**Cleaning and Matching Procurement and Bank of Spain Datasets**

**Documentation**

**Do files**

* cleaning\_boe.do: works with Bank of Spain database. Produces “names\_BoS.dta”.
* cleaning\_proc\_b0.do: works with “batches==0” cases. Reads “temp3.dta” to produce “names\_Proc\_0.dta”.
* cleaning\_proc\_b1\_a.do: works with “batches==1” cases. Manually drops useless observations. Generates “temp4.dta”.
* cleaning\_proc\_b1\_b.do: works with “batches==1” cases. Uses “temp4.dta” and cleans observations. Generates “names\_Proc\_1.dta”.
* cleaning\_proc\_b1\_c.do: works with “batches==1” cases. This version reads directly the data and does **no** manual work. Generates “names\_Proc\_1\_alt.dta”.
* append\_proc.do: appends databases created in “cleaning\_proc\_b0.do” (“names\_Proc\_0.dta”) and “cleaning\_proc\_b1\_a.do” + “cleaning\_proc\_b1\_b.do” (“names\_Proc\_1.dta”). Generates “names\_Proc.dta”.

Here I also make changes in the appended dataset, guided by errors in the matching process.

* append\_proc\_alt.do: appends databases created in “cleaning\_proc\_b0.do” (“names\_Proc\_0.dta”) and “cleaning\_proc\_b1\_c.do” (“names\_Proc\_1\_alt.dta”). This version does **no** manual work. Generates “names\_Proc\_alt.dta”.

Here I also make changes in the appended dataset, guided by errors in the matching process.

**Cleaning—temporary databases**

* temp1.dta: datafile with preliminary work on the dataset.
* temp2.dta: datafile for cases “batches==0” and fixing erroneous cases. This file contains solved cases for observations that start with regular expressions “UTE” (and similar expressions) and “UNION TEMPORAL DE EMPRESAS”.

Date: December 18, 2017

* temp3.dta: datafile for cases “batches==0” and fixing erroneous cases. Similar to temp2.dta, but does the work iteratively and includes cases with “EUROS”, “PESETAS” and numbers as parsers.

Date: January 7, 2018

* temp4.dta: datafile for cases “batches==1” and dropping observations. Cases considered (and completed) are:
  + regexm(winningfirm,"^:[ ]?")
  + regexm(winningfirm,"^,[ ]?")
  + regexm(winningfirm,"^\.[ ]?")

**Cleaning—subsample databases**

* names\_Proc\_0.dta: database corresponding to “batches==0”
* names\_Proc\_1.dta: database corresponding to “batches==1”. Uses manual work in between (“temp4.dta”).
* names\_Proc\_1\_alt.dta: database corresponding to “batches==1”. Uses manual work in between (“temp4.dta”).

**Matching**

* Do file: matching1.do
* Databases: “names\_Proc.dta” and “names\_BoS.dta”
* First attempt:

Date: December 5, 2017

Observations: 83,415 (batches==0 & manual==0)

Matches: 55,567

* Second attempt:

Date: December 22, 2017

Observations: 84,637 (batches==0 & manual==0)

Matches: 56,261

* Third attempt

Date: January 10, 2018

Observations: 241,380 (all batches & all manual)

Matches: 130,088

**Database evolution**

|  |  |
| --- | --- |
| **Bank of Spain**  Relevant do-files:   1. “cleaning\_boe.do” |  |
| * Initial sample | 2,208,697 |
| * Initial filter   + Observations after filter   + Name   + Notes | 2,208,697  “names\_BoS.dta”  This step does not drop any observations. Only re-writes data |
| * Drop duplicates   + Number of observations   + Percentage of total   + Name | 2,026,063  92%  “names\_BoS\_nodup.dta” |
| * Last update January 10, 2018 |  |

|  |  |
| --- | --- |
| **Procurement**  Relevant do-files:   1. “cleaning\_proc\_b0.do” 2. “cleaning\_proc\_b1\_a.do” 3. “cleaning\_proc\_b1\_c.do” 4. “append\_proc\_alt.do” |  |
| * Initial sample | 258,337 |
| * Subsample 1: batches==0   + Number of observations   + Percentage of total * First cleaning   + Number of observations   + Percentage of initial sample   + Notes * Second cleaning   + Number of observations   + Percentage of initial sample   + Notes * Name   Last update | 96,635  37%  95,283  99%  Dropping irrelevant information such as “NO ADJUDICADO” and related  96,749  100,1%  Trying to separate data assigned as single batch, incorrectly. Also, applies general filter (works with “SOCIEDAD ANONIMA”, “SA” and variants)  “names\_Proc\_0.dta”  January 10, 2018 |
| * Subsample 2: batches==1   + Number of observations   + Percentage of total   + Observations after dropping data with no name   + Percentage of total * Cleaning   + Number of observations   + Percentage of total   + Notes   + Name   Last update | 161,702  63%  160,546  99%  158,992  98%  Reading directly the data and avoiding manual work  “names\_proc\_1\_alt-dta”  January 10, 2018 |
| * Appending   + Number of observations   + Number of observations after filter   + Percentage of total   Last update | 255,741 (96,749 + 158,992)  241,380  93%  January 10, 2018 |

|  |  |
| --- | --- |
| **Matching**  Relevant do-files:   1. “matching1.do” |  |
| * Matched   + Number of observations (percentage of final Procurement) * Not matched   + Total   + From Procurement (percentage of final Procurement)   + From Bank of Spain (percentage of final Bank of Spain)   Last update | 130,088 (54%)  2,121,909  111,292 (46%)  2,010,617 (99%)  January 10, 2018 |