09 November 2023

CIRRAU Processing Details

Allan Timmermann[1,2], Ane Marie Closter[1,2], Marie Kempf Frydendahl[1,2]

[1] Centre for Integrated Register-based Research (CIRRAU), Aarhus University, Aarhus, Denmark

[2] National Centre for Register-Based Research, Aarhus University. Fuglesangs Allé 4, 8210 Aarhus V, Denmark

CONTENT

This document is an appendix to "Guidelines and Criteria for CIRRAU and ECONAU servers at Statistics Denmark". The document contains processing details not described in the before mentioned guideline and provides guidelines to best practice for responsible use of these common computer resources.

1. The servers

CIRRAU has several hosted servers at Statistics Denmark, for example: Srvfsencrr12, Srvfsencrr13, Srvfsencrr16.

Each server contains specific research projects, which are named using six digits e.g., 703373, and data are placed in a project according to the permissions of the project.

Access to the servers:

Please see the document "Forbindelse til forskerservere i Danmarks Statistik" at https://www.dst.dk/da/TilSalg/Forskningsservice/Vejledninger for a guide on how to access the remote desktop at Statistics Denmark once you have received your id and password. The remote desktop can be accessed at remote.dst.dk. Here you can also change your password or reset your password if you forgot it. In the feature DDV (Danmarks Datavindue) you can get an overview of the projects you have access to, possibility to browse through available registers and variables and a learning facility that will help you learn more about working with microdata.

Software and configuration:

The three statistical software applications: R, SAS and STATA are installed on all servers, as well as standard applications such as Word, Excel etc. The software application R is updated approximately once a year with all packages from CRAN (and BIOCONDUCTER on some servers). Additional specialized software for example python is available on some servers. The configuration can vary slightly between servers. However, data files are located on either the D:\, E:\, G:\, X:\ or Y:\ drive, which are common disk areas containing the folders: Rawdata and Workdata as well as the personal folder: Kode on some servers.

The Rawdata folder:

Example path: D:\Data\Rawdata\<project number>
Read only, rawdata files are stored here. Users cannot make any changes to data or write new files. Your syntax/code should read data from this location.

The Workdata folder:

Example path: D:\Data\Workdata\<project number>
The Workdata folder is the common workspace on the project. All researchers with access to the project can read and write files. Save your syntax/code, derived datasets and results here. When you get access to a project, the first thing you should do is to create a new folder in the workdata folder for the purpose above. Give the folder a name that corresponds to your researcher ID. Example: your researcher ID is ABCD and you just got access to the project 703301, you create the folder "ABCD" with the following file path: G:\Workdata\703301\ABCD. On the ECONAU servers the folder is automatically generated and named with DST ID and last 4 or 6 digits of the project number.

The Kode folder

Example path: D:\Data\Kode\<your DST ID + last 4/6 digits of the project number>
Available on CIRRAU servers only. For example, D:\Data\Kode\ABCD3373
This is a personal folder where you can save syntax/code that only you can access. Only code and short text documents are allowed to be saved in the Kode folder. Any other types of files, including data files, are not allowed to be saved in the Kode folder. For this purpose, use the Workdata folder instead (see above).

C:\ and F:\ drive

The C:\ drive is a system drive that contains the operating system and installed software. The F:\ drive contains a disk swap area for SAS and STATA temporary work files. It is important that you DO NOT under any circumstances save any files to the C:\ or F:\ drive. There is no back-up of files at C:\ and F:\ and, Statistics Denmark occasionally, deletes files from these drives.

General information about the servers

CIRRAU servers are restarted on the first of each month between 00.00 and 08.00. Please, refrain from running jobs at this time; otherwise, your work will probably be incomplete or lost.

Information specifically regarding CIRRAU server 8 and 13

The servers Srvfsencrr8 and srvfsencrr13 use the same disk area but have separate CPU and memory. Projects located on srvfsencrr13 can login to srvfsencrr8. This setup enables a larger number of users to run analyses at the same time. Server srvfsencrr8 will be closed in 2023.

Statistics Denmark formats

Statistics Denmark provides several SAS formats, which allow different categorizations of many variables, e.g., different levels of education and the groupings used in Statbank Denmark.
As an example, to access the formats in SAS you need to write the following syntax in your SAS editor:

```
libname fmt '\\srvfsenas3\formater\SAS formater I Danmarks Statistik\FORMATKATALOG' access=read only;
Options fmtsearch=(fmt.times_personstatistik fmt.times_bbr fmt.times_erhvervsstatistik fmt.brancher
fmt.uddannelser fmt.geokoder fmt.disco fmt.statistikbank fmt.disced fmt.sundhed);
```

Statistics Denmark has set up guidelines on how to use the formats, see shortcut "DST-Formater" (looks like a star) on the desktop. Here you can also find datasets with most formats in STATA and .TXT format. Researchers are of course welcome to modify a copy of these formats to use in R, STATA etc.

CIRRAU formats for SAS

CIRRAU also provides formats in SAS (NCRR formats), which allow various categorizations of data e.g., country of birth. The location of formats varies slightly on different servers.
As an example, for srvfsencrr13, to access the formats in SAS you need to write the following syntax in your SAS editor:

```
libname format "D:\Data\Formater\sas9_4_2017" access=read only;
Options fmtsearch=(format.cpr format.demografi format.grupper);
```

Batch submit SAS

To batch submit SAS, simply right click and choose batch submit, or locate the directory containing your SAS-program in MS-Dos prompt; Enter "sas94 <my sasfile>".

Signing out of servers:

Click "start", click the user avatar "your username" and click "sign out". Ensure that you sign out properly as soon as your jobs have finished. Otherwise, data stored in memory will use resources, impeding the response times of other users. If you have large jobs running that may require several hours to finish, you can disconnect from the server by closing the remote desktop or click "start", click "power options", click "disconnect". Make sure to login and sign out properly as soon as possible after your jobs have finished.

2. Best practice in data management

Keeping track of the research progress is the most important key to avoid unnecessary permanent datasets, and thereby avoiding unnecessary disk space use. Typically, register-based research utilizes large datasets, and several copies of these datasets will take up disk space, affecting negatively on the disk space for all users.

Keep track of your research progress & use of version control

When you prepare a dataset to perform analyses, e.g., on the relation between income and survival:

1) Save a copy of the SAS/STATA/R program/syntax performing your data management and analyzing tasks, and name this program e.g. inc_surv01.sas (or .r/.do). All file references in the program/syntax should be directly back to rawdata, thereby documenting all changes

needed to prepare data for analyses. The perfect program/syntax includes a description of the reason for each of the tasks performed.

2) Save a copy of the workdataset when it is ready to be analyzed, and name this copy inc_surv01.sas7bdat (or .rdata/.dta). By naming your workdataset the same as your program/syntax, you will indirectly keep track of the program/syntax that generated your dataset for analyses.

3) Perform the analyses by writing code within the existing program/syntax or in an additional program/syntax that contain only the analytical code. This additional program/syntax should initially load the workdataset.

4) If changes are needed, (for example after an update of rawdata) generate a new program/syntax for this task by making an edited version of the former program/syntax. Name the new file inc_surv02.sas. As before, in the new version of the program/syntax all file references should be directly back to rawdata, and never to the former version of these data (e.g., inc_surv01.sas7bdat).

5) Refrain from using version control such as final, very final, second final, definite final, final2, etc. There will never be a final version (except when the paper has been published and you no longer work on the project).

6) You are advised to create a text file called 'readme.txt' in each directory where you shortly describe the contents of files in this directory as well as planned future steps, planned changes, along with dates. This will also help you keep track of your research process.

If you follow these guidelines, you will have a clear track of the research progress, and it will be possible at any stage to implement a new update to your data. In addition, it is possible to delete all intermediate datasets, and to regenerate data from program/syntax if needed.

Unfortunately, based on our experience, many researchers fail to keep track of the research process, making it difficult to later implement changes and update data in the project. Typically, when implementing obvious changes to the analysis the researcher does not find it necessary to document the changes. However, when the publisher returns the reviewed paper 9-12 months later, many researchers have forgotten the underlying reason for the implemented changes.

Ensuring the best performance of our common computer resources

To avoid unnecessary disk space use **never keep more than one copy of a dataset**. Many new users make a new permanent copy of a dataset for each minor change. Worst case scenario is that you get access to data1, and based on this you make a new variable called var2 and save this dataset as data2. Then you decide to make another variable called var3 based on the modified dataset data2 and save the new data as data3. If you end up making 10 new variables, you will have 10 nearly identical permanent copies of the same dataset. After a few years, you may request an update to data1; now you will have to repeat all these 10 steps to implement the update, while also saving additional 10 new permanent copies of the same datasets. This is worst-case scenario for data handling regarding permanent copies of datasets. Please feel free to make as many temporary datasets as needed. 'Temporary datasets' means datasets that are only stored on the server during the ongoing login session.

It is important that users:

1) **Keep only one workdataset for each project/paper.** Do not make copies for each minor change in the dataset. Instead, the user should keep the program/syntax that updates the one working dataset with changes. This will also facilitate documentation and overview in the end.

2) **Keep track of your own disk use**. A single user should not generate more than **50 GB per project.** It is easy to check disk-use by placing the mouse on your personal folder in "WORKDATA", make a right click, scroll down and click on "properties" and look at "size on disk". If the size of the data exceeds 50GB, then you are using too much disk space and you need to delete some of your workdatasets.

3) **Limit the use of virtual memory and CPU.** For example, users are encouraged NOT to run several processes at the same time and if possible, run heavy jobs at night or during the weekends. Users are restricted to use a maximum of **25GB memory** without a preceding agreement with a data manager. Please contact CIRRAU or ECONAU if you plan heavy jobs.

4) **Remember to "sign out" properly through "START"** so that temporary files are deleted. Accumulation of temporary files causes problems for disk space and memory.

5) **Remember to empty the "Recycle Bin" If you delete files.** Found on the desktop in the upper left corner. Simply right click on the recycle bin and choose empty recycle bin.

6) **Do not save files at disk C or F –** only save files in the WORKDATA area or in your personal folder "KODE".

To ensure the performance of our common server resources, CIRRAU and ECONAU management can exclude researchers that repeatedly fail to comply with these guidelines and delete their data.

3. Sending files to the servers

Files with non-personally identifiable information
To place information on the server that do not contain microdata, e.g., syntax/code and documentation, please send the file by email to a CIRRAU Datamanager, see https://cirrau.au.dk/contact for contact details.

The email should have information about server and project number, where to place the file (complete path, typically in the workdata folder), and a statement stating that the file does not contain any personal identifiable information. Please, send only syntax or documentation when it is necessary and too large to practically type in manually.

Files with microdata
Do not send microdata to Statistics Denmark without consent from the project controller responsible for managing the research project. Please contact a CIRRAU Datamanager before sending files to Statistics Denmark.
There are two ways to send data containing microdata to Statistics Denmark, e.g., data from sources outside CIRRAU (external data). Only external data that are needed for a project and described and

approved in the project description can be sent to a project. The project controller needs to be informed and have approved that data will be send to the project. Always include information about server name and project number as well as a description of the data, which variables that needs to be de-identified and whether the file only contains persons in the approved population. If, at all, possible, data should be SAS files, which speeds up the handling time and lowers costs.

1) Use recommended mail to send encrypted files to Statistics Denmark on for example a USB-stick. Address the letter to “Service desk” and “ATTENTION: Susanne Vind”.
2) Up-load files via Statistics Denmark FSE-UPLOAD. All users have access to this service in f5 dynamic webtop with project login. Guidelines for FSE-upload can be found here: https://www.dst.dk/da/TilSalg/Forskningsservice/brug-af-forskermaskiner

The latter method is preferred.