



Guidelines for On-Site Use and Remote Data Access at the FDZ of the BA at the IAB

To be able to process your jobs efficiently standardisation is needed. Thereby we also ensure that every job is processed promptly. For some users our requirements might mean that they have to depart from their accustomed way of programming.

1 General remarks

Files for on-site use and remote data access shall always be sent to the FDZ mailbox iab.fdz@iab.de. The FDZ provides analysis with the software packages Stata and SPSS as standard. If you need other statistical software packages for your analysis please contact the FDZ.

Data at the aggregate level (e.g. unemployment rates by districts) may be added to the data. The data s shall be transmitted to the FDZ with a description of the variables, the aggregation level and the data source.

Stata ado files you want to use have to be transmitted as well. How to include the ado files in your programmes is described in the example on page five.

E-mails are sometimes retained by firewalls or SPAM filters without us being informed. Therefore we acknowledge the receipt of every job submission by e-mail. If you do not get this confirmation within two business days, please tell us.

Before we return your results we do a disclosure limitation review. Its main purpose is the prevention of the risk of identification of individuals or establishments. We first check if your programs and results are documented correctly (see Section 5). Then we look at the results and might delete parts we regard as insecure. The researcher is obligated to refrain from recalculating analysis results in order to recreate deleted values or re-identify individuals or establishments. Should the researcher not comply with these regulations, he will be excluded from any further use of our data for two years. Other research data centres will be notified of any non-compliance with the aforementioned obligations.

We try to return your results as soon as possible, at the latest after 5 business days. Should it take longer than 5 working days we will inform you.

2 On-site use

We recommend that you prepare your programs with our dummy data ahead of your stay. If no dummy data exist and the data you use are available as a scientific use file also, it is advisable to use that. Files you need for on-site use shall be sent to us at least three days ahead of your stay. If they are sent later we cannot ensure that they have been copied to the computer when you arrive so that you might not be able to start using them right away.

The goal of the guest stay is to prepare programs while working with the real data. We will restart these programs at the end of your stay and you will get the results these programs produce. You will not get the results you produce during your research stay.

Please write your programs according to our specifications in Section 5. In particular, please create a master file that documents the sequence of your programs.

3 Remote data access

We do not edit your programs. Hence they must be submitted in a completely workable status. We expect that the programs are tested with our test data as far as possible. Submitted files overwrite existing files if they have the same filename.

To process your jobs we need the following information:

Project number – memory required – expected runtime – expiration of contract – misc

Based on the information on memory required and expected runtime we can enhance the capacity utilisation of our computer. The purpose of asking for the expiration date of the contract is that you and us are reminded on time to extend expiring contracts. "Misc" stands for anything that we should be aware of before we start the programs or before we do the disclosure limitation review (e.g., if MATLAB or TSP are to be used for the analysis). To ensure that your entries are machine-readable, please follow this pattern below exactly.

fdz### - # {mb, gb} - # {m, h, d, w} - yyyy-mm-dd - [Misc]

```
# means numbers
mb = megabyte
gb = gigabyte
m = minutes
h = hours
d = days
w = weeks
yyyy-mm-dd = date in format year-month-day
```

Example Stata: **fdz041 - 3 gb - 14 h - 2011-01-31**

We need „memory required“ only if you use Stata. If you use other software, please insert the number zero.

Example other programs: `fdz041 - 0 gb - 14 h - 2011-01-31`

We ask for your understanding that we return jobs unprocessed if relevant information is missing in the subject line.

4 Structure of the project folders

Each project has designated project folders, which remain the same throughout the project. If your project requires more than one dataset, e.g. BAP and IABS, you will get a separate folder for each dataset.

Each project folder contains the following subfolders:

orig: Here we put in all the original data. Besides the requested data this may include aggregated data that you either brought with you or we put in there. In this folder you may read files but not write.

data: Here you save the data. Please be parsimonious with disk space, do not use more than 20 gigabytes.

prog: Here we put in your programs (do files, ado files, sps files, etc).

log: This folder is for all results, including graphs. Please save Stata results as log files (smcl is not admissible as it hampers our disclosure limitation review).

doc: This folder contains all other files, e.g. Word and Excel files.

It is not permitted to create other folders or subdirectories. There are backups for all project folders according to the standards of the BA.

Please use relative paths only. This way you do not have to change all the paths if the project folder is moved to another server. Use your paths in a way that by double-clicking on your master file (see section 5.1) all programs you want to run start automatically and that all files created are saved in the right folders. We recommend that at the beginning of your master file you move from your *prog* folder one level higher (`cd ..`). After that you can address the subfolders *prog*, *data*, and *log* as described in the example in Section 5.2. The master file is kept in the *prog* folder. If you are not familiar with using relative paths, please talk to us, we will help.

Disk space per project is limited to 20 GB. You can archive up to 30 GB. These files will be stored up to 10 years after the end of your project. You can access archived files for further calculations, but we need up to two weeks to copy them back on our server.

5 Organisation of programs

The purpose of the following rules is to minimize time and effort and to ensure the traceability of the analyses.

5.1 Documentation of program sequence via master file

All programs that you use in your project have to be started by a master file. One purpose of using master files is that we only have to start one program per job even if many programs are supposed to run. The other purpose is that we can trace the steps of your analysis better if we can see the sequence of your programs listed in the master file. For the latter purpose it is also important that you write comments behind every program started from within the master file, which tell us what this program does.

The master file shall include all relevant program calls of the project, so that it is possible at any time to execute all program steps from reading the original data to the creation of final results by one run of the master file. Programs already executed and which are not run again shall be commented out. The name of the master file must be "master.do", this helps us to start your jobs. At the end of your research stay will start your master file which in turn starts the programs you created. Only the hereby resulting files will be sent to you.

Do not use a master file if you work with **SPSS**. This is possible in principle but does not help us because it is hard to control which results are generated by which command. Please create a simple text file instead which lists the sequence of your SPSS programs. We start your programs according to this list. Please send us this list whenever you change it.

This way the master file guarantees full reproducibility of results. Hereby on the one hand an important requirement of the German terms of "good scientific practice" is fulfilled and on the other hand only the original data and the programs have to be archived at the end of the project.

Example for the contents the file master.do in Stata:

```
* Master file of the project "project name"

clear
set more off           // avoid unintentional stops
set linesize 255     // avoid line-breaks
cd ..                 // change out of the prog folder
adopath ++ prog       // tell Stata where ados are
which command         //can be used to make sure that the desired
                       //version of the ado file is used

*do prog/1_PersDat.do // preparation of person data
*do prog/2_BetrDat.do // preparation of establishment data
do prog/3_DatXY.do    // merge ... aggregation to industry level
do prog/4a_AnalyseA.do // regression analysis ...
do prog/4b_AnalyseB.do // matching of ...

capture log close     // create a list of all files in log file
log using log/Dateiliste.log, replace
dir prog\*
dir log\*
dir data\*
dir orig\*
dir doc\*
log close
```

5.2 Sequence of programs

To improve the tracability of the analysis we suggest this sequence of program steps:

- data preparation
- descriptives
- analyses

There may be more than one file on each topic. The last part “analyses” should only contain few tables as a rule they should be placed in the descriptives part.

5.3 Filenames

Stata users must call the master file master.do Include numbers as prefixes of your program filenames in a way that the sequence of your programs becomes visible by sorting the files in the file manager by filename (e.g. see filenames in the master file example above). Assign result files, data files and graphs the same name as the program that creates these results or data. The example program below is called "1_PersDat.do". The created result, the data and the graph files are called "1_PersDat.log“, "1_PersDat.dta“ and "1_PersDat.gph“ accordingly (suffixes do, log, gph and dta refer to Stata). If a program creates more than one result, graph

or data file, their filenames shall differ by suffix only. The output from SPSS is saved by the FDZ with the filename that corresponds to the program.

Example for the contents of a Stata program file for data preparation:

```
* Program file of the project "project name" - 1_PersDat.do
* Contents: preparation of person data

* Level of aggregation: Industries by German federal state
* aggregation in file 1_PersDat.do
* at least 30 establishments per data line
* the number of persons per line is in Faelle

clear
set mem 600m
capture log close
log using log/1_PersDat.log, replace

use orig/originalname, clear

* Generation of variable ...
bysort id: egen varname2 = modus(varname1), max

* Aggregation to the level of industries by German federal state
collapse (count) Faelle = varname2 (mean) varname2, by(branche bula)

* descriptives
summarize varlist

save data/1_PersDat.dta, replace
log close
```

5.4 Documentation within the program files

Besides the documentation of the sequence of program steps in the master file, each program file shall contain headers and remarks that enable the reader to trace the steps taken. We expect:

- headers such as "preparation of variable X" or "aggregation to the industry level"
- a description of the contents/function of loops
- the use of variable labels for all important new created variables
- the use of self-documenting variable names.

If the micro data are aggregated (e.g. to the level of establishments, regions, industries, occupations, etc.) and the further analyses are based on the aggregated data this shall be indicated clearly in the program file. We also ask you to create a variable that contains the number of establishments or persons per data line (e.g. by `bysort branche: gen N = _N`). We

cannot do the disclosure limitation review unless you provide the number of cases. All result files based on aggregated data shall contain the following information in the head of the file:

- the level of aggregation
- in which program file the aggregation was done (this step shall also be marked in the master file)
- how many persons or establishments are at least represented in a data line
- the name of the variable that includes the number of persons or establishments per data line

If the indication of the aggregation is missing in the head of the result file we are likely to delete more numbers than necessary when we do the disclosure limitation review.

5.5 Documentation of the number of cases

One measure to avoid the identification of real persons or establishments is that your results have to pass a disclosure limitation review. In this process all values that are based on less than 20 observations (individuals or establishments) will be replaced by "/". To prevent easy calculation of these values it is often necessary to delete or round other values as well. Whatever statistics you calculate, always display the number of observations each value is based on. If we do not see this number of observations we have to delete the results. If a variable takes only two values (e.g., dummy variables) it is easy to use the mean to calculate the frequencies of the two categories. Therefore we ask you to display minimum, maximum and standard deviation whenever you calculate a mean. With this information we can distinguish variables with only two values from other variables.

If you use sample weights and the command does not display the number of observations you have to do the calculation again without weights. Weighted and not weighted results have to be placed next to each other. If you produce a lot of weighted numbers it is advisable to run two programs, one with weights and one without, everything else equal.

5.6 Permitted file formats

If you use Stata save your results as ASCII text files. Formats like xls, doc, or rtf hamper our disclosure limitation review. To receive results which have been saved in separate files beyond the log files you have to integrate these results in the log file. If you use SPSS in most cases we export the results to Excel to do the disclosure limitation review. In these cases you will get the Excel file.

Stata 10 graphs can sometimes be transformed back into data. To avoid this possibility we will only send you Stata graphs that are generated with the "asis" option:

```
graph ... , ... saving(filename, replace asis) OR  
graph save filename, replace asis
```

The formats wmf, eps and png are allowed also. Please generate graphs only if it is not possible to create them on the basis of tables with e.g. Excel by your own. The number of generated graphs should be as low as possible as the effort of disclosure controls on graphs is very high. Please provide evidence on how many observations every depicted value is based on. This is possible by displaying the observations either right in the graphs or in tables placed directly before or after the generation of the graphs. If you want your results in the form of an aggregated data file talk to us first. We may recommend that you create the file on-site. As a general rule the aggregated data file has to contain an additional variable for each aggregated variable which contains the numbers of valid observations the aggregated variable is based on. The assigned variable names should be meaningful. You may get only one data file per project if the number of observations per cell is low.

6 Publication of results

In publications that contain analyses of FDZ data, the data and documentations have to be cited according to the guidelines on the FDZ website (http://doku.iab.de/fdz/access/Zitierweisen_e.pdf). Please send us also one voucher copy.