Replication manual: “Building State Capacity: Evidence from Biometric Smartcards in India”

1 Overview

1.1 Description of content

The replication archive contains a directory with 5 sub-directories and a master analysis file. This manual describes in detail how one can replicate all analysis-based tables and figures in the published version of “Building State Capacity: Evidence from Biometric Smartcards in India”. “Analysis-based” tables and figures are those which are based on underlying data rather than being illustrations (Table B.1 and Figures 1 and C.1). To begin with, each sub-directories’ content is briefly described (for full detail, see Section 4):

- *analysis_code* contains one Stata do-file or R-code file for each table. For instance, the leakage tables 3a) and 3b) are in the file “Table_3ab.do”’. Note that those do-files will not run without input from the master analysis file (see below) (the two R-code files will run independently).
- *compile_tables_figures* contains a latex-code file which uses as input all exported tables and figures (as .tex, .eps or .pdf) and compiles a pdf-document from it.
- *data* stores all data files used in the analysis. All files are in Stata 11 format (and as such forward compatible). Details on the data is provided below in Section 2.
- *output* contains all .tex, .eps and .pdf files generated by the analysis code. Every time one runs a code file, the respective output file is overwritten. Note that this folder is empty at the outset and can be populated by running the analysis code.
- *utilities* contains user-written Stata programs which were used in the analysis. Details on these are provided in Section 3

Additionally, the base directory contains the master analysis file which creates the computing environment in Stata to execute the analysis

- *master_analysis.do* This is the key-file which one should open in order to execute the analysis. The user has to set a global macro pointing to this replication directory (see next subsection). Then additional paths are automatically set (to data and output directories) and the utility functions are loaded. After this, one can run each do-file separately.

1.2 Example

Suppose one downloaded the zip-archive containing the content described above and unpacked it to “C:/Users/testUser/SmartcardsReplication”. In order to replicate the finding of - say Table 3, the key leakage table - one would open Stata and open the “master_analysis.do” file. The key step is to enter the base directory in line 17 to point to the directory where the extracted contents of the zip-file are. In this example, line 17 should become (note the absence of a trailing slash):
glo root ”C:/Users/testUser/SmartcardsReplication”
This path becomes the “root” path and separate paths for data, output, code and utilities are set (lines 23-26). Finally, one would execute lines 31-37 to load the user-written programs necessary

for some tables. At this point, one can open the "Table_3ab.do"-file and execute the do-file or take a closer look at the analysis. Alternatively, one can follow the content of "master_analysis.do" and evaluate the output of the respective command. In this example, it would be line 60.
Now suppose, one wanted to replicate Figure 2 (Official disbursement trends in NREGS). This figure is created using R[1] and hence one would open the code file Figure_2.R which is in the "analysis_code" directory. Note that a set of packages has to be installed in order for the code to run (see Section 3.2) and those are loaded on top of the R-code. The key step is to set the root directory in line 18 of the code file. Given the example above, the line would become (note the use of forward slashes and the absence of a trailing slash):

```
root <- "C:/Users/testUser/SmartcardsReplication"
```

2 Data

2.1 Description

The sub-directory "data" contains a great many data files. We provide all sections from the household survey, including all variables pertaining to questions that were not used in the analysis. For instance the files "NREGS survey section C2.dta" includes all questions from the household survey section C2 in the NREGS sample. "SSP survey general section.dta" includes all questions from the survey sections E, F, G and H as well other household characteristics. Note that the "general section" has only one observation per household.
However, some tables (mostly in the Appendix) needed a combination of a vast number of different data sources. In those cases, we provide one dataset which can be used to replicate the table. "BLELdata for sample characteristics analysis NREGS.dta" is an example of this - this file can be used to generate table C.9 and is only used for that case. Files pertaining to the Hawthorne effect analysis (E.9) are other examples of this. However, most tables are replicated by merging the respective household survey and leakage files together using jobcard or household identifiers (see next subsection).
Please note that many of the data files contain many additional variables not used in the analysis and many of them are not labelled. However, all variables used in the analysis are labelled and in-code comments give further detail.

2.2 Identifiers

The data used in the paper contains valuable private information such as pension records, NREGS disbursements, age, gender and even addresses. The contents of the surveys we conducted cannot be disclosed in a personally identifiable way either. Therefore all identifiers were removed and replaced by randomly generated ID-variables. These new ID variables however replicate the original structure (i.e. what was unique before is still unique) and can be used to merge different files, e.g., two sections of the household survey.
There are some key ID-variables which are used in almost all of the analysis:

- "_did" is the district-ID. The study was conducted in 8 districts and hence are 8 unique values. This variable is used both to merge datasets and to generate district fixed effects in the regressions

[1] The code is written to work on the most recent version of R - which is 3.3.0 (as of 5/3/2016). If one were to use an older version, the command customizing the x-axis ("scale_x_date") will throw an error. If, for some reason, one preferred to not install the most recent version, this command can be removed. The resulting plot would still show the same information but the x-axis will not correspond to the Figure in the paper.

- “_mid” is the mandal ID. Note that mandal IDs are unique within in a district only. The combination of “_did” and “_mid” is a unique mandal identifier denoted by “_umid” in the analysis and is used to cluster the standard errors.

- “_pid” is the panchayat identifier which is unique within a mandal only. The panchayat identifier is frequently used to merge Gram Panchayat baseline averages into the analysis datasets.

- “_jcid” is the jobcard identifier and identifies a unique household in the NREGS survey.

- “_householdid” exists in both the NREGS and SSP survey. Each identifies a household. In the NREGS survey, “_jcid” and “_householdid” work analogously where one derives from the official data we collected from the government (“_jcid”) and one was used for record keeping in our household survey (“_householdid”). Note that there is no overlap between “_householdid” in the NREGS and SSP sample.

- “memberid” identifies a member within the household in both the NREGS and SSP survey. At few points in the analysis, “worder_code” is used which is analogous to “memberid” but derives from the official data collection - analogous to the distinction between “_jcid” and “_householdid”

Note that in the analysis that makes use of census data, the respective ID-variables occur without the leading underscore. This denotes original ID variables (as they are used in the census and in our survey). Census data is publicly available and hence we did not create random identifiers. Finally, if a dataset contains 100s of thousands of rows of jobcard data (e.g., C.3) then the jobcard identifiers are simply replaced with a running index.

3 Code

3.1 Structure

The directory *analysis_code* contains an individual code-file for each table or figure. The name of the file corresponds to a table or figure in the paper. All do-files are written in away that once the environment is set - as described in section 1.2 - one can work with them independently. As noted above, the two R-code files are independent files and do not rely on inputs from other files (other than data which is loaded). Each code file saves a table or figure file in the “output”-directory. The latex-file in “compile_tables_figure” is then used to compile all tables and figures into a pdf-file.
Note that most of the code is not very computationally intensive and will run quite fast. However, Figures 4a-f and Figure F.1 will take some time ($\sim$ 30 and $\sim$ 5 minutes) because each sub-figure relies on 2000 bootstrap iterations. Table E.2 will take around 3 minutes as well. Overall, the “master_analysis.do” will take around 40-45 minutes to run - the prediction based on a i7-processor with 16gb RAM.

3.2 External packages used

Throughout the analysis, we used a set of external packages which can be downloaded from *SSC* in Stata and *CRAN* in R[2]. We give a brief overview over the packages we use here:

Stata:

- *distinct*: group-wise counts of distinct observations which saves values in local macros for further usage
- *unique*: equivalent to the -*distinct*- command when the latter is used with the option "jointly". The former does not save results in macros for further usage.
- *esttab*: throughout the analysis we're using the combination of -*eststo*- and -*esttab*- to create publication-style tables in .tex format
- *renvars*: an extension to the standard "rename" command allowing users to systematically rename sets of variables

R:

- *foreign* - to load data files of different formats, such as Stata 11 files
- *lubridate* - to create a x-axis scales and labels in custom formats (e.g. dates)
- *ggplot2* - a standard plotting package in R
- *scales* - a standard package to create custom scales in figures
- *reshape* - contains a useful tool ("melt") to restructure data

3.3 Self-written functions and programs

In order to automatize parts of the analysis or provide better visual representation, we defined some Stata programs which are repeatedly called. These programs are included in separate do-files in the *utilities* directory. In addition to this summary, the individual do-files provide details on the code and the purpose the program.

- *AttritionTable*: given a list of variables of interest, this program creates a table with four columns (treatment & control mean, regression-adjusted difference & p-value). Additional arguments to the program is the vector of principal component scores for the regression adjustment as well as a flag for clustered standard errors.
- *BLbalance_program*: essentially equivalent to the above but tailored to the data structure in the household survey baseline balance analysis
- *Perceptions_Table6*: this program creates the table with beneficiary opinions of Smartcards (Table 6)
- *pValueFormatting*: we wrote this program in order to be able to display p-values rounded to two significant digits and add significance stars in regression tables

[2] Packages in Stata can be downloaded using the following command: ssc install packagename
Packages in R are downloaded by calling (incl. the quotation marks): install.packages("packagename")

- *quantileTreatmentEffects*: this program creates all panels in Figure 3. In the process of this data is process, local polynomials are fit and the difference along with its 95%-CI is calculated. Finally, all quantities are plotted.

- *StackingRegressionModel*: unlike the other programs, this program was downloaded (and is openly available[3] from the website of *estout*-project. Given a set of regressions with a common dependent variable, this program stacks the regressions output in order to display a particular coefficient from each regression in a single cell.

- *StudyDistrictComparison*: similar in structure to the first 2 programs, this program is tailored to the data structure of the district comparison.

4 Detailed list of content

We describe the contents of every sub-directory contained in the replication zip-folder:

- *analysis_code*: "Table_1ab.do" through "Table_F3ab.do" create all tables in Stata. Two additional do-files to create figures in Stata and two R-code files to create figures in R

- *compile_tables_figures*: the tex-file "tables_figures_all.tex" which compiles a pdf with all tables and figures. All the tables and figures need to be created in Stata/R first.

- *output*: initially empty and can be populated by running the do- and R-code files

- *utilities*: see Section 3.3

- *master_analysis.do*: see Sections 1.1 and 1.2

- *data*: this folder contains all data files used in the analysis. In alphabetical order
  - all_jc_merged_attritionanalysis.dta: contains information on which jobcards were in the endline sampling frame but not in the baseline frame, or vice versa
  - all_pens_merged_attritionanalysis.dta: same as above, but for pension card records
  - balance for ap mandal comparison from 01 village directory.dta: mandal averages of village facilities from 2001 census
  - balance for ap mandal comparison.dta: 2011 census records at mandal level
  - bl nregs expenditures at gp.dta: Gram Panchayat-wise records from 2010 of funds spent on certain NREGS categories (labor, materials, etc.)
  - bl_result codes for nrj at gp.dta: Gram Panchayat averages of baseline survey response codes (not participate, household not found, etc.)
  - blandel_nrjcards_pergp.dta: GP-wise counts of active jobcards for 2010 and 2012
  - bleldata for sample characteristics analysis nregs.dta: household characteristics from baseline and endline NREGS surveys with a survey indicator
  - bleldata for sample characteristics analysis ssp.dta: same as one above, but for SSP survey
  - cardedgp_nregs.dta: records of GP conversion dates for the Smartcards NREGS implementation
  - cardedgp_ssp.dta: same as the one before for the SSP system
  - cens2011_dist_rural.dta: 2011 census district-wise totals for rural areas

[3] See `http://repec.org/bocode/e/estout/` under "Advanced"; the program is referred to as *-appendmodels-*

- cens2011_dist_total.dta: 2011 census district-wise totals
- district balance from village directory data.dta: district averages of village facilities from 2011 census
- el nregs expenditures at gp.dta: same as "bl nregs expenditures at gp.dta" but for 2012
- el_result codes for nrjs.dta: NREGS household survey response codes (used to identify ghosts for instance) at endline
- el_result codes for nrps.dta: same as the one above, but for the SSP survey
- gp total disbursement.dta: GP-week level records of funds disbursed on NREGS
- hawthorneanalysis audit on nregs muster rolls.dta: combining GP-wise records on NREGS work from official NREGS records with information on our worksite audits
- hawthorneanalysis audits on survey.dta: combining self-reported information on NREGS work (from hhd survey) with information on our worksite audits
- hawthorneanalysis hhd on worksite audit.dta: combining records of NREGS work from our work site audits with information on the household survey as conducted in the GP
- hawthorneanalysis hhd survey on official.dta: combining official NREGS records with household survey information
- hawthorneanalysis recon on nregs muster rolls.dta: conbining official NREGS records with information on worksite reckon missions we conducted as part of the survey
- hh_ssp_survey_attrition.dta: household characteristics along with an indicator for pension record presence in baseline sampling frame (SSP)
- hh_survey_attrition.dta: same as one before, but for the NREGS survey
- jcard muster data complete.dta: work records on all jobcards in treatment and control mandals
- nregs baseline leakage at hhd.dta: household totals for official and survey payments as well as leakage at baseline
- nregs baseline leakage gp mean hh.dta: baseline GP averages of leakage (and payments) household totals
- nregs baseline leakage gp mean indiv.dta: baseline GP averages of individual leakage (and payment) totals
- nregs baseline result codes for nrj.dta: household response codes for NREGS households at baseline
- nregs baseline survey *.dta: these four files pertain to a specific section/specific sections of the NREGS baseline survey. "section general" contains various sections (E, F, G) as well as household characteristics
- nregs leakage.dta: individual-level leakage (and payment) outcomes for NREGS at endline.
- nregs leakage_notnamematched_athhd.dta: leakage (and payment) outcomes where household totals are calculated (rather than name-matched individual records as in "nregs leakage.dta"). Includes district-wise scaling factors computed from average # of jobcards per household
- nregs rollout monthly data.dta: official records on Smartcard rollout by month
- nregs survey *.dta: as for the baseline, these are all variables from the endline NREGS survey organized by section(s)

- nregsofficialusage_t1a.dta: official information on NREGS Smartcard conversion status of GPs (cardedGP) or fraction of carded payments (treatment_level)
- number of workers at worksite audits.dta: records on number of NREGS workers present at certain worksites which we independently surveyed in certain GPs
- officialrecords_survey_vs_nonsurvey_nregs.dta: data from official NREGS records for all sampled househods at endline; used to compare those we could survey and those we could not survey
- officialrecords_survey_vs_nonsurvey_ssp.dta: same as the one above but SSP households
- randomization balance data.dta: mandal-wise records pertaining to NREGS and SSP programs provided by the government for year 2010
- randomization balance from village directory data.dta: mandal averages of village facilities from 2011 census; only treatment and control mandals
- spatial exposure measure at gp.dta: variables pertaining to the spatial analysis indicating the spatial exposure measure at the GP-level
- baseline, endline and leakage for SSP files: see the respective file for NREGS which go by the same. The setup is analogous
- ssp rollout monthly data.dta: official records on various dimension of implementation of Smartcards within the SSP system
- sspofficialusage_t1b.dta: SSP-analog to "nregsofficialusage_t1a.dta" above