The production process of the Global MPI

Nicolai Suppa

German Stata Users Group Meeting
Munich, Germany
May 2019
Outline

1 Introduction

2 Key elements of the production process

3 Concluding Remarks
What is the global MPI?

- a multidimensional **poverty measure**
  - see Alkire and Foster (2011); Sen (1992); Alkire and Santos (2014); Alkire et al. (2018)
- available for 100+ countries (and 1200 sub-national regions)
- developed and published by OPHI and UNDP
- published since 2010
The global MPI

Computational aspects

- all figures are obtained from a *single survey* per country
- **numerous measures** are calculated for each country
  - headcount, intensity, adj. headcount, (un-) censored headcounts,...
- most numbers can be **disaggregated** by area, region, and age group
- (normative) parametric choices require **sensitivity checks**
  - deprivation cutoffs, weighting schemes, poverty cutoff, ...
  - not all measure–parameter–combinations are needed

\[ N: \text{5k–2.7m with } N_{med} \approx 50k; \ # \ of \ estimates \approx 130k \]

Other aspects

- a **highly standardised**, but **not entirely fixed** project.
- **well-defined deliverables**, e.g., excel sheets, country briefings, ...
- relatively small team and not all are Stata experts or even Stata users
Related literature

• Previous work on workflow considerations and programming in Stata:

- The Workflow of Data Analysis Using Stata
- Data Management Using Stata
- Data Analysis Using Stata

The Stata Journal (2005)
5, Number 4, pp. 560–566

Suggestions on Stata programming style

Nicholas J. Cox
Durham University, UK
n.j.cox@durham.ac.uk

Abstract. Various suggestions are made on Stata programming style, under the headings of presentation, helpful Stata features, respect for datasets, speed and efficiency, reminders, and style in the large.

Keywords: pr0018, Stata language, programming style
Motivation

• well-conceived **workflow is vital** for any large-scale project

→ why sharing?

1. transparency: how is the GMPI computed?
2. share some experience and lessons & how to refine this process?
3. illustrate workflow-related problems & implications of coding decisions

• general workflow questions receive **rather little attention**
  ▶ hard to de-contextualise (typically project-specific)
  ▶ often work-flow decisions may (i) not be recognised as such or (ii) alternative solutions make no real difference in practice

• aspects of the present workflow may be **relevant in other settings**
  e.g., other cross-country studies
  e.g., juggling with a plethora of estimates
  e.g., other large scale projects where ‘tiny’ coding tweaks make a difference

• **small ‘innovations’**: results file, reference sheet, spelling sheet, etc.
Desiderata
The 2018 revision

1. improve **efficiency** in general
   - estimation time and storage

2. ensure **replicability** and tractability
   - track down and fix errors

3. achieve **flexibility**
   - re-estimate selected countries or measures

4. low **maintenance costs**
   - Stata skills & feasible revisions

5. develop a more **widely applicable approach** to MPI-estimation

6. increase the number of **default estimates** (e.g., disaggregations, SE)
The basic workflow

- raw micro data
- data prep
- micro data
- estimation
- data dump
- compilation of results
- assembling results
- GMPI2018.dta
- graphs
- country briefings
- data export
- data viz
- reference sheet
- external data
- map production
- labelling

Nicolai Suppa
Munich, Germany, May 2019
The results file

Principle structure

• each estimate is an observation
• each estimate can be uniquely identified using auxiliary variables
e.g., cty, measure, k, wgts, loa, indicator, ...
The master do-file

- designed for **interactive use** (day-to-day work)
  1. **reference sheet** production > extdta prep > spelling sheet
  2. re-run data prep > certification scripts > quality checks
  3. estimation
  4. convert and compile
  5. assemble cleaned results file
  6. deliverables: graphs, excel sheets, **country briefs**, export for data viz

**Tool:** `ctyselect` ➔ returns country codes in `r(ctylist)`

```bash
ctyselect ccty
ctyselect ccty, r(^A)
ctyselect ccty, s(IND)
```
The reference sheet

- contains country and region level information
  - separates estimation from housekeeping (incl. merge of external data)
  - reduces data carried through estimation
  - allows parallel processing
  - simplifies some quality checks
  - key information can be quickly obtained through entire process

<table>
<thead>
<tr>
<th>ccty</th>
<th>ccnum</th>
<th>region</th>
<th>survey</th>
<th>year</th>
<th>region_n-e</th>
<th>fname</th>
<th>fdate</th>
<th>adate</th>
</tr>
</thead>
</table>

Tool: refsh

refsh using path2refsh, rebuild char(ccty survey year) ///
    id(ccty) region(region) path(path2microdata)
Estimation and storing

The principle approach

```
eststo H'{k}'_{'subg'}: svy: mean I_{'k'}, over('subg')
estadd loc measure "H"
estadd loc scalar k = 33
estadd loc loa "{'subg'}"
```

- for `eststo`, `estadd`, see Jann (2005, 2007)

```
estwrite * using path/{'cty'}_{'subg'}.sters , replace
est clear
```

- however, single mega loop is dysfunctional
  - i.e. several nested loops over k, dimensions, subgroup, ...
- **grouping of estimates** to achieve flexibility and avoid Stata limits
  - along cty and loa (national, regional, ...)
  - along auxiliary, main, and dimensional quantities
Estimation and storing

The packaged approach

**Tool: mpi_set, mpi_est**

```bash
mpi_set, d1(d_cm d_nutr, name(hl)) ///
   d2(d_satt d_educ, name(ed)) ///
   d3(d_elct d_sani d_wtr d_hsg d_asst d_ckfl, name(ls)) ///
   name(GMPI)

mpi_est, estsave(path/‘cty’_nat_aux.sters, replace) ///
   name(GMPI) aux(all) addmeta(ccty=‘cty’)

mpi_est, k(01 10 20 33 40 50) weights(equal) name(GMPI) ///
   measures(all) measures(dim(all) kdim(1 20 33 40 50) gen
```

**Tools**

- `gafvars, mpi_setwgts, genwgts, addmetainfo,...`
Dumping and compiling the results
Principle and packaged approach

1. estread each ster-file, and for each estimate
2. dump results into data using _coef_table and xsvmat

<table>
<thead>
<tr>
<th></th>
<th>b</th>
<th>se</th>
<th>t</th>
<th>pvalue</th>
<th>ll</th>
<th>ul</th>
<th>df</th>
<th>crit</th>
<th>eform</th>
</tr>
</thead>
<tbody>
<tr>
<td>c_33_equal:0</td>
<td>0.23306712</td>
<td>0.00900391</td>
<td>25.885091</td>
<td>8.734e-99</td>
<td>0.21538287</td>
<td>0.25075137</td>
<td>580</td>
<td>1.9640625</td>
<td>0</td>
</tr>
<tr>
<td>c_33_equal:1</td>
<td>0.10311217</td>
<td>0.00753559</td>
<td>13.683353</td>
<td>3.882e-37</td>
<td>0.0883118</td>
<td>0.11791255</td>
<td>580</td>
<td>1.9640625</td>
<td>0</td>
</tr>
</tbody>
</table>

3. add locals or scalars from estimates as variables (e.g., loa, k,...)
4. append all dumped estimates of this ster-file

Tool: est2dta

ctyselect cty , s(IND BGD ETH PER)
foreach cty in `r(ctylist)` {
    est2dta, inpath(path2sters) outpath(path2dta) llist(loa indicator measure wgts spec ccty) slist(N k time timedata) clist(`cty')
}

Nicolai Suppa
Munich, Germany, May 2019
Global MPI Country Briefing 2018: India (South Asia)

The Global MPI

The global Multidimensional Poverty Index (MPI) was created using the multidimensional measurement method of Alkire and Foster (AF). The global MPI is an index of acute multidimensional poverty that covers over 100 countries. It is computed using data from the most recent Demographic and Health Surveys (DHS), Multiple Indicator Cluster Surveys (MICS), Pan Arab Project for Family Health (PAPFAM) and national surveys. The MPI has three dimensions and 10 indicators as illustrated in figure 1. Each dimension is equally weighted, and each indicator within a dimension is also equally weighted. Any person who fails to meet the deprivation cutoff of MPI is identified as deprived in that indicator. So the core information the MPI uses is the profile of deprivations each person experiences. Each deprivation indicator is defined in table A.1 of the appendix.

In the global MPI, a person is identified as multidimensionally poor or MPI poor if they are deprived in at least one third of the weighted MPI indicators. In other words, a person is MPI poor if the person’s weighted deprivation score is equal to or higher than the poverty cutoff of 33.33%. Following the AF methodology, the MPI is calculated by multiplying the incidence of poverty (I) and the average intensity of poverty (A).

Specifically, I is the proportion of the population that is multidimensionally poor, while A is the average proportion of dimensions in which poor people are deprived. So, MPI = I × A, reflecting both the share of people in poverty and the degree to which they are deprived.

Table 1. Global MPI in India

<table>
<thead>
<tr>
<th>Area</th>
<th>MPI</th>
<th>I (%)</th>
<th>A (%)</th>
<th>Vulnerable Poverty</th>
<th>Severe Poverty</th>
<th>Population Share</th>
</tr>
</thead>
<tbody>
<tr>
<td>National</td>
<td>0.121</td>
<td>27.5%</td>
<td>43.9%</td>
<td>19.1%</td>
<td>8.6%</td>
<td>100%</td>
</tr>
<tr>
<td>Rural</td>
<td>0.161</td>
<td>36.5%</td>
<td>44.1%</td>
<td>21.8%</td>
<td>11.6%</td>
<td>67.3%</td>
</tr>
</tbody>
</table>

Notes: Source: DHS year 2015-2016, own calculations.

A headcount ratio is also estimated for two other ranges of poverty cutoffs. A person is identified as vulnerable to poverty if they are deprived in 20–33.33% of the weighted indicators. Concurrently, a person is identified as living in severe poverty if they are deprived in 50–100% of the weighted indicators. A summary of the global MPI statistics are presented in table 1 for national, rural and urban areas.

A brief methodological note is published following each round of global MPI update. For example, for the global MPI December 2018 update, please refer to Alkire et al. (2018). The note explains the methodological adjustments that were made while revising and standardizing indicators for over 120 countries. As such, it is useful to refer to the methodological notes with this country brief for specialized information on how the country survey data was managed.

Poverty Headcount Ratios

Figure 2 compares the headcount ratios of the global MPI and monetary poverty measures. The height of the first bar of figure 2 shows the percentage of people who are MPI poor. The second and third bars represent the percentage of people who are poor according to the World Bank’s $1.90 a day and $3.10 a day poverty line. The final bar denotes the percentage of people who are poor according to the national income or consumption and expenditure poverty measures.
Graph and country brief production

- graphs for other countries or parameter choices are easy to obtain
- use (i) \LaTeX-template, (ii) rely on \LaTeX-variables, (iii) ctyselect

```latex
\tempname lc
\fileopen lc using lc.tex , wt replace
\filew lc "\newcommand\ctynam{\texttt{ctynam}}\_n \///
   "\newcommand\ctycode{\texttt{ctycode}}\_n \///
   "\newcommand\calcyyear{\texttt{year}}\_n \///
   ...
\fileclose lc
...
\pdflatex --interaction=nonstopmode --shell-escape
   \input{CB_template.tex}
\mv "CB_template.pdf" "pdfs/CB_\texttt{ctycode}.pdf"
```

- Latex includes country-specific figures and omits entire section if needed.
Other ‘innovations’

- **certification scripts** for cleaned micro data:
  - check existence and data type of key variables (**confirm**), check for sensible values (**assert**), and non-empty data characteristics.
  - reduces the probability of loop breaking
  - saves time, even though other quality checks are still needed

- **spelling sheet:**
  1. clean country and regions names, e.g., using **proper()**
  2. export cleaned region names (and IDs) into dedicated spreadsheet
  3. let copy-editor suggest revised names in separate column (if needed)
  4. generate and update variable for labels

- systematic cross-release **folder structure** (e.g., portability)

- **time stamps** for both estimates and the underlying micro data

- **data characteristics** to hand-over information
‘Innovations’

- flexible results dta
- reference sheet (conditional independence of results & housekeeping)
- certification scripts for cleaned micro data
- spelling sheet (based on reference sheet)
- sensible partitioning of estimations
- data characteristics to hand-over information
Lessons

• a sensible workflow has many benefits
  ▶ often simpler and **cleaner code** (e.g., missing indicators)
  ▶ may allow sensible **packaging** of the code
  ▶ allows instructive **benchmarking** and future **revisions**
  ▶ simplifies documentation
  ▶ ...

• however, developing a sensible work flow was **not trivial**
  ▶ required lots of discussion, experimentation and time
  ▶ ‘pure’ coding decisions can determine the workflow, and therefore, should be recognised as such in the first place.

• anticipate performance relevant issues to easier identify bottlenecks, when project is scaled up
  ▶ variable generation, data types, order of loops and degree of nesting, ...
Open issues

• documentation:
  ▶ Stata help files, desktop companion, paper, presentations, ...

• performance tweaks:
  ▶ so far based on user-experience, little systematic benchmarking

• more comprehensive packaging
  ▶ interesting for other scenarios: i.e. stand-alone toolbox?

• add additional quality checks in certification scripts

• review code and replace re-invented wheels, if more efficient.

• which other aspects could be interesting for a wider audience?
  ▶ ancient coding decisions, which turned out to be problematic
  ▶ difficult trade-offs faced during revision
  ▶ contextual factors
  ▶ ...

Nicolai Suppa
Munich, Germany, May 2019
References


