Small area estimation and Poverty map in Stata

The World Bank

The Poverty and Equity Global Practice

Global Solutions Group on Welfare Measurement and Statistical Capacity

Minh Cong Nguyen

Paul Andres Corral Rodas

João Pedro Wagner De Azevedo

Qinghua Zhao

July 25, 2017

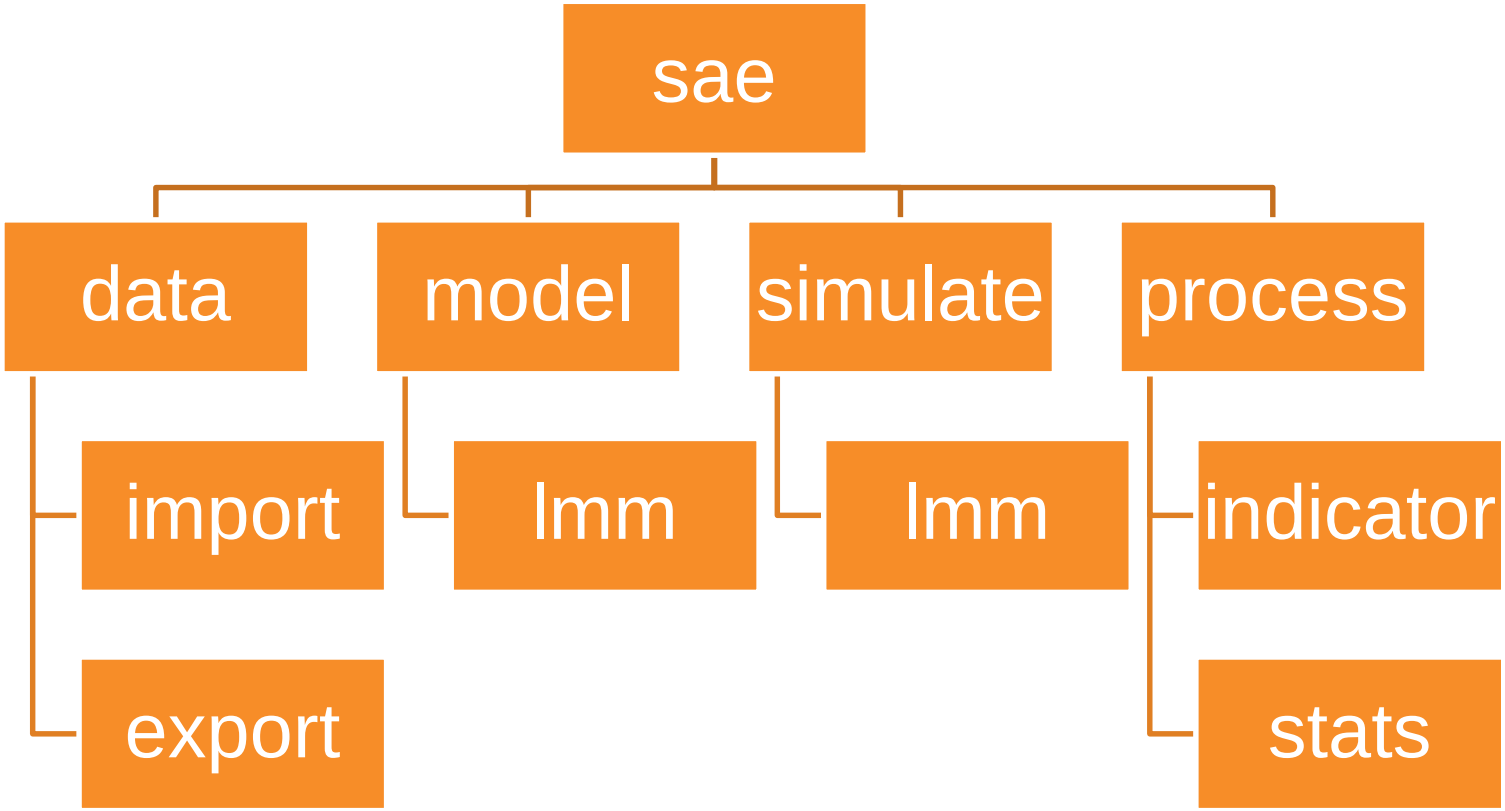
Demand for small estimation are increasing

- New Census round 2020 is coming
- More and more spatial and geo-coded data is available
- Demand for spatial distribution of variables such welfare, consumption/expenditure are on the rise
- Knowledge expansion and desire to advance in this topic drive us to work in details to start the project.
- Replicability of the results and flexibility in modeling, estimations, and simulation are of our interest.
- There is no small area estimation command in Stata.

Contribution

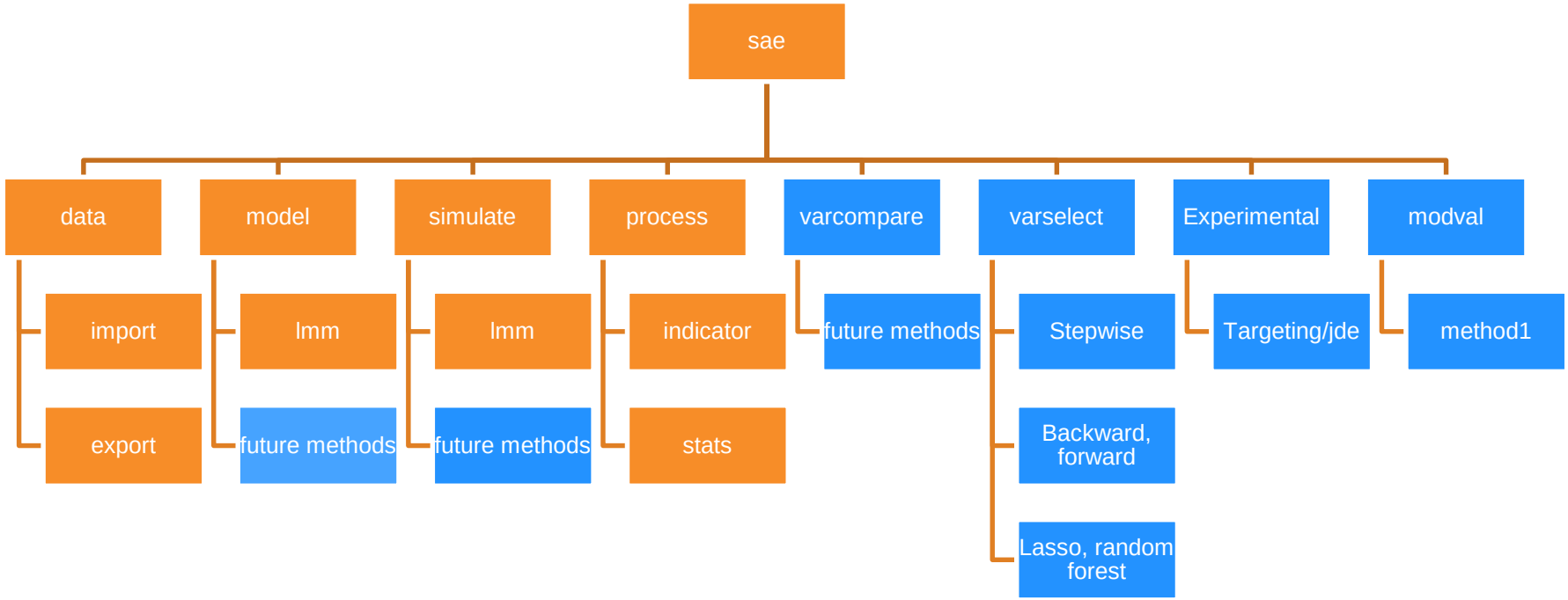
- We introduce a suite of small area estimation commands (**sae**) in Stata that set the base for future work in this topic for the community of Stata users.
- Structure of the commands are intuitive for future integration of the new methods or functions.
- Mata functions and codes are open source and can be linked with new functions or methods by any author or collaborators from the Stata community.
- Using Mata matrix file for storing and retrieving vectors of data quickly are useful when the data is a very large and the method requires intensive matrix computation.

Framework of the Stata sae syntax [*current*]



lmm: linear mixed model

Framework of the Stata sae syntax [future/plan]



WORLD BANK GROUP

Examples of the sae suite commands

- `sae data import`: This is used to import the target dataset to a more manageable format for the simulations (Mata data)
- `sae data export`: Used to export the resulting simulations to dta format
- `sae model lmm`: Used for obtaining the GLS estimates from the 1st stage. Used for testing your model.
- `sae sim lmm`: Obtains the same parameters as the previous step, but performs the Monte Carlo simulation based on census/target data
- `sae proc indicator`: Gets poverty and inequality indicators based on the simulated census vectors
- `sae proc stats`: Gets profile from group classification based on the simulated census vectors

Example code and output with sae

```
. sae model lmm $lhs $selected [aw=weight], area(lid)  varest(h3) ///
>         zvar(dbicycle div_2 highstedu) yhat2(delectric hhsize hhsize2)
You chose H3, parameters must be obtained via bootstrap I changed it for you.
WARNING: 0 observations removed due to less than 3 observations in the cluster.

OLS model:
```

lnrpcexp	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
dbicycle	.0453394	.0447639	1.01	0.311	-.0423964	.1330751
delectric	.1851161	.0420334	4.40	0.000	.1027321	.2675
dfreeze	1.003551	.0820352	12.23	0.000	.8427653	1.164338
div_1	-.193889	.0610794	-3.17	0.002	-.3136025	-.0741755
div_2	.0397289	.0504635	0.79	0.431	-.0591778	.1386357
div_5	-.1042349	.0434166	-2.40	0.016	-.1893298	-.01914
durban	-.1578936	.0463934	-3.40	0.001	-.2488229	-.0669643
hd_age	.0040713	.0013134	3.10	0.002	.0014971	.0066455
hhsize	-.1403914	.0340885	-4.12	0.000	-.2072037	-.0735791
hhsize2	.0062649	.0025765	2.43	0.015	.0012151	.0113147
highstedu	.0360035	.0053334	6.75	0.000	.0255503	.0464568
n15_59yrp	.2646748	.0911977	2.90	0.004	.0859306	.443419
_cons	6.823911	.1355553	50.34	0.000	6.558227	7.089594

Alpha model:

Residual	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
dbicycle	-.6170202	.3138316	-1.97	0.049	-1.232119	-.0019216
div_2	-.7807481	.3468736	-2.25	0.024	-1.460608	-.1008882
highstedu	.1323952	.0396876	3.34	0.001	.054609	.2101814
delectric_yhat2	-.0082707	.0054651	-1.51	0.130	-.018982	.0024407
hhsize_yhat2	.0057893	.0051092	1.13	0.257	-.0042245	.0158032
hhsize2_yhat2	-.0005463	.0003732	-1.46	0.143	-.0012778	.0001852
_cons	-6.518535	.6360782	-10.25	0.000	-7.765225	-5.271844

GLS model:

lnrpcexp	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
dbicycle	.0850408	.0372959	2.28	0.023	.0119421	.1581395
delectric	.2220866	.036005	6.17	0.000	.151518	.2926552
dfreeze	.9344732	.0819585	11.40	0.000	.7738375	1.095109
div_1	-.1783579	.1003378	-1.78	0.075	-.3750165	.0183006
div_2	.0507601	.0750079	0.68	0.499	-.0962526	.1977728
div_5	-.1062534	.067863	-1.57	0.117	-.2392624	.0267556
durban	-.1667263	.066672	-2.50	0.012	-.2974009	-.0360516
hd_age	.0041828	.0010604	3.94	0.000	.0021043	.0062612
hhsize	-.1544192	.0221914	-6.96	0.000	-.1979136	-.1109248
hhsize2	.0067948	.0014894	4.56	0.000	.0038756	.009714
highstedu	.0358874	.0043579	8.23	0.000	.027346	.0444288
n15_59yrp	.2870317	.0701307	4.09	0.000	.1495782	.4244853
_cons	6.839225	.1107267	61.77	0.000	6.622205	7.056246

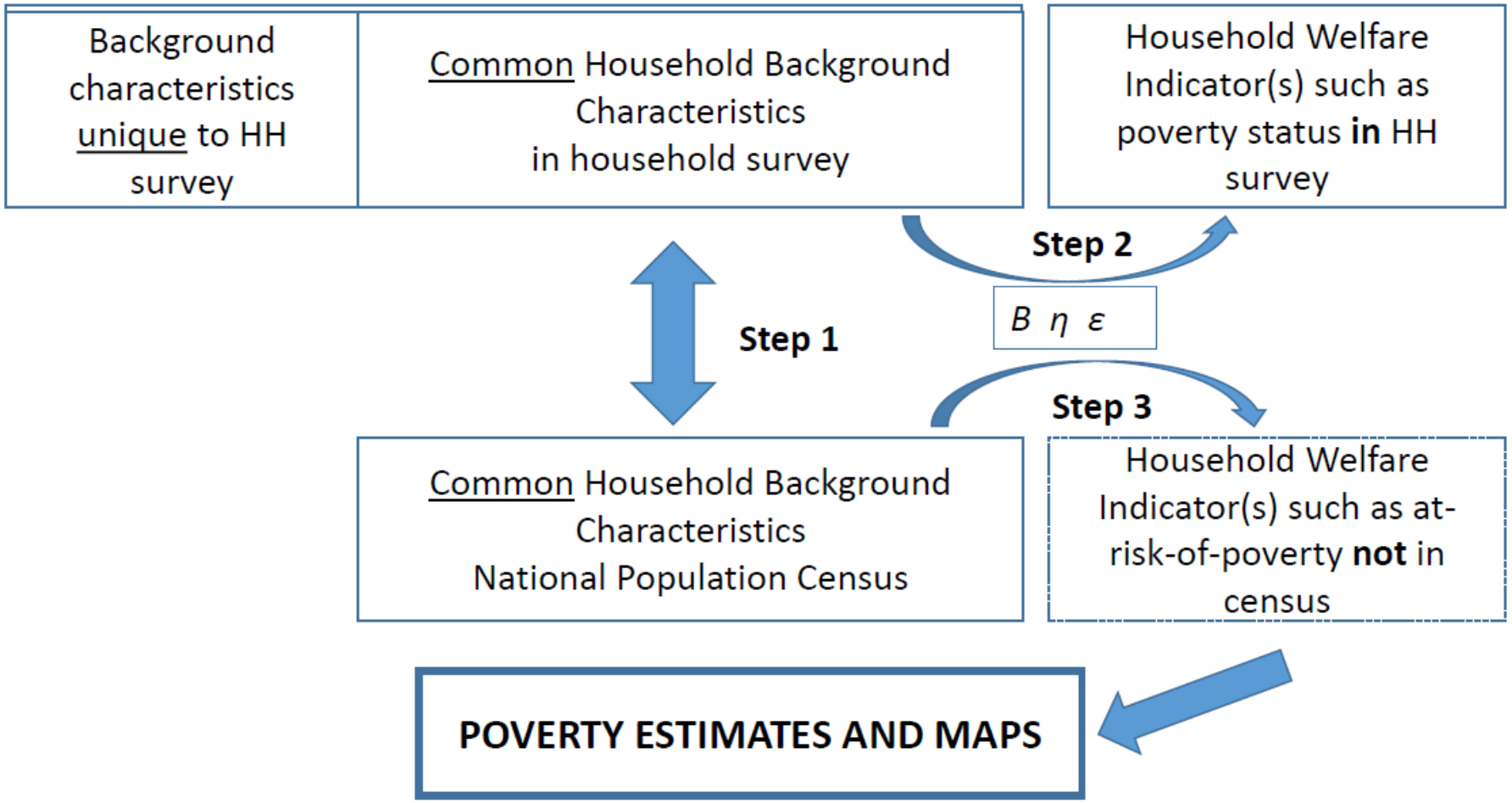
Example code and output with sae

```
Model settings
------------------------------------------------------------------------------
Error decomposition                             H3

Beta model diagnostics
------------------------------------------------------------------------------
Number of observations                =       640
Adjusted R-squared                    =  .49700659
R-squared                             =  .50645247
Root MSE                              =  .37553107
F-stat                                =  53.616198

Alpha model diagnostics
------------------------------------------------------------------------------
Number of observations                =       640
Adjusted R-squared                    =  .03921974
R-squared                             =  .04824115
Root MSE                              =  2.6955813
F-stat                                =  5.3474065

Model parameters
------------------------------------------------------------------------------
Sigma ETA sq.                         =  .029914
Ratio of sigma eta sq over MSE        =  .21212057
Variance of epsilon                   =  .11498459
------------------------------------------------------------------------------
                    <End of first stage>
```

```
Initializing the Second Stage, this may take a while...

Bootstrapped drawing of betas and parameters

Number of simulations: 100
Each dot (.) represents 1 simulation(s).
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100

Finished running the Second Stage
------------------------------------------------------------------------------


Opening plugin
Output bas been loaded into Stata. Please see the results.
```

Variable	Obs	Mean	Std. Dev.	Min	Max
Unit	647	1.34e+11	8.38e+10	0	2.61e+11
nHHLDs	647	77.89799	505.1247	20	10080
nDroppedHHLD	647	0	0	0	0
nIndividuals	647	378.4312	2454.563	60	48969
YTrimed	647	0	0	0	0
nSim	647	100	0	100	100
Min_Y	647	245.2176	58.81206	110.6127	584.1427
Max_Y	647	9579.258	9204.94	2550.198	76258.54
Mean	647	1228.435	411.247	657.3038	4044.871
StdErr	647	257.1216	143.0396	30.54054	1501.675
avg_FG~84208	647	.3800809	.1478174	.0006122	.8074
se_FGT~84208	647	.1703329	.0547565	.0043073	.2921397
avg_FGT~1000	647	.522021	.1608339	.0009184	.9112
se_FGT0_1000	647	.1697267	.0524297	.0052484	.2701606

Overview of the flows



WORLD BANK GROUP

Model specification - 1st stage, aka Beta model

- Estimate the via OLS: $\boldsymbol{y_{ch} = x_{ch}^T \beta + u_{ch}}$
- Units within an area are not independent from one another, where $\widehat{u_c.}$ is the is the average of $\hat{u}_{ch}$ for a specific cluster we get: $\boldsymbol{\hat{u}_{ch} = \hat{u}_{c.} + (\hat{u}_{ch} - \hat{u}_c) = \hat{\eta}_c + \hat{e}_{ch}}$
- We estimate the following: $y_{ch} = x_{ch}^T \beta + \eta_c + e_{ch}$
- Obtain GLS estimates, where $\mathrm{Var}(\boldsymbol{\hat{\eta}_c}) = \sigma_\eta^2$ and $\mathrm{Var}(\boldsymbol{\hat{e}_c}) = \sigma_{e_{ch}}^2$

WORLD BANK GROUP

Elbers, Lanjouw, and Lanjouw (2003)

The ELL method accounts for spatial correlation by allowing for part of the model error to be shared by all households living in the same locality – This common error is referred to as the location error:

$$u_{ch} = \eta_c + \varepsilon_{ch}$$

Households in the same municipality share the same η. The resulting decomposed variance is:

$$E\left[u_{ch}^2\right] = \sigma_\eta^2 + \sigma_\varepsilon^2$$

The larger the variance of η, the less precise the estimates of welfare

- Variance of the location (η) may be lowered by inclusion of cluster level variables (cluster means from the census, satellite, or administrative data)

The variances can be estimated via ELL's proposed methodology or via Henderson's method III

- How the residuals are split under Henderson's method III is different (see ELL, 2002 and Van der Weide, 2014)

Heteroskedasticity – different variances across households

ELL introduces different variances for different households (σ_ε)

- The literature often shows variances of expenditures among rich households are larger than those among poor households
  - In reality, this is an empirical question
  - ELL method estimates variances of errors at the household level from household/individual characteristics and location variables "alpha model"

ELL specify a parametric form of heteroskedasticity, but simplify it by setting $B = 0$ and $A = 1.05max(e_{ch}^2)$

$$E\left[e_{ch}^2\right] = \sigma_{e_{ch}}^2 = \left[\frac{A\exp^{Z'_{bh}\alpha} + B}{1+\exp^{Z'_{bh}\alpha}}\right] \approx \ln\left[\frac{e_{ch}^2}{A - e_{ch}^2}\right] = Z'_{ch}\alpha + r_{ch}$$

Heteroskedasticity – different variances across households

The alpha-model matters for the point estimates as well as for the standard errors

This is because measures of poverty and inequality are non-linear functions of household incomes, and thereby non-linear functions of the error terms

- As a result, the expected value of poverty and inequality measures will be a function of all moments of the error distribution functions

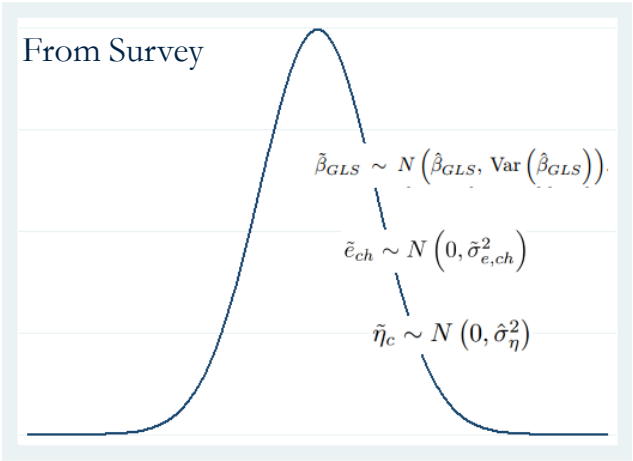
In practice, the adjusted R-squared of the alpha-model is often very modest

- Even so, the estimated poverty rates are not insensitive to the choice of the alpha-model

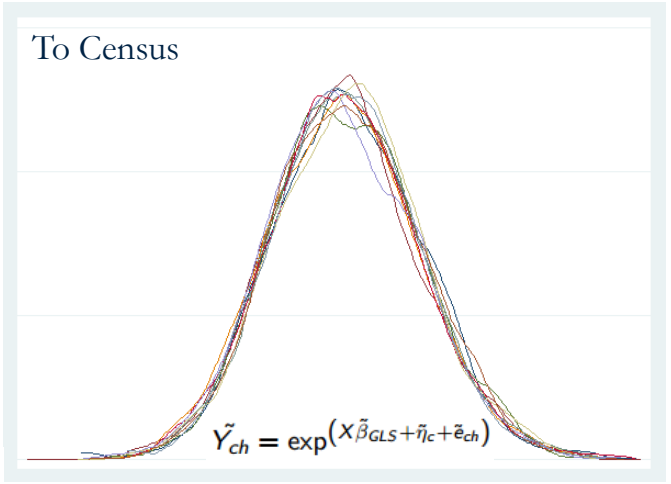
By defining $exp^{z'\alpha} \equiv D$ and using the Delta Method (Taylor expansion for $E[\sigma_{ch}^2]$) we get:

$$\hat{\sigma}_{e,ch}^2 \approx \left[\frac{AD}{1+D}\right] + \frac{1}{2}\widehat{\mathrm{Var}}(r)\left[\frac{AD(1-D)}{(1+D)^3}\right]$$

Monte Carlo Simulation (aka Second Stage)



- The goal is to simulate a sufficiently large number of census vectors of welfare to allow for reliable estimates of poverty (usually 100)
- From the first stage parameters it is possible to take random drawings from the assumed distributions
  - Alternatively it is possible to get bootstrapped samples of the survey data to yield all parameters needed for simulating census vectors



- Take the drawn parameters and apply these to the *X* matrix of characteristics in the census, and simulate the residuals
- This yields *R* simulated vectors in the census data
- From these vectors we get *R* poverty rates per area of interest, the standard deviation of these yields the standard errors

$$\tilde{Y}_{ch} = X\tilde{\beta}_{GLS} + \tilde{\eta}_c + \tilde{e}_{ch}$$

WORLD BANK GROUP

Practical issues

- Working with very big data (census)
- Working with limited computing power (32 bit, slow processing power, small RAM)
- Performance with sorting and large matrix operations in Mata
- Installation and updates

Practical issues – Working with very big data/limited computing power

- Powerful computers with sufficient RAM might open the large data. However, census are often large and contains many variables and Stata might not be able to open it.
- Operations on the large data take time, especially with sorting.
- We use Mata matrix binary data file for storing and retrieving vectors of data. The size of the Mata matrix file is often very large (8*N*K) but accessing vectors from Mata is fast.
- Questions for Stata:
  - Is there a way to compress the Mata matrix file? Mata matrix file definition?
  - Is there a way to read a matrix in Mata from plugin? How to combine Mata functions and plugins?

Practical issues – Performance with sorting

- Sorting in Mata is slow compared with other languages such as R. We created a plugin that reads the Mata matrix file and performs several operations including sorting.

Number of obs	Mata	R
1 million vector	64 seconds	12 seconds

- Other users also show the performance in sorting as well as other data manipulation functions

Source:

- https://www.statalist.org/forums/forum/general-stata-discussion/general/425307-comparison-with-r
- http://www.matthieugomez.com/pictures/1e7.png

Practical issues – Performance with large matrix operation in Mata

- When performing the operations with many simulations from the equation below, it is faster and more efficient with vector based calculations, one vector at a time. In addition, we are looking into OpenMP or Cilk for multithreaded parallel computing.
- Groups created based on the sorted hierarchical location ID are very useful when calculating simple statistics aggregated at those group levels.

$$\tilde{Y}_{ch} = X\tilde{\beta}_{GLS} + \tilde{\eta}_c + \tilde{e}_{ch}$$

- Running sum (reading the vector only once) is useful to get statistics for different groups defined by the hierarchical location ID.

Hierarchical location ID

Example of Hierarchical location ID (lid) with 12 digits = RDDZZMMMMMMM

Digit in Loc. ID	Number of digits to shift in ID (right to left)	Census	Survey
R	11	Rural/Urban (2)	
DD	9	Admn. Division (6)	
ZZ	7	Zila (64)	
MMMMMMM	0	Mauza (504)	Mauza (64)
		20 HH	10 HH

Example of running sum with groups (sorted by lid)

seq	lid	group1	group2	x	runningsum(x)	mean_group1	Formula	mean_group2
1	111	11	1	2	2			
2	112	11	1	3	5			
3	113	11	1	5	10	3.33	=10/3	
4	121	12	1	8	18			
5	122	12	1	5	23			
6	123	12	1	8	31	7	=(31-10)/(6-3)	
7	131	13	1	5	36			
8	132	13	1	2	38			
9	133	13	1	4	42	3.67	=(42-31)/(9-6)	4.67
10	211	21	2	4	46			
11	212	21	2	6	52			
12	213	21	2	2	54	4	=(54-42)/(12-9)	
13	223	22	2	7	61			
14	224	22	2	3	64			
15	225	22	2	8	72	6	=(72-54)/(15-12)	5

info_group1	Column1	Column2
11	1	3
12	4	6
13	7	9
21	10	12
22	13	15

info_group2	Column1	Column2
1	1	9
2	10	15

Hierarchical location ID

Example of Hierarchical location ID (lid) with 12 digits = RDDZZMMMMMMM

- Running sum is useful for some statistics such as poverty headcount, mean log deviation, and General Entropy indicators, or statistics as function of means or weighted means
- You read the vector once for those calculations. If you read up to the N*th* observations, you should be able to get all statistics (defined above) for different aggregated levels from those first N observations. [Another faster way to collapse in Mata]
- However, it is not possible when statistics/indicators need the whole vector such as Gini (requires sorting) or decile distribution.
- We wrote a plugin that reads the Mata matrix binary file and calculates those indicators in C.

Using poverty maps to improve the efficiency of transfers

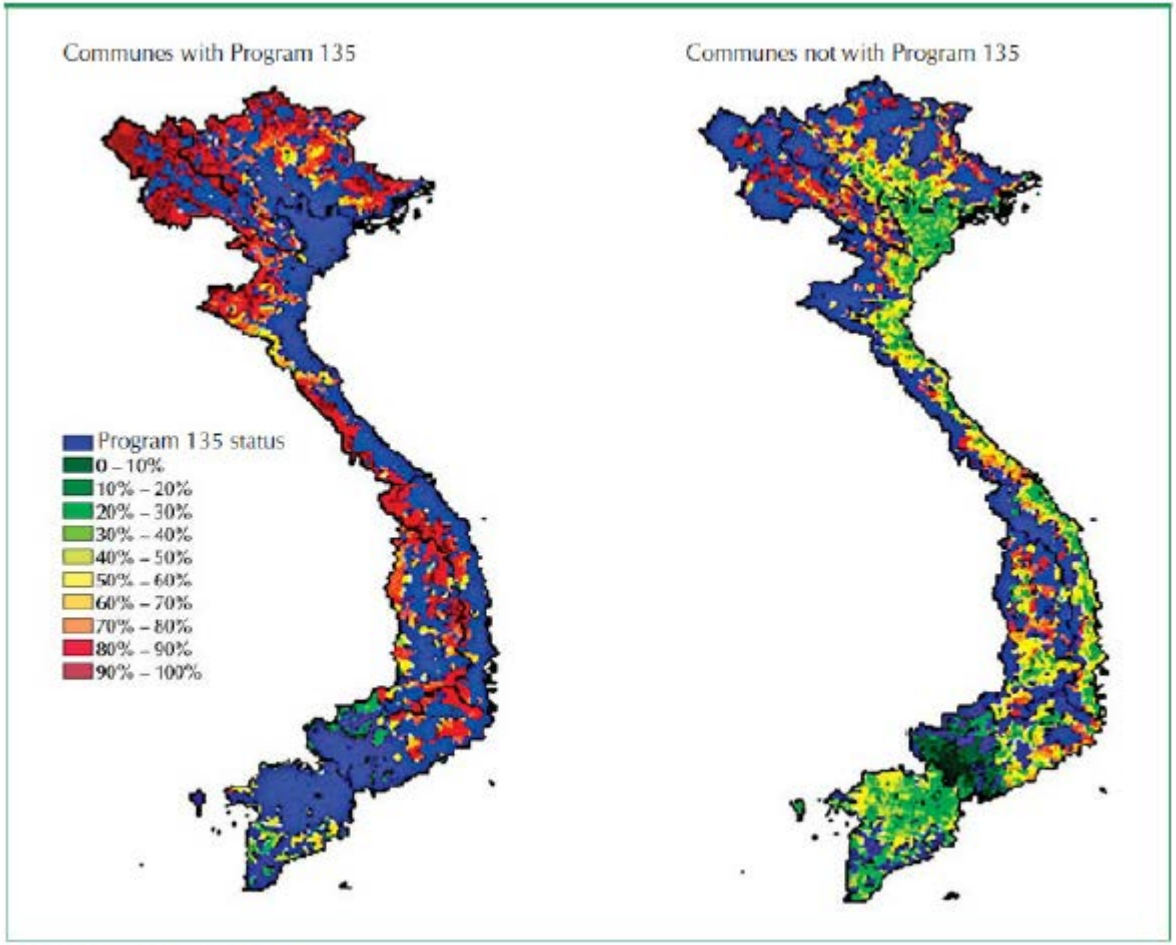
	Ecuador (Rural)	Madagascar (Urban & Rural)	Cambodia (Urban & Rural)
Uniform transfer	100	100	100
Optimal targeting (1st admin level)	76.0	60.7	54.5
Optimal targeting (2nd admin level)	66.7	46.4	41.4
Optimal targeting (3rd admin level)	58.4	37.6	30.8

Source: Elbers et al., Journal of Development Economics, 2007

Using poverty maps for Program monitoring

Program 135 in Vietnam targets using multiple criteria.

Comparing program map with poverty map showed good coverage, but also revealed unexpected cases of exclusion.

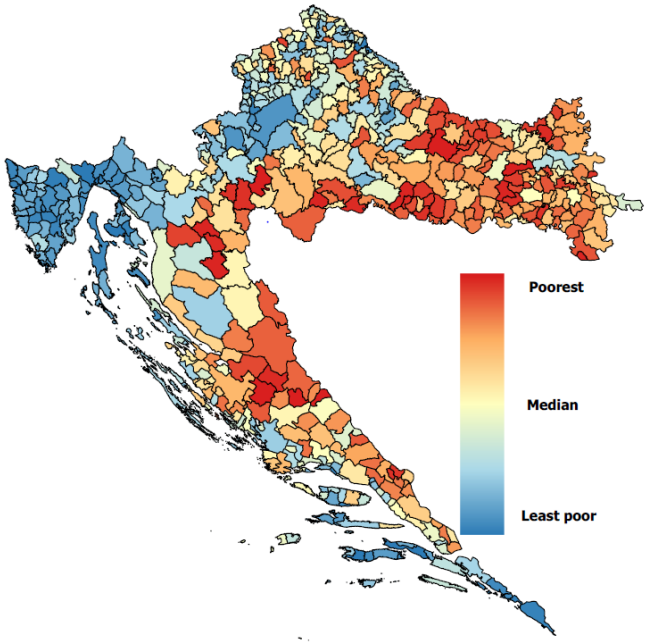


Sources: Nguyen et al. 2004b.
Note: On left map, communes without Program 135 are shown in blue. On right map, communes with Program 15 are in blue.

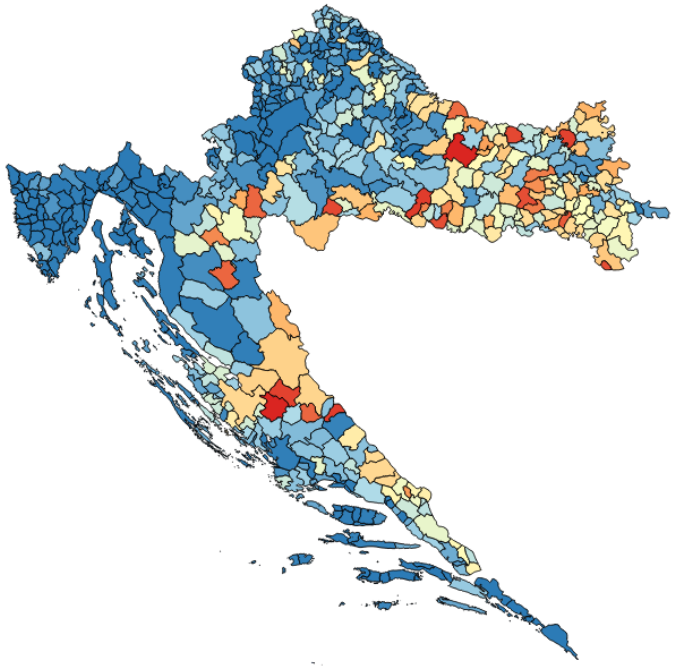
WORLD BANK GROUP

Ex-post evaluation – Informing the efficiency of transfers

Poverty ranking before transfers



Ranking of post-transfer values compared to values before transfers values



WORLD BANK GROUP

Thank you!

Minh Cong Nguyen mnguyen3@worldbank.org
Paul Andres Corral Rodas pcorralrodas@worldbank.org
Joao Pedro Wagner De Azevedo jazevedo@worldbank.org
Qinghua Zhao qzhao@worldbank.org

WORLD BANK GROUP