates the variable `east`, which takes the value of one if the workplace of an individual is in the Eastern part of Germany and zero if it is in the Western part of the country.

2.9 do-file: 07_wages_marginal.do

Another important threshold is the marginal part-time income threshold. Jobs with wages below this threshold are either exempt from social security contributions (before 1999) or subject to a lump-sum contribution

[4] For details on the BHP see Schmucker et al. (2016).

[5] The Research Data Center of the IAB provides an exhaustive list of the earnings limits: http://doku.iab.de/fdz/Bemessungsgrenzen_de_en.xls.

to be paid by the employer (1999 or later). In principle, those jobs are only covered in the data from 1999 on (Antoni et al. 2019). We provide the do-file `07_wages_marginal.do` to mark employment spells with wages below this threshold. This do-file creates the variable `limit_marginal`, which stores the marginal part-time income threshold. Moreover, the do-file flags observations with wages below the marginal part-time income threshold using the dummy variable `marginal`.

2.10 do-file: 08_wages_deflation.do

To make wages comparable across different time periods, we calculate real wages (`wage_defl`). For the deflation, we divide nominal wages by the consumer price index from the Federal Statistical Office (Statistisches Bundesamt 2019). In addition, the file `08_wages_deflation.do` also deflates the assessment ceilings (`limit_assess_defl`) and the marginal part-time income thresholds (`limit_marginal_defl`). We use a specific consumer price index for West Germany for the years 1975–1991 and a general price index for the whole country for years that are more recent. The base year is 2015.

2.11 do-file: 09_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% of all spells for workers between 1975 and 2017, the proportion of censored observations within certain subgroups is substantial. For instance, 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 the 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, the do-file `09_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 assessment ceiling for all observations.[6] 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).[7] By default, the do-file `09_wages_imputation.do` first clusters observations by year, East and West Germany, and three education groups. For each of these clusters, we fit Tobit wage equations, controlling for worker characteristics X.[8]

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 and store imputed wages in the variable `wage_imp`. Although it is

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

[7] 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.

[8] This is done by saving the dataset in a temporary file and then running the Tobit regressions on smaller subdatasets, each containing only one cluster. After imputation, these subdatasets are appended to obtain the full dataset with imputed wages. An alternative would be to keep the full dataset 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 subdatasets.

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. Note also that to speed up the computations, our imputation procedure stores local images of the data on the disk. When working with large samples of the IEB, one might exceed the local storage limits. In such cases, users need to manually assign a storage location on their network drive.

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 the do-file `09_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. The potential bias depends on the correlation between wage and the omitted covariates.[9] Additionally, it might also be reasonable to choose different subgroups from the ones we suggest in `09_wages_imputation.do`.

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.[10]

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

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

2.12 do-file: 10_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. The do-file `10_parallel_episodes.do` restricts the dataset to the main episodes but retains potentially relevant information from parallel spells. In particular, we count the number of parallel spells and jobs (`nspell`, `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`). Information from the main spell remains unchanged. Depending on the research question it might be preferable to investigate data from the main spell or summary data of parallel spells.[11]

Various approaches exist to identifying the main episode. By default, `10_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.

2.13 do-file: 11_yearly_panel.do

Since spell data are 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 dataset, we 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`).

By default, `11_yearly_panel.do` uses June 30 as the cutoff date because data from the BHP are measured on the last day of June (Schmucker et al. 2016). Of course, there might be situations where a different cutoff date

[11] 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. 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.

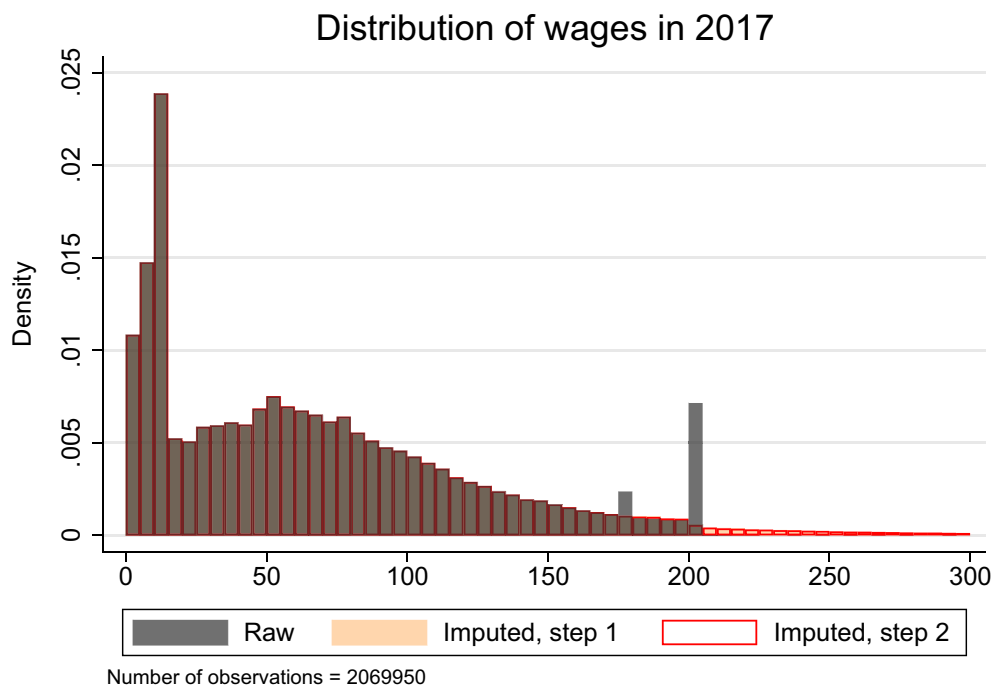

Fig. 1 Distribution of daily wages in 2017. Note: The figure reports the distribution of the censored daily wage (in 2015 Euros) and its imputed equivalent after the first and second imputation steps

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 dataset 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 dataset 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.

2.14 do-file: 12_restrictions.do

One important part of the preparation of the SIAB for scientific analysis is to restrict the data to a sensible subsample. For instance, many projects are only interested in specific time periods, regions, employment statuses, or sociodemographic groups. To simplify the restricting of the data, we gather typical sample restrictions in the do-file `12_restrictions.do`. By default, all restrictions are commented out, and the program leaves the data unchanged.

Note that file `12_restrictions.do` is also useful when working on a research agenda with several related projects. For instance, scholars could generate one clean base version of the SIAB and then create "offsprings" of the base version by executing `12_restrictions.do` with different settings tailored to the specific project.

2.15 do-file: 13_clean_up.do

Finally, we use the do-file `13_clean_up.do` to sort the data, declare the panel structure and compress variables.[12] If variables are irrelevant for the data analysis, it might also be advisable to drop them using `13_clean_up.do`.

The master file saves the final dataset as `siab_clean.dta`.

3 Case study: using the SIAB to study long-run costs of job loss

In this section, we demonstrate the SIAB's usefulness for carrying out analyses that require comprehensive information on individual employment biographies. One recurring and currently active field of research in labor economics is to analyze the long-run effects of involuntary job loss. In an early and very influential study, Jacobson et al. (1993) analyze the earnings losses of previously high-tenured workers who lost their jobs during a mass layoff. They employ an event-study design to analyze not only the magnitude of those earnings losses but also the temporal dynamics of earnings losses before and after the layoff. The authors find that, in the long run, displaced workers lose approximately 25% of their annual income. Remarkably, this effect is very persistent: On average, laid-off workers never catch up to their previous income level.

The paper by Jacobson et al. (1993) offers two main innovations. First, the study concentrates on workers who were displaced by a mass layoff. Administrative datasets usually contain no information on the reasons for a separation, i.e., whether it was voluntary, self-inflicted by a worker's low productivity or misbehavior, or caused by reasons a worker cannot influence. Mass layoffs, by contrast, are arguably exogenous from an individual worker's perspective. Second, the authors employ an event-study estimation to examine the temporal dynamics of the effect of a layoff on an individual's employment outcomes. An event study is related to the more general difference-in-differences framework. It allows us to examine the impact of experiencing an "event", e.g., a layoff, at several points in time before and after the event. Since the seminal contribution by Jacobson et al. (1993), a large number of studies have adopted this framework to analyze various aspects of job loss. For example, Davis and von Wachter (2011) use the same approach to corroborate the earlier finding of long-lasting negative effects of layoffs and show that the effects are even worse if the layoff happens during an economic downturn. Several recent working papers use IAB data and show that income losses after mass layoffs are of similar magnitude and persistence in Germany and in the US (Schmieder et al. 2012; Fackler et al. 2017; Schmieder et al. 2020). Since these papers only apply the event-study approach, they are often somewhat vague regarding the details of the empirical model. By contrast, Schmidheiny and Siegloch (2019) present a precise explanation of the interpretation of this technique and how to apply it in practice. Many more as yet unpublished works use some version of the event-study approach. In fact, in the last few years, it has been almost impossible to attend a workshop or conference in labor economics without at least one paper using this method. Given that the method is so *en vogue* at the moment, we demonstrate how to carry out such an analysis using the SIAB.

The goal of the following case study is to analyze the individual consequences of displacement due to the closure of a plant. Previous studies on this topic typically consider displacements due to mass layoffs. Aside from permanent closures, this term also covers workforce reductions where a firm loses a significant share of its initial size. Both cases raise the danger of mistaking restructuring of firms with several plants for actual mass layoffs. IAB datasets have unique plant identifiers but do not offer any information on which plants belong to the same firm. A large number of workers leaving one plant might look like a mass layoff but could also stem from workers being assigned to a different plant within the same firm. Hethey-Maier and Schmieder (2013) propose to look at the "maximum clustered outflow", i.e., the group of workers from one plant entering the same new plant. If this cluster is very large compared to the plant's previous size, then one would suspect that this is because of restructuring. Replicating the analysis of Hethey-Maier and Schmieder (2013) would require access to the full sample of all employees of each plant and not just the SIAB's two-percent sample. Unfortunately, it is not possible to use the SIAB to identify cases where plants fire a large share of their workforce but continue to exist. It is also not possible to directly observe whether a plant disappears. However, this information can be inferred from an auxiliary dataset of the *Establishment History Panel* (BHP), which offers information on incidents where plant IDs disappear from the IAB records. These data are also part of the shortened version of the BHP described in "Data preparation procedure" section but need to be explicitly requested and justified in the original application for working with the SIAB. The file `SIAB_7517_v1_bhp_exit_v1.dta` consists of observations of all plants on June 30 of a given year that disappear from IAB data before June 30 of the following year. Merging

[12] 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 dataset to the SIAB produces only matches for those cases where a worker is employed at a plant that ceases to exist within the subsequent 365 days and therefore identifies potential cases of plant closures. This auxiliary dataset adds five new variables: the number of employees in the existing plant, the size of the largest cluster (i.e., the number of workers moving to the same new plant), the number of employees of the plant where the largest cluster moves to, a dummy that indicates if this plant is newly established, and a categorical variable called `austritt`, the type of exit. `austritt` is classified according to the heuristic of Hethey-Maier and Schmieder (2013). We define a person who experienced a plant closure within the next year as a person who was employed at a plant that appears in the auxiliary dataset and where `austritt` either indicates an "atomized death" (the maximum clustered outflow is no more than 30% of the plant's original size) or "chunky death" (between 30 and 80 percent of the plant's original size).[13] While we cannot rule out that some cases in the other categories of disappearances of plant IDs are also related to true plant closures, this procedure minimizes the risk of including false-positive cases.

In the remainder of this section, we describe how we prepare the SIAB for an event study and how we carry out the actual analysis. The do-files for this exercise can be found in the folder `case_study` in the Additional file 1. Throughout this case study, we follow the "manual" on event studies by Schmidheiny and Siegloch (2019).

3.1 do-file: case_study/master.do

After running the SIAB preparation from the Additional file 1 with the default settings, our event study analysis starts again with a master do-file `master.do`. Here, we configure the Stata environment and launch all parts of the preparation and analysis in the correct order. The structure of the subfolders is the same as in the previous data preparation procedure to ensure that our code also works within the FDZ environment.

First, we define macros to delineate the time period of the analysis. We are ultimately interested in plant closures that occurred from 1990 onwards. Note that we cannot observe the exact timing of the plant closure. We define the event of a plant closure to occur in year t if a plant existed in the IAB data on June 30 of year t but not on June 30 of year $t+1$. Since the current version of the SIAB includes information up to the year 2017, this implies that the last possible year for an "event" is 2016. The observation window for events is therefore 1990–2016. As laid out by Schmidheiny and Siegloch (2019), the actual window of analysis must be shorter than the observation window for events. The intuition behind this is that a plant closure that occurred before the start of the window of analysis can already influence the outcome. Analogously, to test for pretrends, one must also allow for events to happen after the end of the window of analysis. In this analysis, we want to observe pretrends of up to 3 years before the plant closure and the evolution of the outcome up to 5 years after the event. We therefore define the window of analysis from 1995 ($= 1990 + 5$) to 2013 ($= 2016 - 3$). The final dataset then contains mostly individuals observed before and after the plant closure but also a number of individuals observed only after or only before the event.

Next, we load the yearly panel version of the SIAB prepared in the first part of this paper. To use server resources efficiently, we only load observations that lie within the previously defined observation window for events.

3.2 do-file: case_study/1_find_layoffs.do

In the first do-file, we identify workers who experienced a plant closure according to our definition any year between 1990 and 2016. We first merge the auxiliary BHP information on the nature of disappearing plant IDs. Then, we create a variable that tags workers who experienced one of those events and a variable with the year of the respective event (the year before the plant disappears).

Next, we further restrict the sample to individuals who held a normal full-time job at the same plant for two consecutive years before the plant closure and who were young enough to still be on the labor market for ten more years before retirement. This ensures that we only compare workers who were attached to their original plant and would not have left the plant shortly anyway. The resulting dataset has 478,710 observations of 30,719 treated individuals.

3.3 do-file: case_study/2_event_study.do

In this section, we explain how we estimate an event study according to Schmidheiny and Siegloch (2019):

$$y_{it} = \sum_{j=-3}^{5} \beta_j b_{it}^j + \mu_i + \theta_t + \varepsilon_{it} \tag{1}$$

where y_{it} is the outcome of individual i in calendar year t. μ_i is an individual fixed effect, θ_t is a calendar year fixed effect, and ε_{it} is the error term. The variables of interest are the time-to-/since-event dummies b_{it}^j, where $j = -3, -2, 0, 1, 2, 3, 4, 5$. These variables indicate

[13] We acknowledge that it is disputable whether the latter category is already susceptible to being related to restructuring. We leave a deeper analysis of this classification to the researcher applying our code.

whether an individual is observed j years before/after plant closure. The coefficients of those dummies represent the difference in the outcome j years before/after the plant closure relative to the reference point one year prior to the plant closure. Note that the endpoints at $j = -3$ and 5 are treated specifically. These dummies indicate whether the individual is observed 3 *or more* years before and 5 *or more* years after the event. This procedure is called "binning of the endpoints of the observation window for events" and ensures that the β_j can be interpreted as dynamic treatment effects even in the absence of any never-treated individuals in the data (Schmidheiny and Siegloch 2019).

Before we carry out the actual analysis, we first define the outcome variable. Potential outcomes of interest are the days in employment per calendar year, the (log) annual earnings, or the (log) average daily wage. Even though we work with the yearly panel version of the SIAB, the outcome variables have daily precision. We construct them using the variables `year_labor_earn` and `year_days_emp`, which were generated in "Data preparation procedure" section using the original spell data. Finally, we generate the time-to-/since-event dummies.

The actual event study estimation is a regression analysis of the outcome variable on the time-to-/since-event dummies and fixed effects for calendar years and individuals. We restrict the estimation to the years $1995 - 2013$, so that the window of analysis is smaller than the observation window for events. We cluster the standard errors to allow the error terms to be correlated among all observations for the same individual. Note that the estimation does not contain any further control variables. Many individual characteristics and the heterogeneity of the closing plants are at least covered to a large extent by the individual fixed effects. A linear age term cannot be identified since it is perfectly collinear with the combination of individual and calendar year fixed effects. One might be tempted to control for time-varying characteristics such as tenure or occupation. We refrain from including those variables in the model since for all observations with $j > 0$, the values might be the outcome of having experienced a plant closure.

After the regression analysis, we create a new variable that runs from -3 to 5 for the first nine observations and has missing values thereafter. We then define three further variables where we save the coefficients of the time-to-/since-event dummies and the upper/lower bounds of their 95-percent confidence intervals. These four new variables can easily be plotted using Stata's 'twoway' commands, resulting in a comprehensible visualization of the dynamic treatment effects of a plant closure on individuals' average outcomes.

Figure 2 reports the results of an event study for the annual days in employment as the outcome variable. Since $j = -1$ is the reference point, the respective coefficient is equal to zero by construction. The coefficient at $j = -2$ is zero as well since we have restricted the sample to workers with at least two years of tenure at the closing plant. The coefficient at $j = -3$ is remarkably smaller. This stems from the fact that there is substantial mobility in and out of employment in the SIAB and that stable employment biographies are somewhat less common than what is often assumed. However, this coefficient is even smaller (i.e., more negative) than the actual difference between $j = -3$ and $j = -1$ due to the binning of the endpoints, which means that the coefficient of $j = -3$ represents *all observations* 3 *or more* years after the event.

There is a sharp drop in the annual employment duration of approximately 18 days in the year of the event and of approximately 61 days in the year after the event. The coefficient of $j = 0$ should be interpreted with caution since the actual plant closure occurred on some day between June 30 of year 0 and June 29 of year 1. From $j = 2$ onwards, the average worker seems to recover from the initial shock. However, this recovery is neither fast nor complete. Even four years after the plant closure, the displaced individuals spend 40 fewer days in employment per year than before the event. The coefficient of $j = 5$ drops slightly, which again stems from the binning of the endpoints. The overall picture corroborates the previous results on the long-run effects of involuntary displacement. Aside from the direct costs of displacement, this shock leaves a deep scar in an individual's employment biography. The average worker never fully recovers from the layoff.

This case study demonstrates how the prepared SIAB dataset can be used to carry out a state-of-the-art empirical analysis in labor economics. Of course, this analysis is only exemplary, and we do not claim that it is comprehensive or innovative. We leave it to the researchers applying our preparation procedure for the SIAB to push the frontier of labor market research.

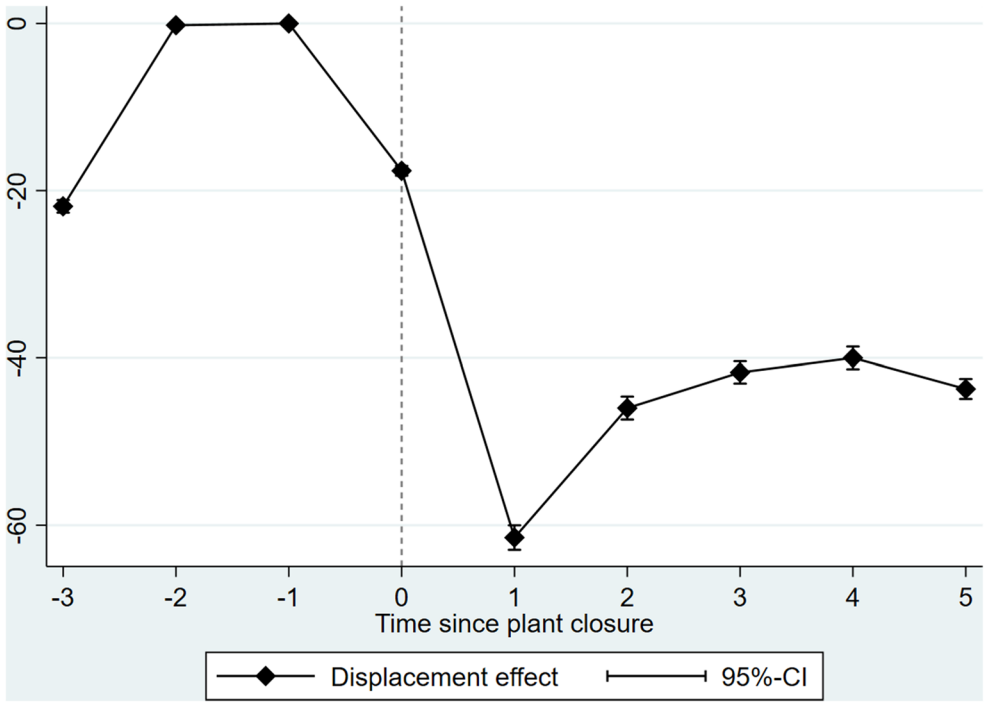

Fig. 2 Event study results for annual employment days. Note: The figure reports the coefficients of an event study analysis. The dependent variable is the days in employment in the calendar years before/after a plant closure. The actual plant closure occurred between June 30 of year 0 and June 29 of year 1. $N = 478,710$ observations of 30, 719 individuals

4 Concluding remarks

Ever since the first version of the SIAB was made available to the scientific community, people have been asking the IAB to provide either a fully prepared version of the SIAB or a field manual that explicitly spells out the whole process of preparing the data. While having such a manual is certainly convenient, it also brings along a number of problems. It is extremely difficult to account for the idiosyncratic needs of each research project. There are several ways to deal with challenges in the data, such as parallel spells or censored wages, and often there is no consensus on which way is the correct one. Providing a manual on data preparation would raise the risk of consolidating a status quo where decisions should actually be made by the researcher.

While we are aware of these problems, we are also aware that preparing the SIAB is a very complicated task, especially for researchers who have little experience in preparing administrative data. Even experienced researchers might not be aware of some issues faced when preparing the SIAB for scientific analysis.

Our goal is not to provide an exhaustive manual on the preparation of the SIAB. We rather seek to point out which steps are necessary to prepare these data and to provide some examples of possible ways in which the preparation could be done.

This collection of best practice examples is a combination of codes used by the authors and some of their colleagues. Running our do-files on the SIAB will result in a prepared version of this dataset that could potentially serve as the starting point for a large number of different research projects in applied micro labor economics or sociology. We demonstrate one of those possible applications in our case study on the long-run effects of involuntary displacement due to plant closure.

We have made enormous efforts to minimize the number of mistakes in our code. However, we can guarantee neither that there are no mistakes left nor that this collection is complete. We strongly encourage all users of this collection to check our code for mistakes and adjust it to suit the requirements of each specific project. We assume no responsibility for any problems that result from using our code.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12651-020-00275-9.

Additional file 1. The supplementary material contains extensively commented Stata do-files to replicate our data preparation and the subsequent analysis.

Acknowledgements
We thank Manfred Antoni, Melanie Arntz, Ann-Christin Bächmann, Johanna Eberle, Andreas Ganzer, Nina Gläser, Peter Haller, 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, Philipp vom Berge, Florian Zimmermann, and three anonymous referees for many useful comments and suggestions and for sharing their code. We also thank the Research Data Centre (FDZ) of the Federal Employment Agency at IAB to help us verify that our do-files also run within the JoSuA-environment. All errors remain our own.

Authors' contributions
Both the authors read and approved the final manuscript.

Funding
This study was conducted while both authors were employees of the IAB and one author was a scholarship holder in the joint graduate program of the IAB and the University of Erlangen-Nuremberg (GradAB). However, the authors did not receive any specific funding for this particular study.

Availability of data and materials
The datasets generated and analyzed during the current study are not publicly available due to data protection laws (social security data) but can be accessed on-site at the Institute for Employment Research (IAB) subject to an individual data protection agreement. All program codes necessary to replicate the study are available in the Additional file 1.

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

Author details
[1] Institute for Employment Research (IAB), University of Würzburg, IZA, Regensburger Str. 100, 90478 Nuremberg, Germany. [2] Institute for Employment Research (IAB), Regensburger Str. 100, 90478 Nuremberg, Germany.

Appendix

Description of necessary changes to use our programs with the SUF of the SIAB-Regionalfile

do-file: Eberle_Schmucker2017/01_SIAB_bio.do

Instead of unique establishment identifiers, the SUF only contains individual-specific establishment counters. To compute *days in establishment* and *days in job*, Eberle and Schmucker (2017) use the establishment identifiers in the SIAB. To make their program applicable to the SUF, users could use the individual-specific establishment counter `bnn` instead of the establishment identifier `betnr`.

do-file: 02_merge_BHP.do

Because the scientific use file of the SIAB does not contain establishment identifiers, it is not possible to merge establishment data. Hence, the program `02_merge_BHP.do` does not work with the SUF.

do-file: 03a_industries_1digit_destatis.do and 03b_industries_1digit_iab.do

The SUF does not contain time-consistent industry classification codes from 1993 (`w93_3_gen`). Thus, our mapping of 3-digit industry codes to broader industry codes does not work. However, users of the SUF could instead use the variable `w08_gen_gr`, which already contains a broad set of industries based on the classification scheme from 2008.

do-file: 04_occ_blossfeld.do

To create occupational classifications based on Blossfeld (1987), we use 3-digit occupation codes from 1988. These codes are not available in the SUF. However, users could instead use the broader set of occupations `beruf_gr` to create occupational classifications similar to those in Blossfeld (1987).

do-file: 06_wages_assessment_ceiling.do

The Eastern and Western parts of Germany have different earnings limits for statutory pension insurance. To distinguish between workplaces in East and West Germany, we create an indicator for East German workplaces `east` from federal state codes in `ao_bula`. Since this variable is not available in the SUF, users should use the information on locations in `ao_region` instead.

do-file: 09_wages_imputation.do

The procedure to impute top-coded wages described in "Data preparation procedure" section computes the *leave-one-out* mean wages of workers and establishments. Since there is no unique establishment identifier in the SUF, it is not possible to compute leave-one-out means for establishments. Users therefore have to skip the corresponding step in the imputation procedure. Accordingly, users should adapt the regression model in the second part of the imputation procedure.

List of generated or modified variables

See Table 1.

Table 1 Generated and modified variables

Variable name	Short description	Do-file
Age	Age (in years)	Master.do, 01_split_episodes.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)	09_wages_imputation.do
East	1 if workplace in East Germany (incl. Berlin); 0 if West	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
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	03a_industries_1digit_destatis.do
Industry1_estpanel	Industry; 1-digit; IAB establishment panel; based on w93_3_gen	03b_industries_1digit_iab.do
Jahr	Year	Master.do, 01_split_episodes.do
Limit_assess	Contribution assessment ceiling	06_wages_assessment_ceiling.do
Limit_marginal	Marginal part-time income threshold	07_wages_marginal.do
Limit_assess_defl	Contribution assessment ceiling, deflated (2015)	08_wages_deflation.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	10_parallel_episodes.do
Occ_blo	Blossfeld occupations	04_occ_blossfeld.do
Parallel_benefits	Indicator for receipt of UI benefits	10_parallel_episodes.do
Parallel_jobs	Number of parallel jobs	10_parallel_episodes.do
Parallel_wage	Total wage of all parallel employment spells	10_parallel_episodes.do
Parallel_wage_imp	Total imputed wage of all parallel employment spells	10_parallel_episodes.do
Wage	Daily wage, not imputed, top-coded wages replaced by assessment ceiling (−4 EUR), deflated (2015)	09_wages_imputation.do
Wage_defl	Daily wage, deflated	08_wages_deflation.do
Wage_imp	Imputed daily wage, deflated (2015)	09_wages_imputation.do
Year_days_benefits	Total days benefit receipt per calendar year	11_yearly_panel.do
Year_labor_earn	Total labor earnings per calendar year	11_yearly_panel.do
Year_days_emp	Total days employed per calendar year	11_yearly_panel.do

Received: 8 July 2019 Accepted: 25 July 2020
Published online: 26 August 2020

References

Antonczyk, D., DeLeire, T., Fitzenberger, B.: Polarization and rising wage inequality: comparing the US and Germany. Econometrics **6**(2), 20 (2018)
Antoni, M., Schmucker, A., Seth, S., vom Berge, P.: Sample of integrated labour market biographies (siab) 1975 - 2017. Institute for Employment Research, Nuremberg. FDZ-Datenreport 02/2019, (2019) http://doku.iab.de/fdz/reporte/2019/DR_02-19.pdf
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. Quarterly. J. Econ. **128**(3), 967–1015 (2013)
Dauth, W., Findeisen, S., Südekum, J.: The rise of the east and the far east: German labor markets and trade integration. J. Eur. Econ. Assoc. **12**(6), 1643–1675 (2014)
Davis, S.J., von Wachter, T.: Recessions and the costs of job loss. Brookings. Papers. Econ. Activity. Fall. **2011**(1), 1–55 (2011)
Dustmann, C., Ludsteck, J., Schönberg, U.: Revisiting the German wage structure. Quarterly. 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: Simple heuristic based on the Etstablishment History Panel (BHP). Institute for Employment Research, Nuremberg. FDZ-Methodenreport, (2011) http://doku.iab.de/fdz/reporte/2011/MR_05-11_EN.pdf
Eberle, J., Jacobebbinghaus, P., Ludsteck, J., Witter, J.: Generation of time-consistent industry codes in the face of classification changes. Institute for Employment Research, Nuremberg. FDZ-Methodenreport 05/2011, (2014) http://doku.iab.de/fdz/reporte/2011/MR_05-11_EN.pdf
Eberle, J., Schmucker, A.: Creating cross-sectional data and biographical variables with the Sample of Integrated Labour Market Biographies 1975-2014 - programming examples for Stata. Institute for Employment Research, Nuremberg. FDZ-Methodenreport 06/2017, (2017) http://doku.iab.de/fdz/reporte/2017/MR_06-17_EN.pdf
Fackler, D., Müller, S., Stegmaier, J.: Explaining wage losses after job displacement: Employer size and lost firm rents. IWH Discussion Papers, No. 32/2017 (2017)
Fitzenberger, B., Osikominu, A., Völter, R.: Imputation rules to improve the education variable in the IAB Employment Subsample. Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften **126**(3), 405–436 (2006)
Fitzenberger, B., Seidlitz, A.: The 2011 break in the part-time indicator and the evolution of wage inequality in Germany. J. Lab. Market. Res. **54**(1), 1–14 (2020)
Gartner, H.: The imputation of wages above the contribution limit with the german iab employment sample. Institute for Employment Research,

Nuremberg. FDZ-Methodenreport 02/2005, (2005) http://doku.iab.de/fdz/reporte/2005/MR_2.pdf
Gehrke, B., Weber, E.: Identifying asymmetric effects of labor market reforms. Eur. Econ. Rev. **110**, 18–40 (2018)
Goos, M., Manning, A., Salomons, A.: Explaining job polarization: routine-biased technological change and offshoring. Am. Econ. Rev. **104**(8), 2509–26 (2014)
Hethey-Maier, T., Schmieder, J.F.: Does the use of worker flows improve the analysis of establishment turnover? Evidence from German administrative data. Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften **133**(4), 477–510 (2013)
Hutter, C., Möller, J., Penninger, M.: Reducing the need for heuristic rules-an iterative algorithm for imputing the education variable in SIAB. Schmollers Jahrbuch. Zeitschrift für Wirtschafts-und Sozialwissenschaften **135**(3), 355–388 (2015)
Jacobson, L.S., LaLonde, R.J., Sullivan, D.G.: Earnings losses of displaced workers. Am. Econ. Rev. **83**(4), 685–709 (1993)
Jaenichen, U.: Do we measure employment durations correctly?: The case of German administrative employment data. Technical report, Institute for Employment Research, Nuremberg. FDZ-Methodenreport 10/2018, (2018) http://doku.iab.de/fdz/reporte/2018/MR_10-18_EN.pdf
Kohlbrecher, B., Merkl, C., Nordmeier, D.: Revisiting the matching function. J. Econ. Dyn. Control **69**, 350–374 (2016)
Kruppe, T., Müller, E., Wichert, L., Wilke, R.A.: On the definition of unemployment and its implementation in register data-the case of Germany. J. Contextual. Econo. **128**(3), 461 (2008)
Ludsteck, J., Thomsen, U.: Imputation of the working time information for the employment register data. Institute for Employment Research, Nuremberg. FDZ-Methodenreport 01/2016, (2016) http://doku.iab.de/fdz/reporte/2016/MR_01-16_EN.pdf
Müller, D., Strauch, K., et al.: Identifying mothers in administrative data. Institute for Employment Research, Nuremberg. FDZ-Methodenreport 13/2017, (2017) http://doku.iab.de/fdz/reporte/2017/MR_13-17_EN.pdf
Rendall, M., Weiss, F.J.: Employment polarization and the role of the apprenticeship system. Eur. Econ. Rev. **82**, 166–186 (2016)
Riphahn, R.T., Schnitzlein, D.D.: Wage mobility in east and West Germany. Labour. Econ. **39**, 11–34 (2016)
Sanchez, M., Wellschmied, F.: Modeling life-cycle earnings risk with positive and negative shocks. Rev. Econ. Dyn. (2020)
Schimpl-Neimanns, B.: Mikrodaten-tools: Umsetzung der Berufsklassifikation von Blossfeld auf die Mikrozensen 1973-1998. ZUMA, Mannheim. ZUMA-Methodenbericht 2003/10 (2003)
Schmidheiny, K., Siegloch, S.: On event study designs and distributed-lag models: Equivalence, generalization and practical implications. CEPR Discussion Paper No. DP13477 (2019)
Schmieder, J.F., Von Wachter, T., Bender, S.: The long-term effects of UI extensions on employment. Am. Econ. Rev. **102**(3), 514–19 (2012)
Schmieder, J.F., von Wachter, T., Heining, J.: The costs of job displacement over the business cycle and its sources: Evidence from germany. Working Paper (2020)
Schmucker, A., Seth, S., Ludsteck, J., Eberle, J., Ganzer, A.: Establishment history panel, 1975-2014. Institute for Employment Research, Nuremberg. FDZ-Datenreport 03/2016, (2016) http://doku.iab.de/fdz/reporte/2016/DR_03-16_EN.pdf
Schönberg, U.: Does the IAB employment sample reliably identify maternity leave taking?. A data report. Zeitschrift für ArbeitsmarktForschung **42**(1), 49–70 (2009)
Statistisches Bundesamt (2002). Klassifikation der Wirtschaftszweige, Ausgabe 1993. Statistisches Bundesamt, Wiesbaden. https://www.destatis.de/DE/Methoden/Klassifikationen/GueterWirtschaftklassifikationen/Content75/KlassifikationWZ93.html
Statistisches Bundesamt (2019). Preise - Verbraucherpreisindizes für Deutschland (Lange Reihe ab 1948). Statistisches Bundesamt, Wiesbaden. https://www.destatis.de/DE/Themen/Wirtschaft/Preise/Verbraucherpreisindex/_inhalt.html
Thomsen, U., Ludsteck, J., Schmucker, A.: Skilled or unskilled-improving the information on qualification for employee data in the iab employee biography. Institute for Employment Research, Nuremberg. FDZ-Methodenreport 09/2018, (2018) http://doku.iab.de/fdz/reporte/2018/MR_09-18_EN.pdf

Publisher's Note