The production process of the Global MPI

Nicolai Suppa

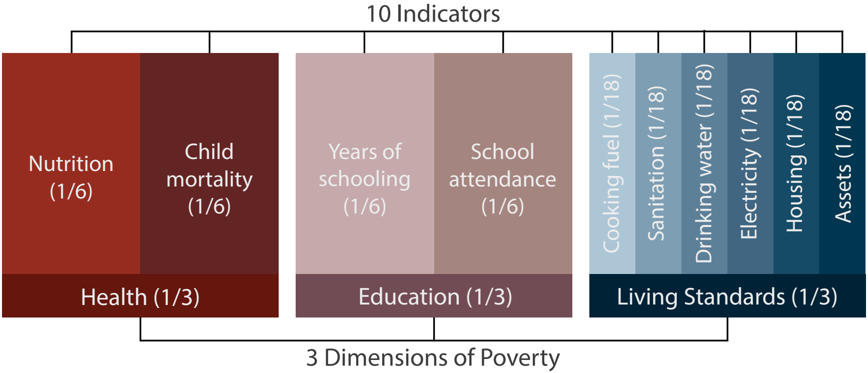
German Stata Users Group Meeting
Munich, Germany
May 2019

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

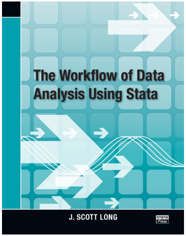
- all figures are obtained from a **single survey** per country
- **numerous measures** are calculated for each country
  - headcount, intensity, adj. headcount, (un-) censored headcounts,...
- most numbers can be **disaggregated** by area, region, and age group
- (normative) parametric choices require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

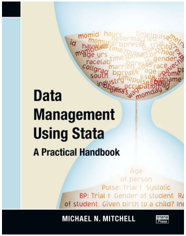
➜ N: 5k–2.7m with $N_{med} \approx$ 50k; # of estimates $\approx$ 130k

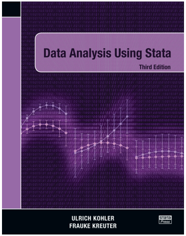
Other aspects

- 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

- Previous work on workflow considerations and programming in 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

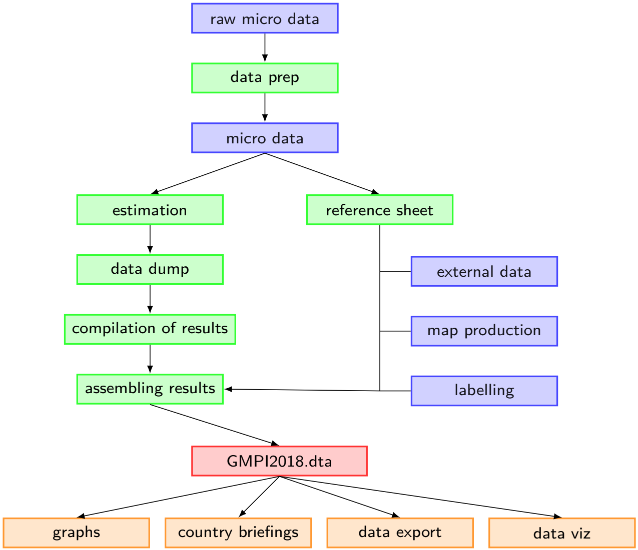
  e.g., other large scale projects where 'tiny' coding tweaks make a difference
- **small 'innovations'**: results file, reference sheet, spelling sheet, etc.

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

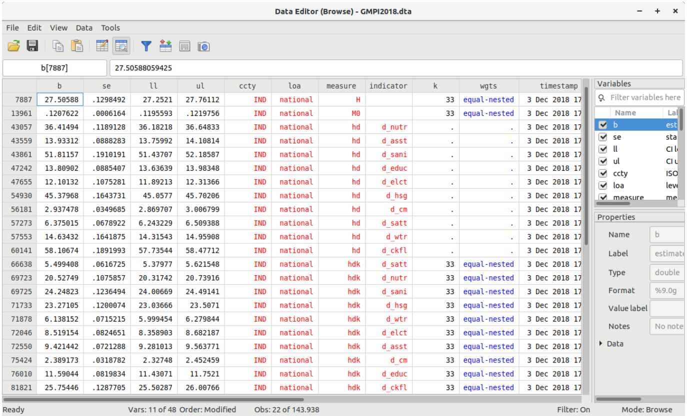


The results file

Principle structure

- each estimate is an observation
- each estimate can be uniquely identified using auxiliary variables

  e.g., `cty`, `measure`, `k`, `wgts`, `loa`, `indicator`, ...



	b	se	ll	ul	ccty	loa	measure	indicator	k	wgts	timestamp
7887	27.50588	.1298492	27.2521	27.76112	IND	national	H		33	equal-nested	3 Dec 2018 17
13961	.1207622	.0006164	.1195593	.1219756	IND	national	M0		33	equal-nested	3 Dec 2018 17
43057	36.41494	.1189128	36.18218	36.64833	IND	national	hd	d_nutr	.	.	3 Dec 2018 17
43559	13.93312	.0888283	13.75992	14.10814	IND	national	hd	d_asst	.	.	3 Dec 2018 17
43861	51.81157	.1910191	51.43707	52.18587	IND	national	hd	d_sani	.	.	3 Dec 2018 17
47242	13.80902	.0885407	13.63639	13.98348	IND	national	hd	d_educ	.	.	3 Dec 2018 17
47655	12.10132	.1075281	11.89213	12.31366	IND	national	hd	d_elct	.	.	3 Dec 2018 17
54930	45.37968	.1643731	45.0577	45.70206	IND	national	hd	d_hsg	.	.	3 Dec 2018 17
56181	2.937478	.0349685	2.869707	3.006799	IND	national	hd	d_cm	.	.	3 Dec 2018 17
57273	6.375015	.0678922	6.243229	6.509388	IND	national	hd	d_satt	.	.	3 Dec 2018 17
57553	14.63432	.1641875	14.31543	14.95908	IND	national	hd	d_wtr	.	.	3 Dec 2018 17
60141	58.10674	.1891993	57.73544	58.47712	IND	national	hd	d_ckfl	.	.	3 Dec 2018 17
66638	5.499408	.0616725	5.37977	5.621548	IND	national	hdk	d_satt	33	equal-nested	3 Dec 2018 17
69723	20.52749	.1075857	20.31742	20.73916	IND	national	hdk	d_nutr	33	equal-nested	3 Dec 2018 17
69725	24.24823	.1236494	24.00669	24.49141	IND	national	hdk	d_sani	33	equal-nested	3 Dec 2018 17
71733	23.27105	.1200074	23.03666	23.5071	IND	national	hdk	d_hsg	33	equal-nested	3 Dec 2018 17
71878	6.138152	.0715215	5.999454	6.279844	IND	national	hdk	d_wtr	33	equal-nested	3 Dec 2018 17
72046	8.519154	.0824651	8.358903	8.682187	IND	national	hdk	d_elct	33	equal-nested	3 Dec 2018 17
72550	9.421442	.0721288	9.281013	9.563771	IND	national	hdk	d_asst	33	equal-nested	3 Dec 2018 17
75424	2.389173	.0318782	2.32748	2.452459	IND	national	hdk	d_cm	33	equal-nested	3 Dec 2018 17
76010	11.59044	.0819834	11.43071	11.7521	IND	national	hdk	d_educ	33	equal-nested	3 Dec 2018 17
81821	25.75446	.1287705	25.50287	26.00766	IND	national	hdk	d_ckfl	33	equal-nested	3 Dec 2018 17

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)`

```
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

ccty	ccnum	region	survey	year	region_n~e	fname	fdate	adate
BGD	050	1	DHS	2014	barisal	bgd_dhs14.dta	22 Apr 2019 12:22	29 Apr 2019 17:36
BGD	050	2	DHS	2014	chittagong	bgd_dhs14.dta	22 Apr 2019 12:22	29 Apr 2019 17:36
BGD	050	3	DHS	2014	dhaka	bgd_dhs14.dta	22 Apr 2019 12:22	29 Apr 2019 17:36
BGD	050	4	DHS	2014	khulna	bgd_dhs14.dta	22 Apr 2019 12:22	29 Apr 2019 17:36
BGD	050	5	DHS	2014	rajshahi	bgd_dhs14.dta	22 Apr 2019 12:22	29 Apr 2019 17:36
BGD	050	6	DHS	2014	rangpur	bgd_dhs14.dta	22 Apr 2019 12:22	29 Apr 2019 17:36
BGD	050	7	DHS	2014	sylhet	bgd_dhs14.dta	22 Apr 2019 12:22	29 Apr 2019 17:36
BGD	050	.	DHS	2014		bgd_dhs14.dta	22 Apr 2019 12:22	29 Apr 2019 17:36

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

```
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) measuresdim(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`

	b	se	t	pvalue	ll	ul	df	crit	eform
c_33_equal:0	.23306712	.00900391	25.885091	8.734e-99	.21538287	.25075137	580	1.9640625	0
c_33_equal:1	.10311217	.00753559	13.683353	3.882e-37	.0883118	.11791255	580	1.9640625	0

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 ccty , 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')
}
```

Graph and country brief production

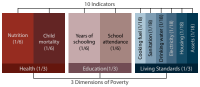
Oxford Poverty and Human Development Initiative (OPHI)
Oxford Department of International Development
Queen Elizabeth House, University of Oxford
www.ophi.org.uk

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).[1] 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.[2] Any person who fails to meet the deprivation cutoff 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.

Figure 1. Structure of the Global MPI



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 (H) and the average **intensity** of poverty (A). More specifically, H 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, $MPI = H \times A$, reflecting both the share of people in poverty and the degree to which they are deprived.

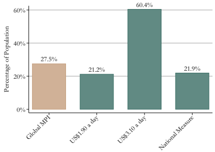
Table 1. Global MPI in India

Area	MPI	H	A	Vulnerable	Severe Poverty	Population Share
National	0.121	27.5%	43.9%	19.1%	8.6%	100.0%
Urban	0.039	9.0%	42.6%	13.7%	2.4%	32.7%
Rural	0.161	36.5%	44.1%	21.8%	11.6%	67.3%

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

[1]A formal explanation of the method is presented in Alkire and Foster (2011). An application of the method is presented in Alkire and Santos (2014).
[2]It should be noted that the AF method can be used with different indicators, weights and cutoffs to develop national MPIs that reflect the priorities of individual countries. National MPIs are more tailored to the context but cannot be compared.

Figure 2. Headcount Ratios by Poverty Measures



Notes: Source for global MPI: DHS, year 2015-2016, own calculations. Monetary poverty measures are the most recent estimates from World Bank (World Bank, 2018). Monetary poverty measure refer to 2011 ($1.90 a day), 2011 ($3.10 a day), and 2011 (national measure).

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 100 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.[3]

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.

[3]Previous methodological notes, published for each round of update, are made available on the OPHI website:http://ophi.org.uk/multidimensional-poverty-index/mpi-resources/.

- 1 for each country, 9–12 pages, up to 9 figures and 2 tables
- some countries lack section 'Subnational Analysis'

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`

```
tempname lc
file open `lc' using lc.tex , w t replace
file w `lc' "\newcommand\ctyname{`ctyname'}" _n ///
    "\newcommand\ctycode{`ctycode'}" _n ///
    "\newcommand\calcyear{`year'}" _n ///
    ...
file close `lc'
...
!pdflatex --interaction=nonstopmode --shell-escape
   \input{CB_template.tex}
!mv "CB_template.pdf" "pdfs/CB_`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
  - ...

References

Alkire, S. and Foster, J. (2011). Counting and multidimensional poverty measurement. *Journal of Public Economics*, 95(7-8):476–487.

Alkire, S., Kanagaratnam, U., and Suppa, N. (2018). The global multidimensional poverty index (MPI): 2018 revision. OPHI MPI Methodological Notes 46, Oxford Poverty and Human Development Initiative, University of Oxford.

Alkire, S. and Santos, M. E. (2014). Measuring acute poverty in the developing world: Robustness and scope of the multidimensional poverty index. *World Development*, 59:251–274.

Cox, N. (2005). Suggestions on stata programming style. *The Stata Journal*, 5(4):560–566.

Jann, B. (2005). Making regression tables from stored estimates. *The Stata Journal*, 5(3):288–308.

Jann, B. (2007). Making regression tables simplified. *The Stata Journal*, 7(2):227–244.

Kohler, U. and Kreuter, F. (2012). *Data Analysis Using Stata, Third Edition*. Stata Press.

Long, J. S. (2008). *The Workflow of Data Analysis Using Stata*. Stata Press.

Mitchell, M. N. (2010). *Data Management Using Stata*. Stata Press.

Sen, A. K. (1992). *Inequality Reexamined*. Russell Sage Foundation book. Russell Sage Foundation, New York, 3 edition.