Research Data Centre (FDZ)
of the German Federal
Employment Agency (BA)
at the Institute for
Employment Research (IAB)

FDZ-Methodenreport

Methodological aspects of labour market data

12/2017
EN

# Combining the Waves of the IAB Establishment Panel

## a Do-file for the Basic Data Preparation of a Panel Data Set in Stata

Matthias Umkehrer

Bundesagentur für Arbeit

# Combining the Waves of the IAB Establishment Panel

## a Do-file for the Basic Data Preparation of a Panel Data Set in Stata

Matthias Umkehrer (Institute for Employment Research)

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

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

## Contents

List of tables

List of figures

## Abstract

This FDZ-Methodenreport introduces the do-file *BP_panelgen_93YY_v1_en* for combining all waves of the IAB Establishment Panel from 1993 to the most recent year to one panel data set in Stata. It explains the initial problem, outlines the structure of the do-file, describes the panel data set and provides operating instructions. The variables in the resulting panel data set are consistently named, coded and labelled to provide a standardized basis for panel analyses. The program code is kept simple and transparent so that users can easily adjust it if necessary. However, caution has to be taken as generated panel variables might still differ in content across waves due to changes in the questionnaire, for instance.

## Zusammenfassung

Dieser FDZ-Methodenreport stellt das Do-file *BP_panelgen_93JJ_v1* vor, welches alle Wellen des IAB-Betriebspanels von 1993 bis zum aktuellsten Jahr zu einem Paneldatensatz in Stata zusammenfügt. Er erläutert die Ausgangslage, skizziert den Aufbau des Do-files, beschreibt den Paneldatensatz und bietet Bedienhinweise. Die Variablen im generierten Paneldatensatz sind konsistent benannt, kodiert und beschriftet, um eine standardisierte Grundlage für Panel-Analysen zu schaffen. Der Programmcode ist einfach und transparent gehalten, sodass Nutzerinnen oder Nutzer bei Bedarf leicht Anpassungen vornehmen können. Allerdings gilt es zu beachten, dass sich die generierten Panelvariablen, zum Beispiel aufgrund von Änderungen der Fragestellung, noch immer in ihrem Inhalt über die Wellen hinweg unterscheiden können.

**Keywords:** IAB Establishment Panel; data preparation; panel data; data management

### Notes and thanks

The individual authors are responsible for the content of the FDZ Datenreports and Methodenreports. Matthias Umkehrer produced the core of the Stata syntax and the FDZ Methodenreport. Sandra Dummert helped completing the syntax. Special thanks are due to Jochen Späth for sharing his routines for data preparation, to Stefanie Wolter, Dana Müller, Jens Stegmaier and Tina Hinz for valuable comments and suggestions, as well as to Jasmin Werle, Falk Voigt, Lisa Schmidtlein, Florian Heinritz, Marie Sophie Jestadt and Stefan Scholz for their supporting work.

### Data availability

The data set described in this paper is available for use for professional researchers. For further information, see ‘Establishment Data’ on the website: http://fdz.iab.de/en.aspx.

# 1 Existing situation

The IAB Establishment Panel is a representative survey among establishments carried out annually since 1993. By providing comprehensive information on a large number of employment policy related topics, such as personnel development, workforce structure, business policy, investment activities, public funding, or vocational training, among many others, the IAB Establishment Panel is the central basis for the analysis of labour demand in Germany. See Fischer et al. (2008) for a description, or visit the FDZ's homepage: http://fdz.iab.de/en.aspx for further information on data access, methodology and working tools.

The IAB Establishment Panel is conceived as a longitudinal survey. On the one hand, a majority of establishments takes part in the survey in subsequent years. On the other hand, large parts of the questionnaire remain constant over time. This longitudinal nature enables not only investigation of developments over time by contrasting cross sections but also within establishments.

However, the IAB Establishment Panel is not prepared as a panel data set but the single cross sections are available. Harmonizing cross-sectional variables can be tedious and is not always unproblematic. Variable names are wave-specific. They combine the wave, indicated by lower case characters, and the number of the question in the questionnaire. a07 for example indicates the seventh question of the first wave's questionnaire in 1993. Moreover, many related variables differ in content across waves as changes in wording of questions, answer categories, coding of single values, measurement unit and reference period, among others, occur.

To assist the users with combining the waves of the IAB Establishment Panel the FDZ offers the Stata do-file *BP_panelgen_93YY_v1_en*.[1] This do-file creates a panel data set of all years from 1993 to the most recent year and for all variables of the IAB Establishment Panel. The resulting panel variables are consistently named, coded and labelled across years. As far as possible, the original information is maintained and no artificial information is added, e.g. by forward/backward extrapolation.

While the goal of the data preparation is to provide a standardised basis for panel analyses with the IAB Establishment Panel, caution has to be taken as some panel variables might still differ in content across waves due to the problems mentioned earlier. It is the responsibility of the user to verify the generated data is suited for his or her research purposes. Particularly, the user should always make sure that the underlying questions and/or re-coding are sufficiently consistent across waves given the research question at hand. Therefore, we strongly recommend using the questionnaires and the Excel-file 'Variablenliste_en' (List of variables), both available from the FDZ's homepage, in addition to the work with the do-file.

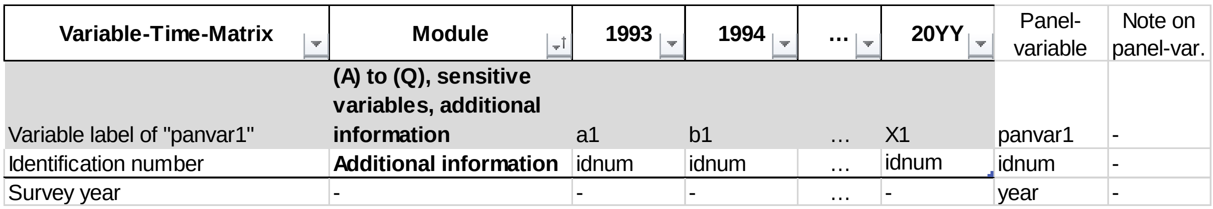
[1] The IAB Establishment Panel is also part of the different models of the Linked-Employer-Employee-Data of the IAB (LIAB), see Klosterhuber et al. (2016) on the cross-sectional model and Heining et al. (2016) on the longitudinal model, as well as of WeLL and LPP. Users of these data can also apply the do-file in exactly the same way as described in this Methodenreport.

This Methodenreport is structured as follows: The next section describes changes in the data structure when combining the waves to a panel as well as basic steps of harmonization. Sections 3 and 4 outline the structure of the do-file and provide operating instructions. The final section contains some remarks on the preparation of specific panel variables.

## 2 Basics of data preparation

To provide an overview of the variables collected in at least two waves of the IAB Establishment Panel with sufficient similarity, the FDZ offers a variable-time-matrix.[2] Table 1 presents an illustrating example of this matrix.

**Table 1 - Illustration of variable-time-matrix**



| **Variable-Time-Matrix** | **Module** | **1993** | **1994** | **…** | **20YY** | Panel-variable | Note on panel-var. |
|---|---|---|---|---|---|---|---|
| Variable label of "panvar1" | **(A) to (Q), sensitive variables, additional information** | a1 | b1 | … | X1 | panvar1 | - |
| Identification number | **Additional information** | idnum | idnum | … | idnum | idnum | - |
| Survey year | - | - | - | … | - | year | - |

Each row of the variable-time-matrix contains the original variable names from the cross sections, e.g. a1 to X1 as a hypothetical example. The first column holds a rough description of the content of these variables. This description will also serve as the variable label of the newly created panel variable. The second column indicates the thematic module each set of variables is assigned to. Organizational variables, like idnum, are part of the module ‘Additional information’, for instance.[3] The variable year, in contrast, is not included in any cross section but is created during the panel generation. The names of the panel variables appear in the penultimate column, like panvar1 in this hypothetical example. Finally, the last column provides some notes on specific panel variables, see Section 5 for details.

The original cross-sectional data sets are organized in wide format, meaning that there is one data set per wave with one row per establishment included. The tables in the upper panel of Figure 1 provide a fictitious example of this format: The rows contain all realisations of variables belonging to a given establishment, like establishment identifier idnum and all values $wq_j$ collected in the current survey wave, with $j$ = [1; N] for the establishment, $w$ = [a; X] for the survey wave and $q$ = [1; Q] for the question (assume, for simplicity, that questions $q$ are identical across waves).

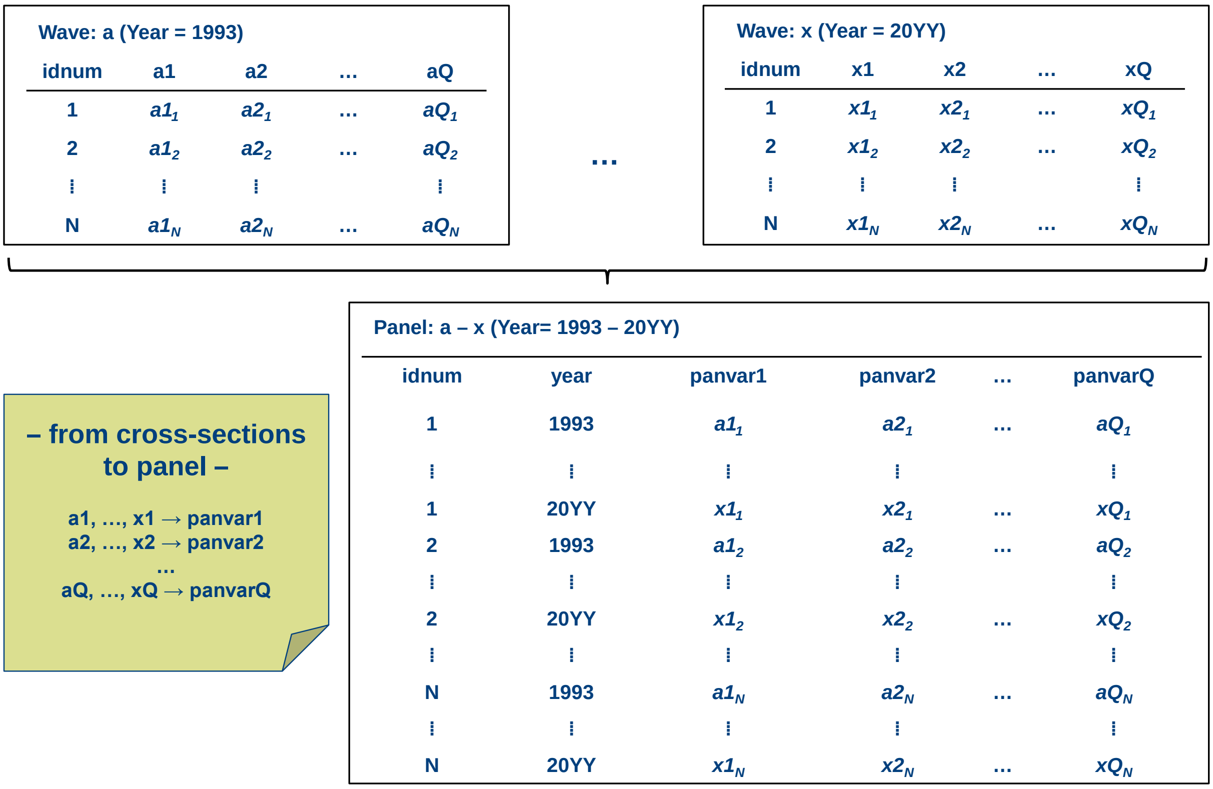
The panel preparation changes the format of the data from wide to long, as illustrated in the lower panel of Figure 1. This implies that the cross sections are merged in a way that the arising panel data set now holds one row per establishment for each year the establishment was included. At the same time, the single cross-sectional variables a*q* to X*q* are subsumed into one panel variable. In this illustrating example the panel variable is called panvar*q*. Consequently, the columns of the panel data set, next to idnum and year, contain the new panel

[2] This matrix can be found in the sheet ‘Panel variables’ of the Excel-file ‘Variablenliste_en’ (List of variables). The sheet ‘Additional variables’ lists those variables surveyed in one single wave only, usually as part of wave-specific focal topics.

[3] Section 4 describes how the user can restrict the preparation of the panel data set to subsets of modules.

variables, which subsume all related cross-sectional variables (originally the entries of the rows of the variable-time-matrix).

Figure 1 - Panel structure - from wide to long format



Wave: a (Year = 1993)

| idnum | a1 | a2 | … | aQ |
|---|---|---|---|---|
| 1 | $a1_1$ | $a2_1$ | … | $aQ_1$ |
| 2 | $a1_2$ | $a2_2$ | … | $aQ_2$ |
| ⁞ | ⁞ | ⁞ | | ⁞ |
| N | $a1_N$ | $a2_N$ | … | $aQ_N$ |

…

Wave: x (Year = 20YY)

| idnum | x1 | x2 | … | xQ |
|---|---|---|---|---|
| 1 | $x1_1$ | $x2_1$ | … | $xQ_1$ |
| 2 | $x1_2$ | $x2_2$ | … | $xQ_2$ |
| ⁞ | ⁞ | ⁞ | | ⁞ |
| N | $x1_N$ | $x2_N$ | … | $xQ_N$ |

– from cross-sections to panel –

a1, …, x1 → panvar1
a2, …, x2 → panvar2
…
aQ, …, xQ → panvarQ

Panel: a – x (Year= 1993 – 20YY)

| idnum | year | panvar1 | panvar2 | … | panvarQ |
|---|---|---|---|---|---|
| 1 | 1993 | $a1_1$ | $a2_1$ | … | $aQ_1$ |
| ⁞ | ⁞ | ⁞ | ⁞ | | ⁞ |
| 1 | 20YY | $x1_1$ | $x2_1$ | … | $xQ_1$ |
| 2 | 1993 | $a1_2$ | $a2_2$ | … | $aQ_2$ |
| ⁞ | ⁞ | ⁞ | ⁞ | | ⁞ |
| 2 | 20YY | $x1_2$ | $x2_2$ | … | $xQ_2$ |
| ⁞ | ⁞ | ⁞ | ⁞ | | ⁞ |
| N | 1993 | $a1_N$ | $a2_N$ | … | $aQ_N$ |
| ⁞ | ⁞ | ⁞ | ⁞ | | ⁞ |
| N | 20YY | $x1_N$ | $x2_N$ | … | $xQ_N$ |

During the process of panel generation the do-file takes a couple of basic steps for aligning the cross-sectional variables:

- assignment of unique names to panel variables
  (column ‘Panel-variable’ of the variable-time-matrix),

- assignment of variable labels to panel variables
  (first column of the variable-time-matrix),

- consistent labelling of categories common to all relevant waves, i.e. consistent value labelling,

- adding of a category ‘no’ / ‘0’ to variables for which only the category ‘yes’ / ‘1’ was coded in the original data,

- grouping of various missing-categories into one missing-category with value ‘-9’ and label ‘n.s.’ (not specified) – if a variable does not exist the standard missing value ‘.’ is coded –, and

- various types of case-specific adjustments as described in Section 5.

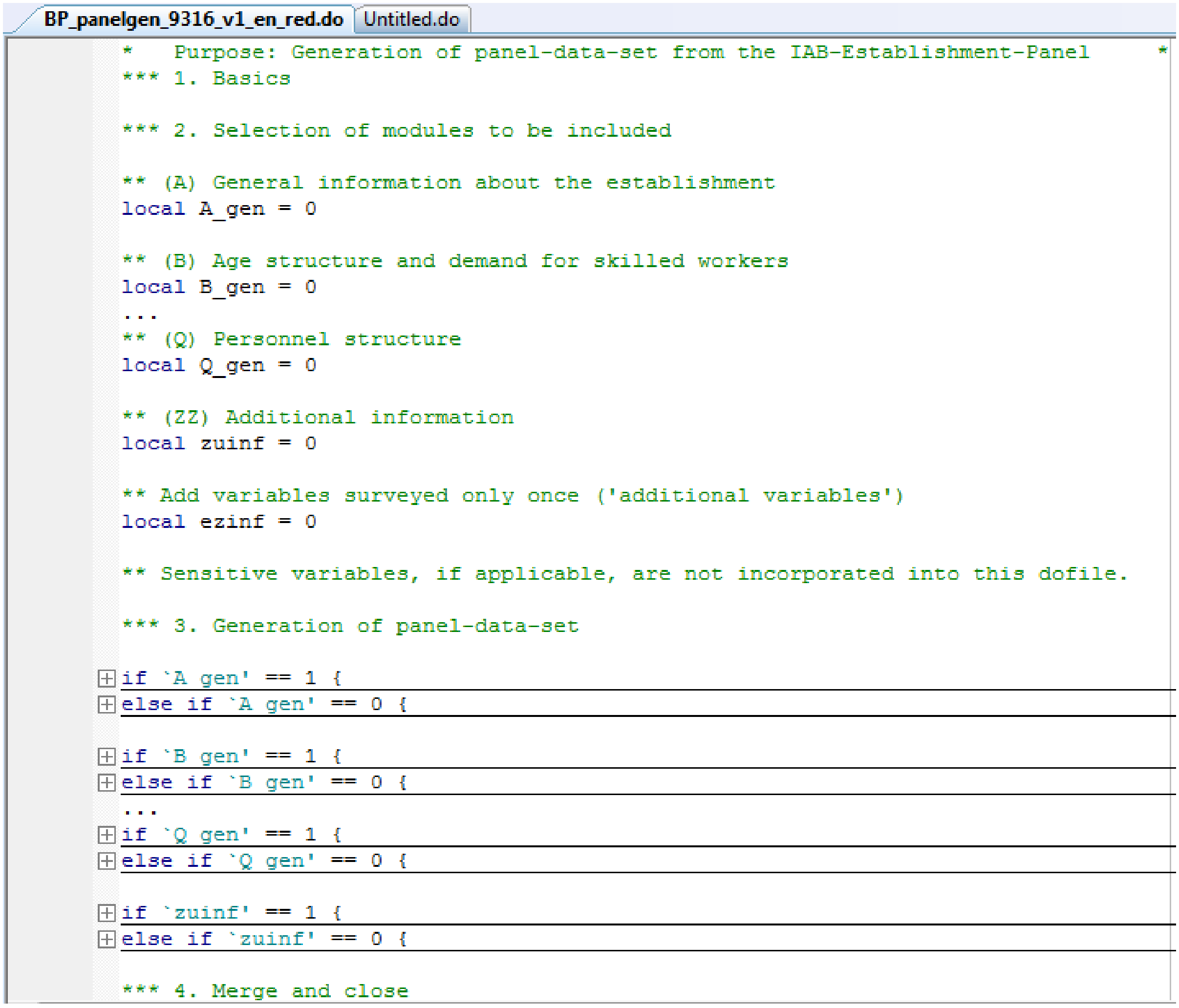
## 3 The structure of the do-file

The program code of *BP_panelgen_93YY_v1_en* is kept simple and transparent so that users can easily adjust the code if necessary. There is only one do-file with its basic structure outlined in Figure 2 and Figure 3.

Figure 2 shows that the do-file is organized in four sections. The first section configures basic settings such as Stata version, directories and naming of log-/ and data-files.

In the second section the modules (A) to (ZZ), which are to be included in the panel data set, are defined by setting the associated local macro from ‘0’ to ‘1’. By default, only variables collected in multiple waves will be included in the final panel data set. However, it is possible to add ‘Additional variables’, i.e. variables surveyed only once, to the selected modules by setting the local macro ezinf from ‘0’ to ‘1’.

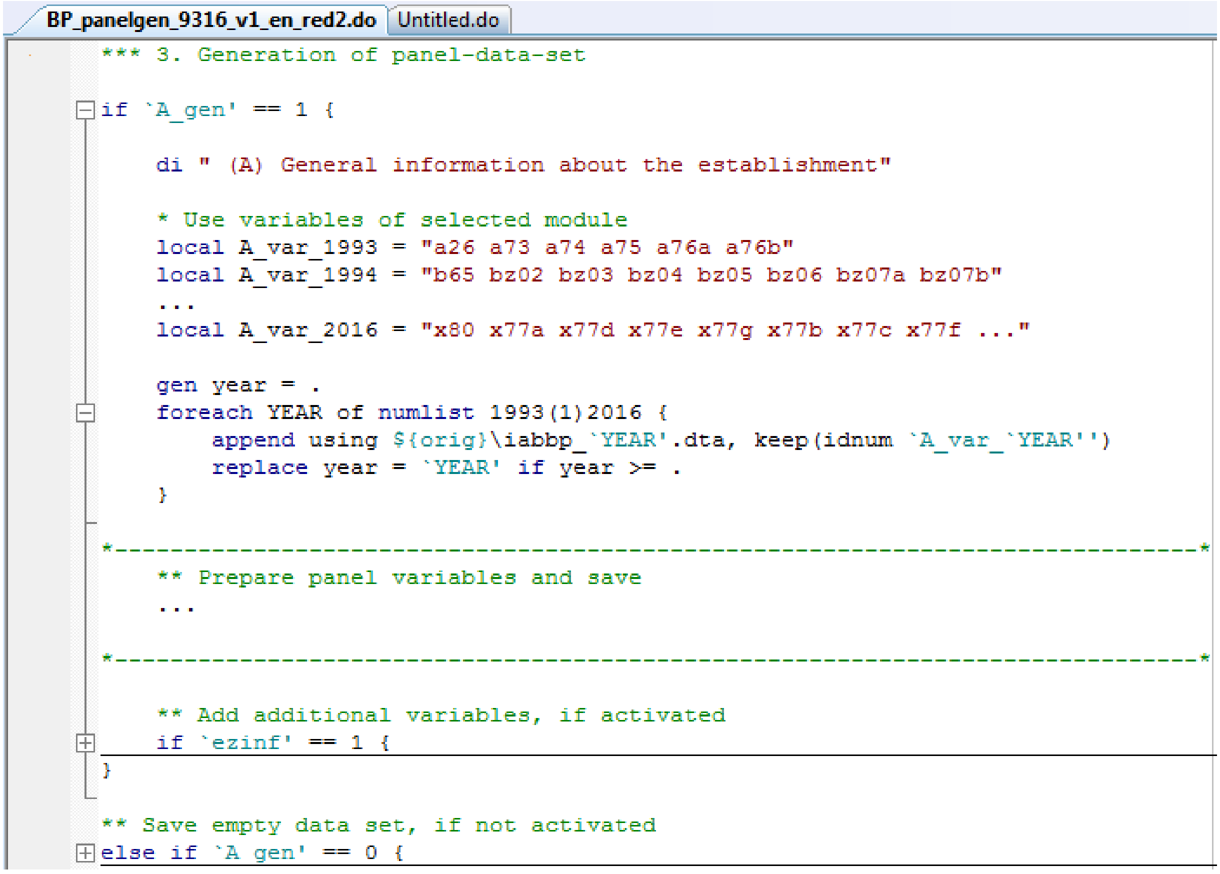
**Figure 2 - Illustration of do-file structure (level 1)**



```
*   Purpose: Generation of panel-data-set from the IAB-Establishment-Panel     *
*** 1. Basics

*** 2. Selection of modules to be included

** (A) General information about the establishment
local A_gen = 0

** (B) Age structure and demand for skilled workers
local B_gen = 0
...
** (Q) Personnel structure
local Q_gen = 0

** (ZZ) Additional information
local zuinf = 0

** Add variables surveyed only once ('additional variables')
local ezinf = 0

** Sensitive variables, if applicable, are not incorporated into this dofile.

*** 3. Generation of panel-data-set

if `A_gen' == 1 {
else if `A_gen' == 0 {

if `B_gen' == 1 {
else if `B_gen' == 0 {
...
if `Q_gen' == 1 {
else if `Q_gen' == 0 {

if `zuinf' == 1 {
else if `zuinf' == 0 {

*** 4. Merge and close
```

In a third section, the program prepares temporary subsets for each module. An empty data set is saved if the module is not activated. If the module is activated the program prepares the data as illustrated in Figure 3: First, local macros for each wave are used to load all related variables of the module, i.e. variables shown in the rows of the variable-time-matrix. Second, the variables are appended and the time-variable year is created. As variable names are unique in the cross sections, the appended subset now contains one column per cross-sectional variable holding the survey information if the year equals the wave and missing '.' otherwise. Third, the cross-sectional information is transferred into one column with the help of the egen-function rowmin(). This new column is kept as the new panel variable and prepared as described at the end of the previous and in the final sections.

**Figure 3 - Illustration of do-file structure (level 2)**



```
*** 3. Generation of panel-data-set

if `A_gen' == 1 {

    di " (A) General information about the establishment"

    * Use variables of selected module
    local A_var_1993 = "a26 a73 a74 a75 a76a a76b"
    local A_var_1994 = "b65 bz02 bz03 bz04 bz05 bz06 bz07a bz07b"
    ...
    local A_var_2016 = "x80 x77a x77d x77e x77g x77b x77c x77f ..."

    gen year = .
    foreach YEAR of numlist 1993(1)2016 {
        append using ${orig}\iabbp_`YEAR'.dta, keep(idnum `A_var_`YEAR'')
        replace year = `YEAR' if year >= .
    }

*-------------------------------------------------------------------------------*
    ** Prepare panel variables and save
    ...

*-------------------------------------------------------------------------------*

    ** Add additional variables, if activated
    if `ezinf' == 1 {
}

** Save empty data set, if not activated
else if `A_gen' == 0 {
```

The module-specific subsets are merged by idnum and year to one panel data set in the fourth step. The name of the panel data set is identical to the name of the do-file used for generating it, i.e. *BP_panelgen_93YY_v1_en.*

The attached Stata do-file has been tested with IABBP_9316_v1 using Stata 14.1. The FDZ does not guarantee that the specifications chosen in the provided codes can be applied to all research interests. We strongly advise users to check if the specifications can be transferred to their research project before adopting the routines. Furthermore, we cannot completely rule out mistakes in the codes and the variable-time-matrix. If you find a mistake, please send an e-mail to the FDZ-box (iab.fdz@iab.de). For a general introduction in data analysis with Stata we recommend Kohler und Kreuter (2012a und 2012b).

# 4 How to work with the do-file

Users should consider the following points when working with the do-file *BP_panelgen_93YY_v1_en*:

- Download of the do-file

  Users can download the do-file from the FDZ homepage (see also the Link in the Appendix). The do-file is also available on-site during guest visits.

- Data requirements

  The routine requires all waves of the IAB Establishment Panel indicated by the name of the do-file, e.g. waves 1993 to 2016 in the case of *BP_panelgen_9316_v1_en*. Currently, there is no option to select specific years. Waves not needed for the research project can be dropped by the user after panel data preparation.

- Use within FDZ environment

  When working on-site or via remote-execution with JoSuA no directories have to be specified. The program automatically assumes a directory structure as set up in the FDZ-Gästenetz and JoSuA, respectively, and defines the associated macros.

  To execute the program for panel generation the following line

  ```
  do “$prog\BP_panelgen_93YY_v1.do"
  ```

  (with the year of the version used inserted for ‘YY’) has to be written in the *master.do*. In case of remote execution both *master.do* and *BP_panelgen_93YY_v1_en.do* have to be uploaded to JoSuA.

- Working with test data

  The program also runs with the test data of the IAB Establishment Panel. Before execution users have to create directories structured in the same way as in the FDZ environment. The following directories are used by the panel do-file:

  - Main directory – the user’s project repository (e.g. ‘C:\fdzXXXX’)
    - prog – where all do-files are located
    - orig – where the original data files / test data are stored
    - data – where the created panel data set is saved to
    - log – where the corresponding log-file is saved to

  Additionally, the associated macros have to be defined. Conveniently, the do-file already contains the required code under the point ‘** Directories’. By default, this code is commented out. Once activated only the global macro main has to be defined by inserting the path of the main directory. Afterwards, the do-file operates in the same way as described under ‘Use within FDZ environment’.

- Choice of modules to be included in the panel data set

  The modules to be included in the panel data set have to be defined by the user in section 2 of the do-file. This is done by setting the associated local macros from ‘0’ to ‘1’. Due to memory constraints the number of selected modules has to be kept as small as possible. By default, only variables collected in at least two waves are considered.

- Adding ‘Additional variables’ to selected modules

  To add variables surveyed only once to the selected modules the local macro ezinf has to be set from ‘0’ to ‘1’.

- Adding sensitive variables

  Sensitive variables are not considered in the data preparation. They have to be added by the users themselves, e.g. by merging one-to-one on idnum and year. However, it has to be noted that the variable year is not included in the respective source files but hast to be generated during the preparation of the merge.

- Generated files

  The program saves a log-file documenting the choice of modules to the directory ‘log’ and a .dta-file of the panel data set to the directory ‘data’. By default, both log-file and data-file are named in exactly the same way as the do-file.

## 5 Notes on prepared panel variables

The last column of the variable-time-matrix contains some notes on specific panel variables either related to data preparation or time-consistency of the underlying cross-sectional information. However, if there is no note this does not necessarily mean that the variable is problem-free. Examples for such notes include:

- *Integrated* – Missing values were replaced in the panel variable if stored in separate variables.

- *Information aggregated in some years* – Some categories have been grouped in at least one year while generating the panel variable, implying that the cross-sectional information is more detailed in at least one year than in the panel.

- *Generated using information from multiple rows* – Variables from different rows of the variable-time-matrix have been combined when forming the panel variable.

- *Classification changes over time* – More than one panel variable have been created as a result of changes in classification being too sever in order to be harmonized.

- *Until 2001 (incl.) division by 1.95583* – Values in DM have been converted to EUR.

- *Question deviates: YYYY* – The question in the given year deviates from related questions in other years.

- *Question changes: from YYYY* – The wording or content of related questions changes from the given year onwards.

- *Rows marked with an asterisk* – Descriptives of the panel variable show noticeable deviations in at least one year. However, these deviations do not necessarily imply inconsistencies.

Finally, the last digit of the name of each panel variable indicates whether the variable is

- **_b* – binary; typically ‘yes/no’,

- **_d* – discrete or

- **_c* – continuous.

## References

Alda, Holger; Dundler, Agnes; Müller, Dana; Spengler, Anja (2006): *Development of a panel dataset from the IAB Establishment Panel cross-sectional data.* FDZ-Datenreport, 02/2006 (en)

Fischer, Gabriele; Janik, Florian; Müller, Dana; Schmucker, Alexandra (2008): *The IAB Establishment Panel – from Sample to Survey to Projection.* FDZ-Methodenreport, 01/2008 (en)

Heining, Jörg; Klosterhuber, Wolfram; Lehnert, Patrick; Seth, Stefan (2016): *Linked Employer-Employee Data from the IAB: LIAB Longitudinal Model 1993-2014 (LIAB LM 9314).* FDZ-Datenreport, 10/2016 (en)

Klosterhuber, Wolfram; Lehnert, Patrick; Seth, Stefan (2016): *Linked Employer-Employee Data from the IAB: LIAB Cross-sectional Model 2 1993 – 2014 (LIAB QM2 9314).* FDZ-Datenreport, 05/2016 (en)

Kohler, Ulrich; Kreuter, Frauke (2012a): *Datenanalyse mit Stata * allgemeine Konzepte der Datenanalyse und ihre praktische Anwendung.* 4. Auflage. München: Oldenbourg Wissenschaftsverlag

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

## Appendix

Download of the Stata do-file:

- http://doku.iab.de/fdz/iabb/BP_panelgen_en.zip

# Imprint

FDZ-Methodenreport 12/2017 (EN)

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

**Editorial staff**
Dana Müller, Dagmar Theune

**Technical production**
Dagmar Theune

**Download**
http://doku.iab.de/fdz/reporte/2017/MR_12-17_EN.pdf

**Internet**
http://fdz.iab.de/

**Corresponding author:**
Matthias Umkehrer
Institute for Employment Research (IAB)
Research Data Centre (FDZ)
Regensburger Str. 104
D-90478 Nürnberg
Phone: +49-911-179-6211
Email: Matthias.Umkehrer@iab.de