SIPP Program Documentation and Instructions[1]

Last Updated: December 8, 2015

Contents

[1]Please address questions, comments or errors to Rob Dent, robcdent@gmail.com or Laura Pilossoph, pilossoph@gmail.com. Any errors or typos within the programs are those of the authors, use at your own discretion.

1 Overview of SIPP Programs

1.1 Bash scripts

1. `sipp_scrape.bash`
   - The main script for downloading SIPP data and submitting Stata jobs. All user options (proxy, Stata version, OS, etc.) should be set at the top of this file.
   - Downloads SIPP data by panel – default is set to 1990-2008 but the user can pre-select any specific panel year. **Note** 1990-1993 must always be downloaded together and in sequential order (i.e. don't submit 1992 by itself or 1990, 1992, 1991, 1993).
   - Refer to Section 2 for the outline of how the program moves through each SIPP panel. The panels vary by what kind of datafile is available, whether or not they require topical modules or other external files, etc.
   - User options should be set in lines 10-21:
     (a) `os` – this should be `mac`, `linux` or `windows`. Certain bash commands behave different across operating systems
     (b) `proxy` – should be `on` or `off` depending on how your machine's Internet functions. If `proxy = on`, set the `http_proxy` parameter to the correct host and port.
     (c) `stata` – include how your machine calls Stata from bash (see Section 3 for more detail)
     (d) `panels` – which years of the SIPP you want to pull. **Note** that 1990-1993 panels must always be pulled at the same time and in order, i.e. populating the `panels` value with `1991 1990 1992 1993` will result in an error. The default value is to download all panels: `1990 1991 1992 1993 1996 2001 2004 2008`.
   - The program operates inside one major loop and relies primarily on the `wget` and `sed` commands for bash, submitting Stata jobs along the way.
2. `controls.bash`
   - Downloads all controls for the analysis (in this case, just PCE deflator for earnings variables).

1.2 Stata .do files

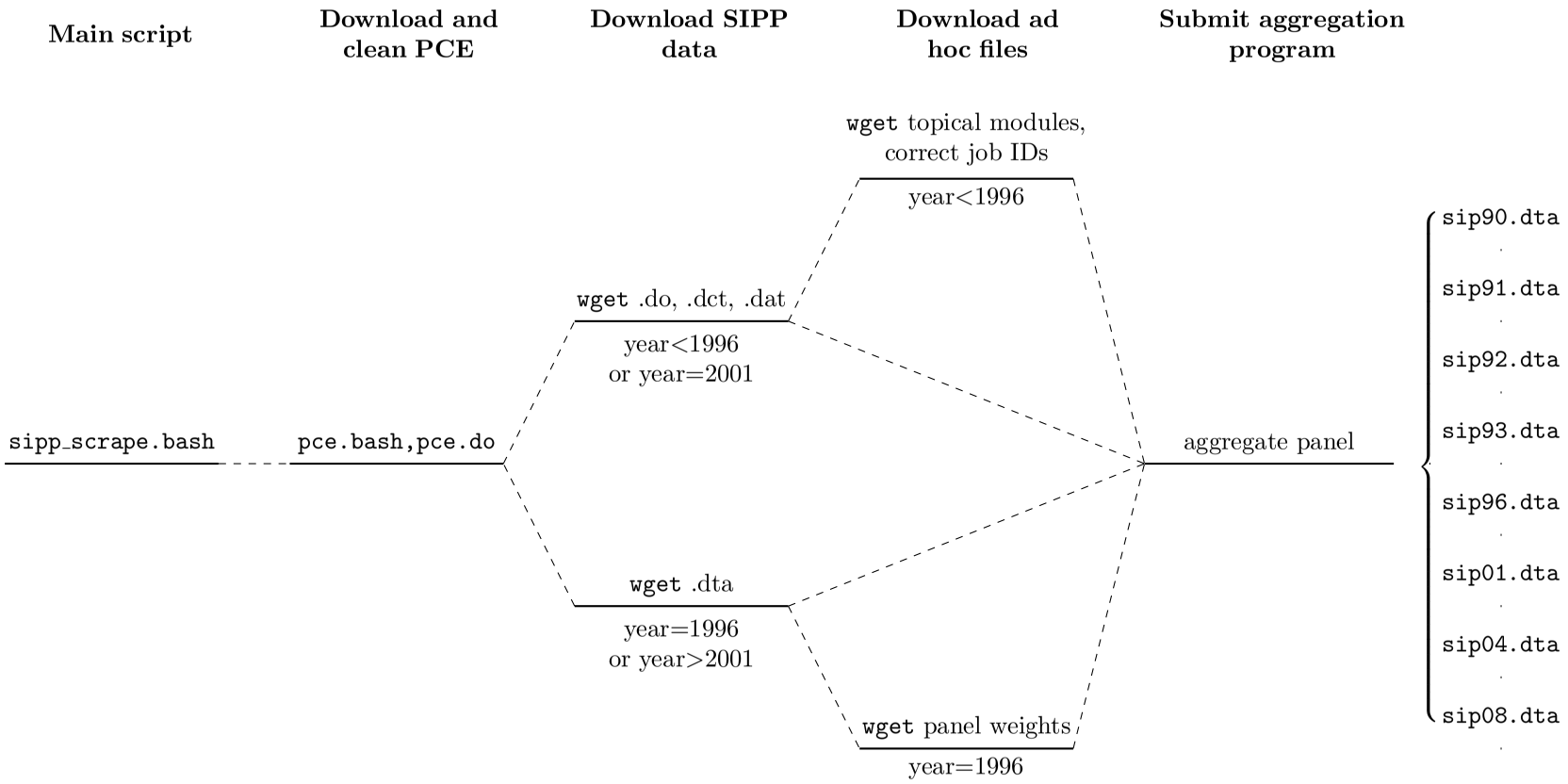
Note on Stata Files: Be sure to `ssc install carryforward` before running.

1. `dta_make.do`
   - Operates only on the 1990-93 and 2001 panels
   - Submits the .do files that call .dct/.dat files downloaded from the NBER to extract SIPP data wave by wave
   - Line 9 contains "`local year = `" – this will change (via `sed`) for each panel so that the correct waves for each panel are extracted
2. `extract_sipp_all.do`
   - This file takes the outputted .dta files (either directly downloaded or from `dta_make.do` above) and appends them together for each specific panel

- Line 6 contains the same "`local year = `" to change with each panel that the bash script is looping through
- `extract_sipp_all.do` calls `keepvars.do` to select which variables to keep and what to rename them to (more on this below)

3. `final_90_93.do`
   - Cleans the 1990-93 waves and fixes some of their panel-specific characteristics
   - For each panel in 1990-93, brings in the topical module that contains the start year and month of each respondent's *first* job (this is contained in the core waves in later panels)
   - The job IDs for 1990-93 have corrected files posted by the Census – cleans these and merges them in according to Stinson (2003)
   - Renames the 1990-93 variables to the 1996-2008 variable names (after the 1996 reorganization)
   - Extracts correct start date from the topical modules (this is only done for the 1990-93 panels, from `start_date_1990_93.do`)
4. `keepvars.do`
   - Keeps relevant variables, merges in longitudinal weights and renames variables for convenience
   - Works depending on the `${panel}` global passed to it from `extract_sipp_all.do` above
5. `controls.do`
   - Brings in and cleans data downloaded from `controls.bash`

2 Workflow for Aggregating and Cleaning SIPP Data



3 Instructions for Running Bash Script

In general, the only thing that changes across operating systems is the way

3.1 Linux

- Make sure `wget` is installed
- `os` should be set to "`linux`" in the user options section
- Populate other user options according to user-specific details

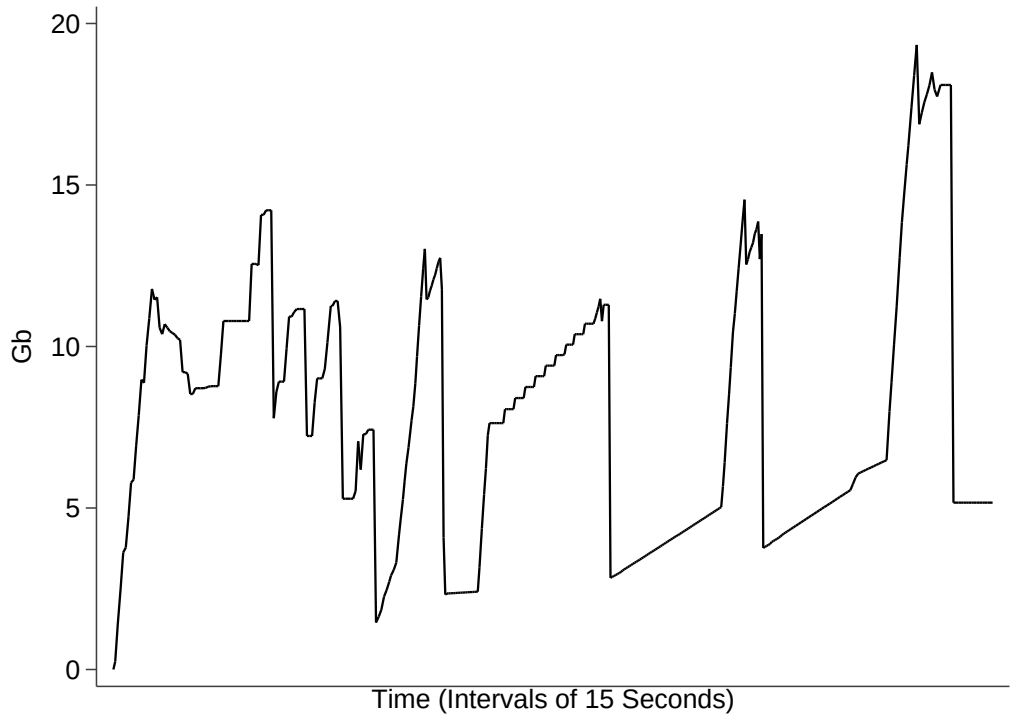
3.2 Mac

- `os` should be set to "`mac`" in the user options section
- Make sure you can submit .do files in batch mode. Open Terminal and type "`bash`", "`echo $PATH`"
- If you don't see "`/Applications/Stata/Stata{MP/SE}.app/Contents/MacOS/.`" then type "`export PATH=${PATH}:/Applications/Stata/Stata{MP/SE}.app/Contents/MacOS/.`" (be sure to replace the MP/SE with whatever Stata package you're running)
- You should now be able to run Stata in batch mode if by calling `stata -b do` (so populate the `stata` parameter with that)
- Populate other user options according to user-specific details

3.3 Windows

- `os` should be set to "`windows`" in the user options section
- Make sure `wget` is installed and specify the `stata` parameter correctly (see below)
- Calling Stata from batch mode can be done by specifying where the Stata executable file resides on your machine. For example:
  ``` ``C:\Program Files (x86)\Stata13\StataSE-64'' -e do test.do ```
  would work. The "`-e`" option here forces Stata to run in the background and print everything to a "`test.log`" file.
- Populate other user options according to user-specific details
  ```

4 Disk Space



The plot above shows a time series of the total disk space taken up for the entire `sipp_scrape` folder while downloading all panels from 1990-2008. The program periodically erases intermediate input files while running in an effort to economize on space but it still hits upwards of 20Gb before finishing up with just about 5Gb. Be sure your machine has enough disk space to handle all of the files before running the script.