SMALL AREA ESTIMATION OF POVERTY IN THE EU: LESSONS LEARNED

JOÃO PEDRO AZEVEDO

LEAD ECONOMIST AND GLOBAL LEAD ON WELFARE MONITORING AND STATISTICAL CAPACITY
POVERTY AND EQUITY GLOBAL PRACTICE
JAZEVEDO@WORLDBANK.ORG @JPAZVD

WORLD BANK GROUP

Workshop on Small Area Methods and living conditions indicators in European poverty studies in the era of data deluge and Big data

Pisa, May8th 2018

Why is SAE more important than ever?

- 2020 Census round is coming
- More and more spatial and geo-coded data is available
- Demand for spatial distribution of variables such unemployment, welfare, consumption/expenditure are on the rise
- Administrative records are also biased
- Fiscally constrained governments demand better information to monitor their expenditures

Main Features of the Project

Construct poverty maps for all EU Member States (NUTS 3 or lower)

- World Bank responsible for ten new Member States
- Consortium of Nordic research centers covering the other 17 Member States

Two phases to the project:

- Pilot in Denmark and Slovenia to compare poverty mapping methodologies
- peer reviewed by Steering Committee that includes Eurostat and other European technical experts
- Produce maps for remaining member states using agreed methodology

Within member states, the main partners are national statistical institutes (NSIs).

- Working with data before it is sent to Eurostat ➔ getting national buy-in, working collaboratively on-site in NSIs, strengthening NSI capacity
- Full census microdata not available in time in most countries ➔ using aggregate data in some countries, and possible refinements as census microdata becomes available.
- Cultivating interest of line ministries, especially Ministries of Labor and Ministries of Regional Development.

European Commission / World Bank Poverty Mapping Project

Objective: identify the small areas (e.g., municipalities) most likely to have the highest risk of poverty rates. That is, show the regional disparities within EU Member States.

Purposes:

- Inform European Commission negotiations with Member States for 2014-2020 budget cycle, using high-resolution poverty statistics
- Inform national and sub-national policies and programs

Collaboration among EC (DG Employment, DG Regional Policy, Eurostat), World Bank, and the national authorities in Member States

The Challenge: Obtaining Poverty Indicators for Small Sub-national Areas

Household surveys such as EU-SILC are the main source of indicators of living conditions, poverty, and social exclusion.

- Detailed information on multiple indicators
- Sample sizes are too small to be representative for disaggregated sub-national units.

Population censuses

- 100% coverage permits assessment for small areas
- Typically do not have much information on the usual poverty and social exclusion indicators

The Solution: Small Area Estimation (SAE) applied for poverty Combines the Census and the Survey

Survey:

- Welfare measure
- Not representative at lower level

+

Census:

- Full coverage
- No monetary welfare measure

Administrative Records

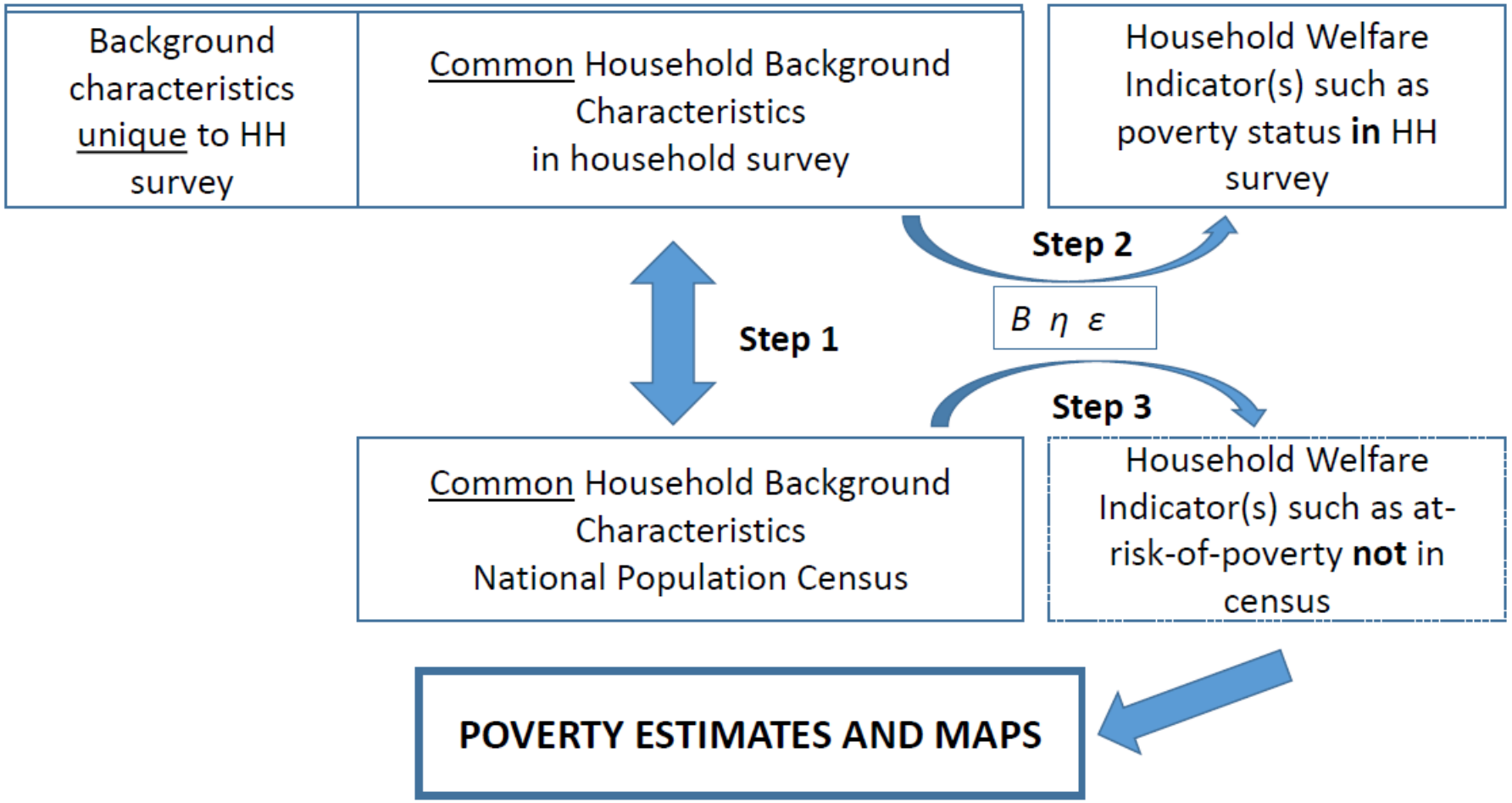
Tax / Revenues / Population

→

Poverty Mapping:

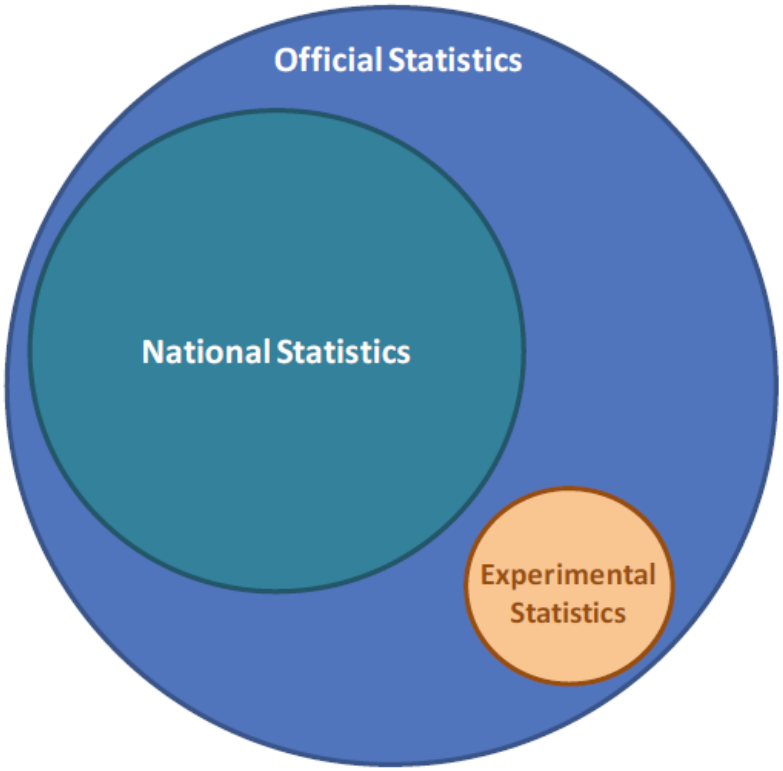
- Geographically disaggregated estimates of: poverty, number of poor, average income/consumption, inequality
- Simulating a measure of welfare from household survey into census, using statistical methods
- Considerable data requirements

Overview of the workflow



WORLD BANK GROUP

National Systems of Statistics that brings together Official, National and Experimental statistics



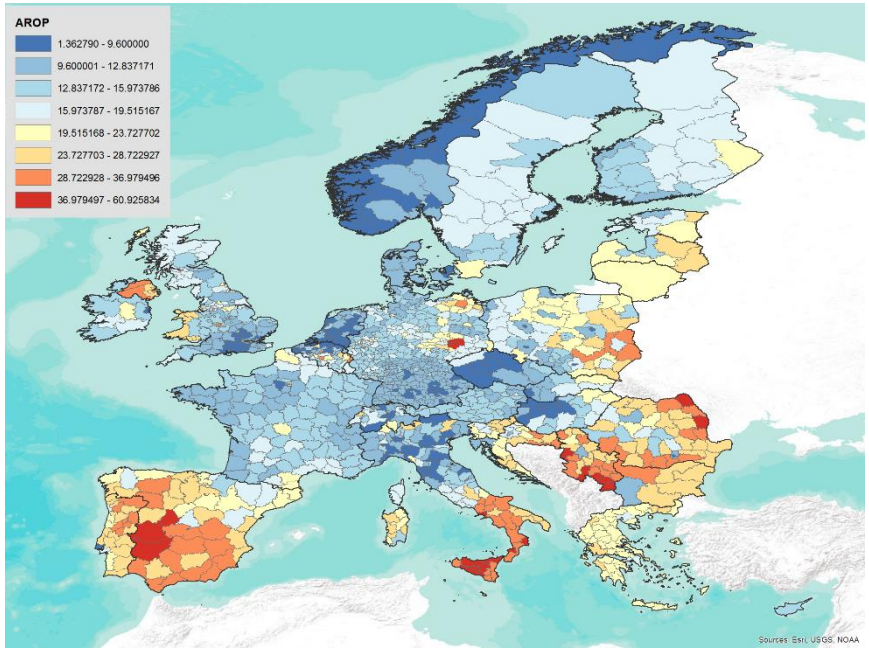
Source: “Assessment and Designation of Experimental Statistics”. UK Government Statistical Service

Creating the space for and recognizing the value of innovation in any Statistical System

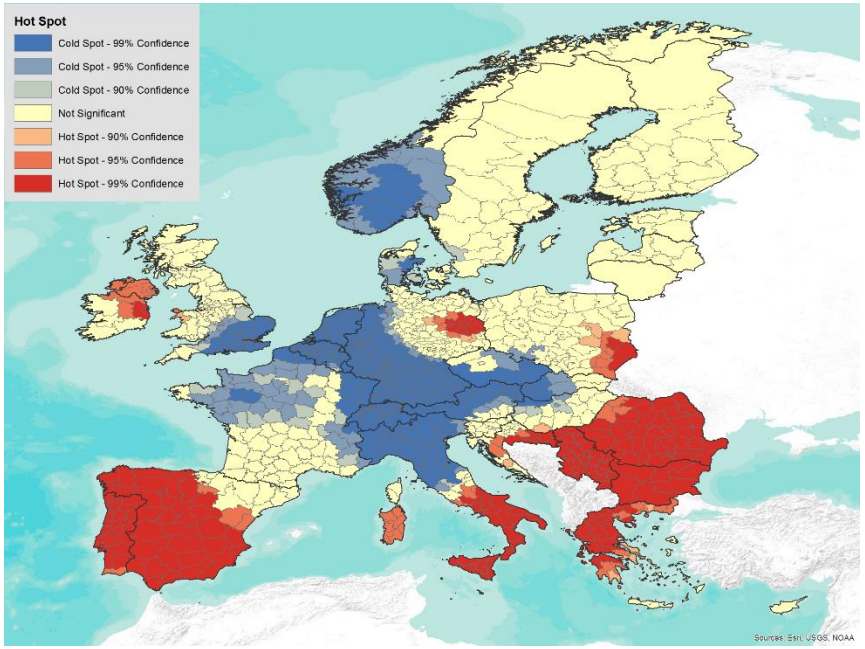
WORLD BANK GROUP

Understanding the spatial distribution of poverty is extremely important, and small area estimation methods are now being used as both official and experimental statistics

EU At-Risk-of-Poverty Maps



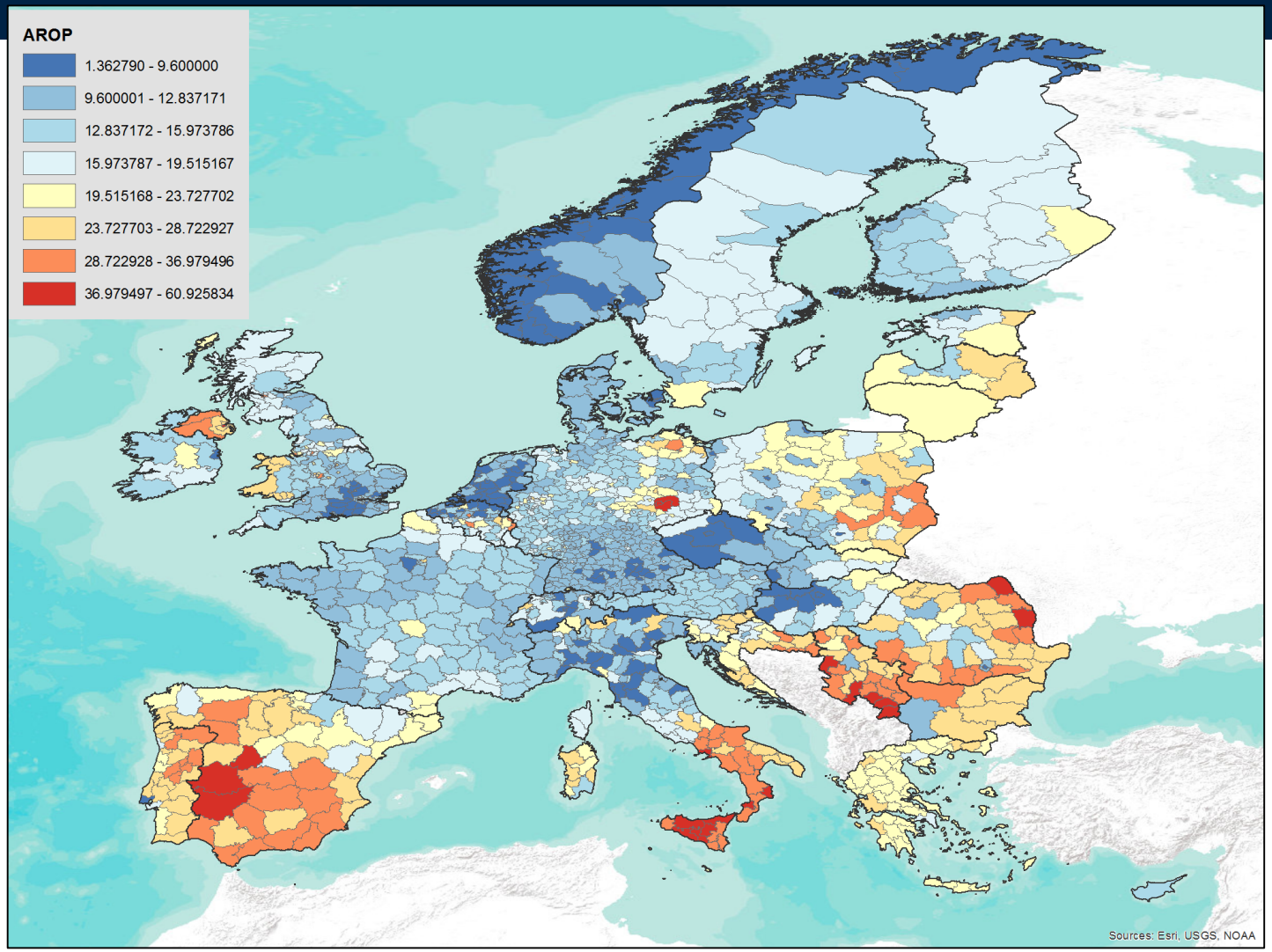
Hot and Cold Spots of the places At-Risk-of-Poverty



Note: EU Poverty Map 2011 produced by National Official of Statistics in AT, BE, BG, CH, CY, CZ, DE, DK, EE, EL, FI, FR, HR, HU, IE, IT, LT, LU, LV, MT, NL, NO, PL, PT, RO, SE, SI, SK and UK in collaboration with DGREGIO/TiPSE/World Bank.

WORLD BANK GROUP

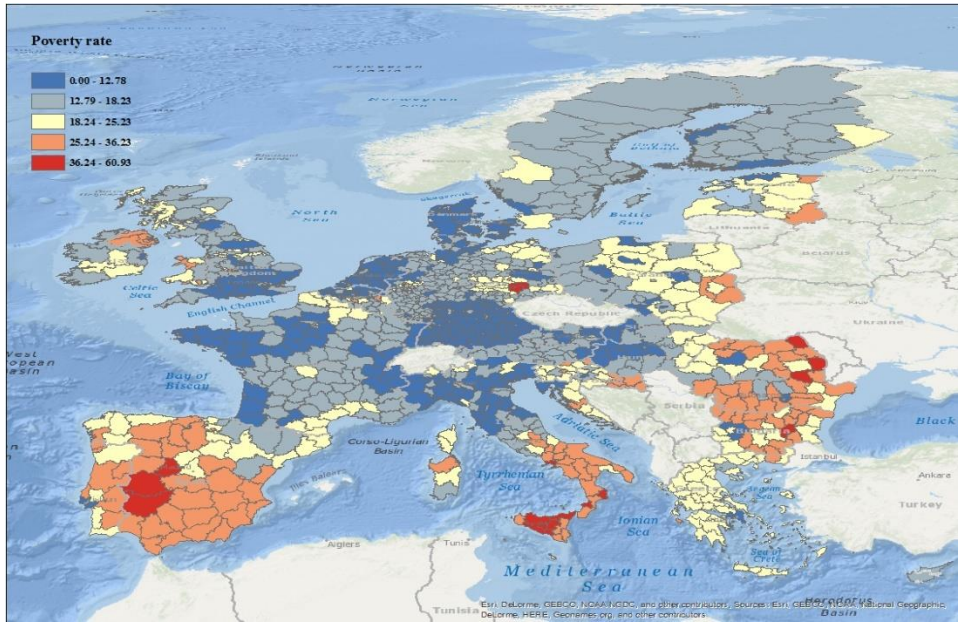
And special insights can be gained when we look at the big picture…



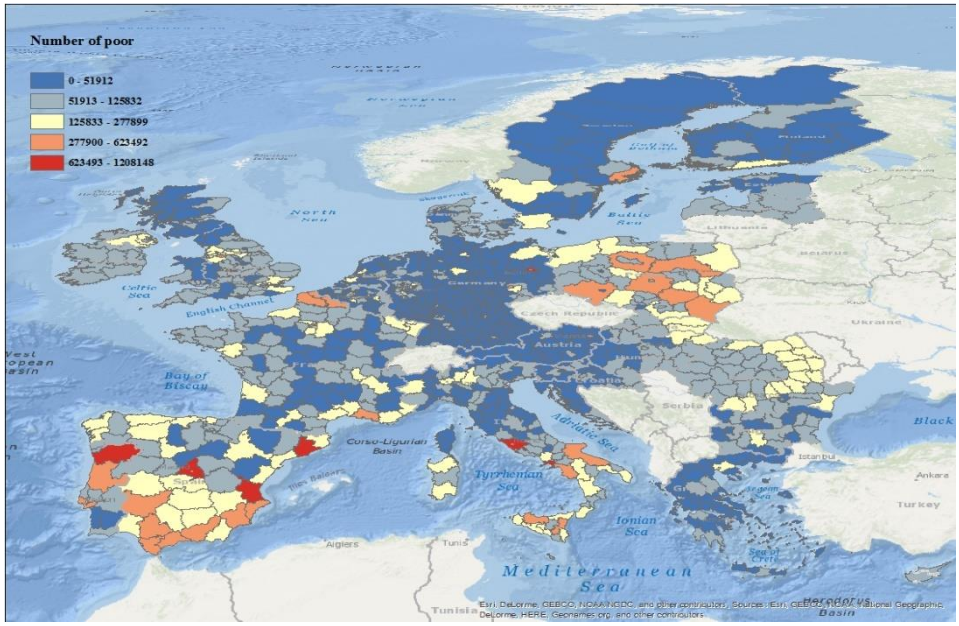
WORLD BANK GROUP

Such as the poverty rate and the density of poverty are not always aligned

EU At-Risk-of-Poverty at NUTS3



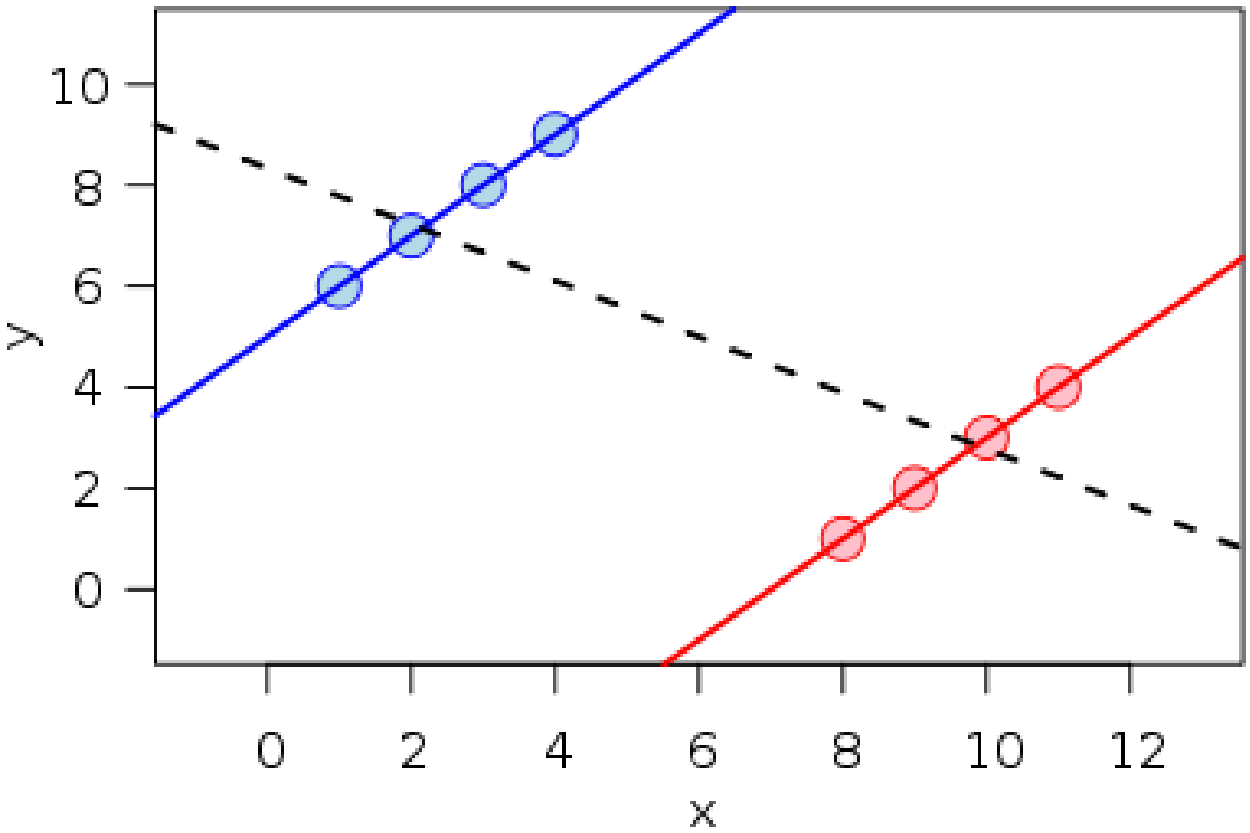
EU Number of Population At-Risk-of-Poverty at NUTS3



WORLD BANK GROUP

And the level of spatial analysis matters…

A trend appears in two different groups of data but disappears or reverses when these groups are combined

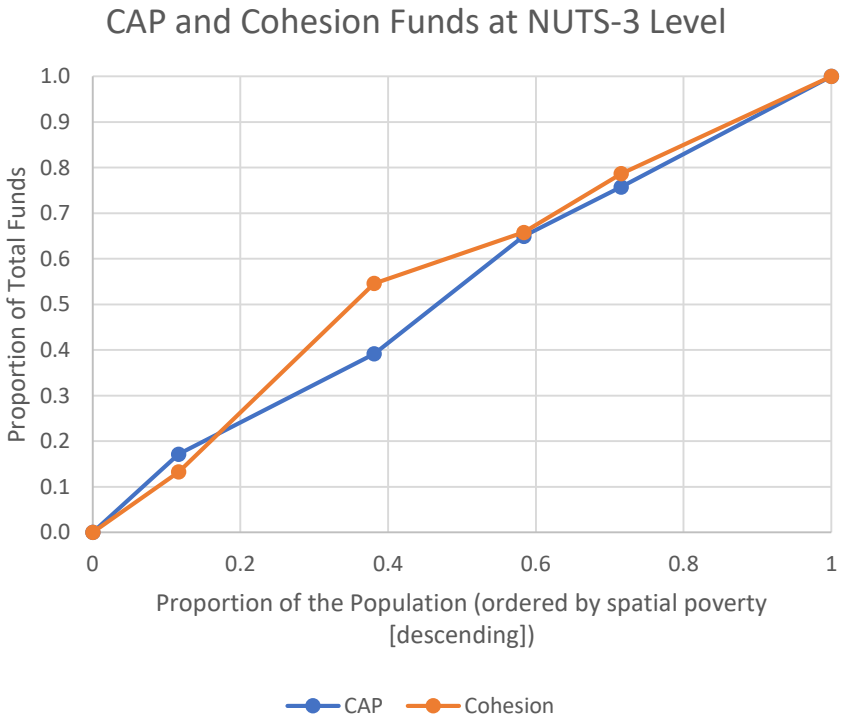


Simpson's paradox for quantitative data: a positive trend (red line, blue line) appears for two separate groups, whereas a negative trend (dashed line) appears when the groups are combined.

WORLD BANK GROUP

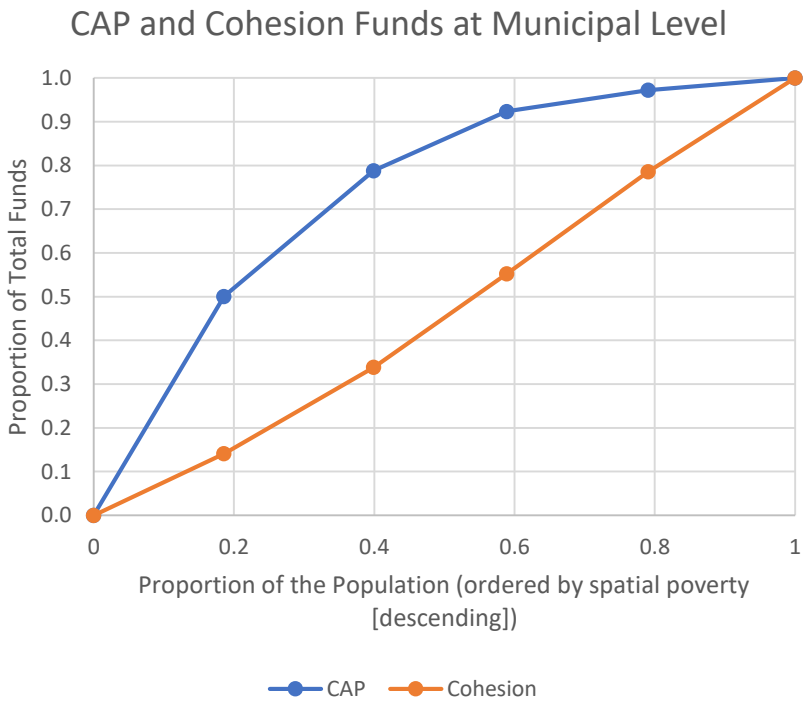
…and it can impact our understanding of facts.

EU cohesion funds at the NUTS 3 level seem to be aligned with the spatial distribution of poverty



Note: Graph makes use of BOOST data for a selected member state

A more granular view suggests that funds are allocated quite differently within each NUTS3

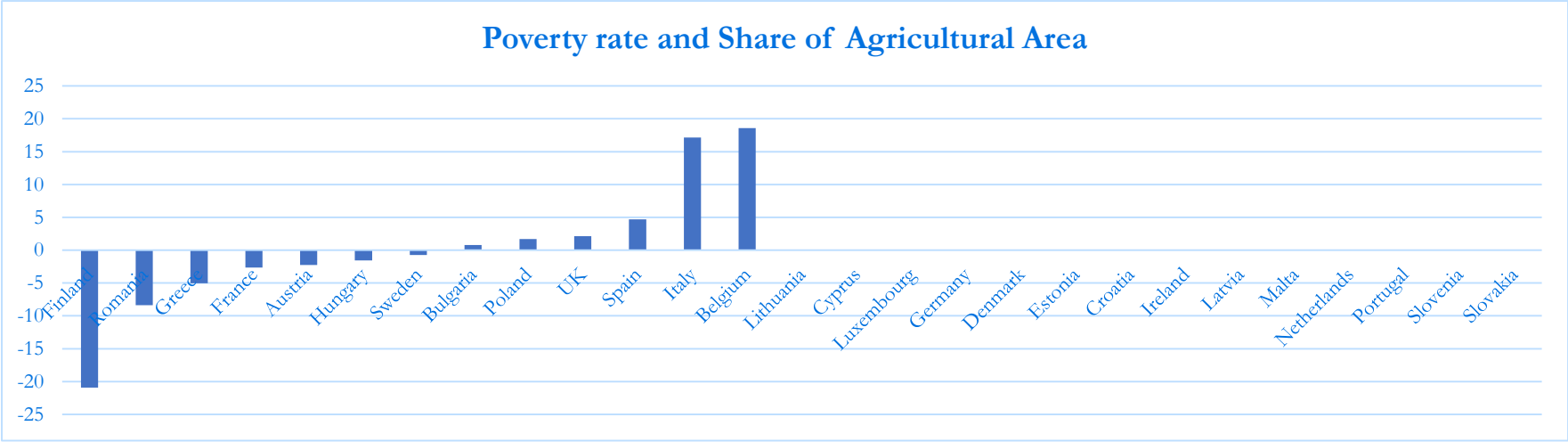


Note: Graph makes use of BOOST data for a selected member state Locations are ranked in order of poverty level. The first quintile corresponds to 20 percent of the population who reside in the poorest locations

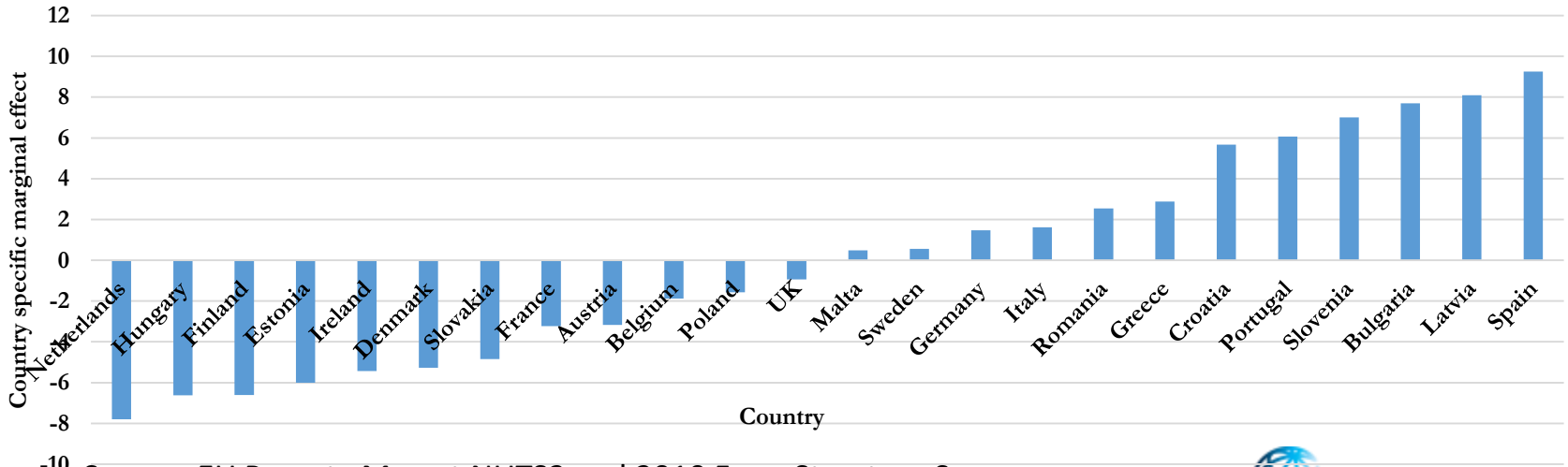
WORLD BANK GROUP

As the level of spatial disaggregation matters for our understanding of relationships at the country level.

the analysis of the relationship between share of agriculture area and poverty is largely different, if we are working at SILC level of Statistical Representation or at the NUTS3 level



Source: EUSILC at NUTS1 and NUTS3 and 2010 Farm Structure Survey



Source: EU Poverty Map at NUTS3 and 2010 Farm Structure Survey

WORLD BANK GROUP

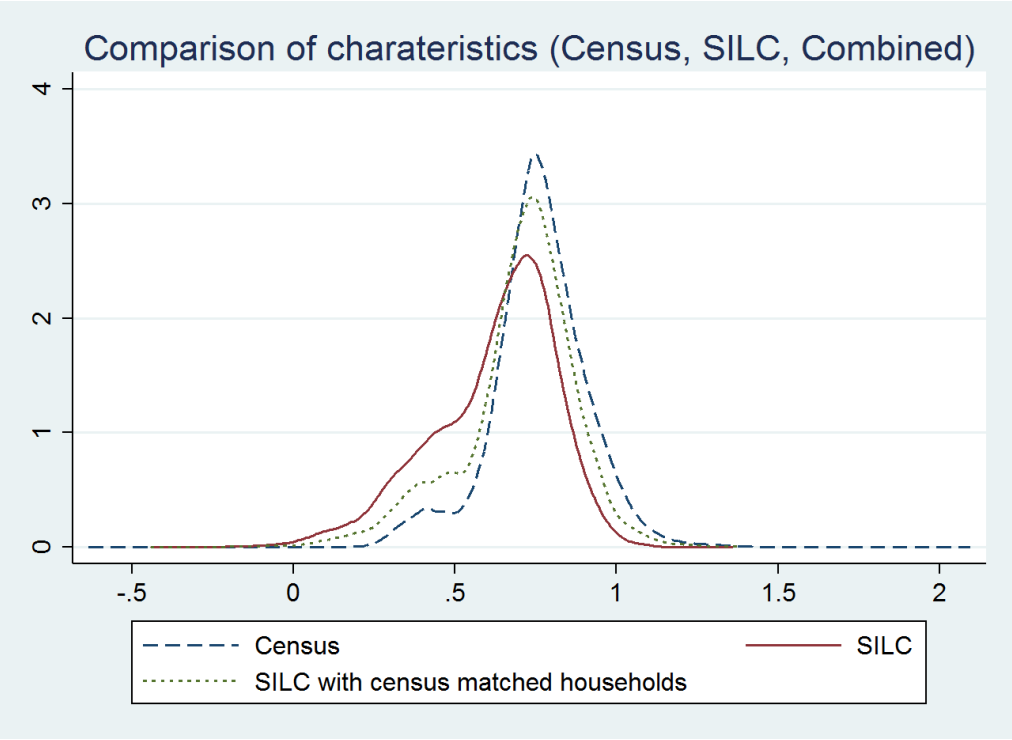
Validation: And the quality of small area estimations can be greatly improved as responses across different statistical operations are linked

For the **2020 round of the population Census** it is critical to enable the link of responses from surveys with Census data.

Bulgaria and Latvia are using this information to improve the quality of their small area estimation of poverty with great success.

This possibility can applied on a number of **other indicators**

This process can be of value even in places where tax records can be link with Census respondents, since in some countries surveys can better capture **informal sources income**.



Source: NSI Bulgaria / World Bank Poverty Map

WORLD BANK GROUP

Validation
Household size (Census)

Table 1: Bulgaria Census Figures (Share of households)

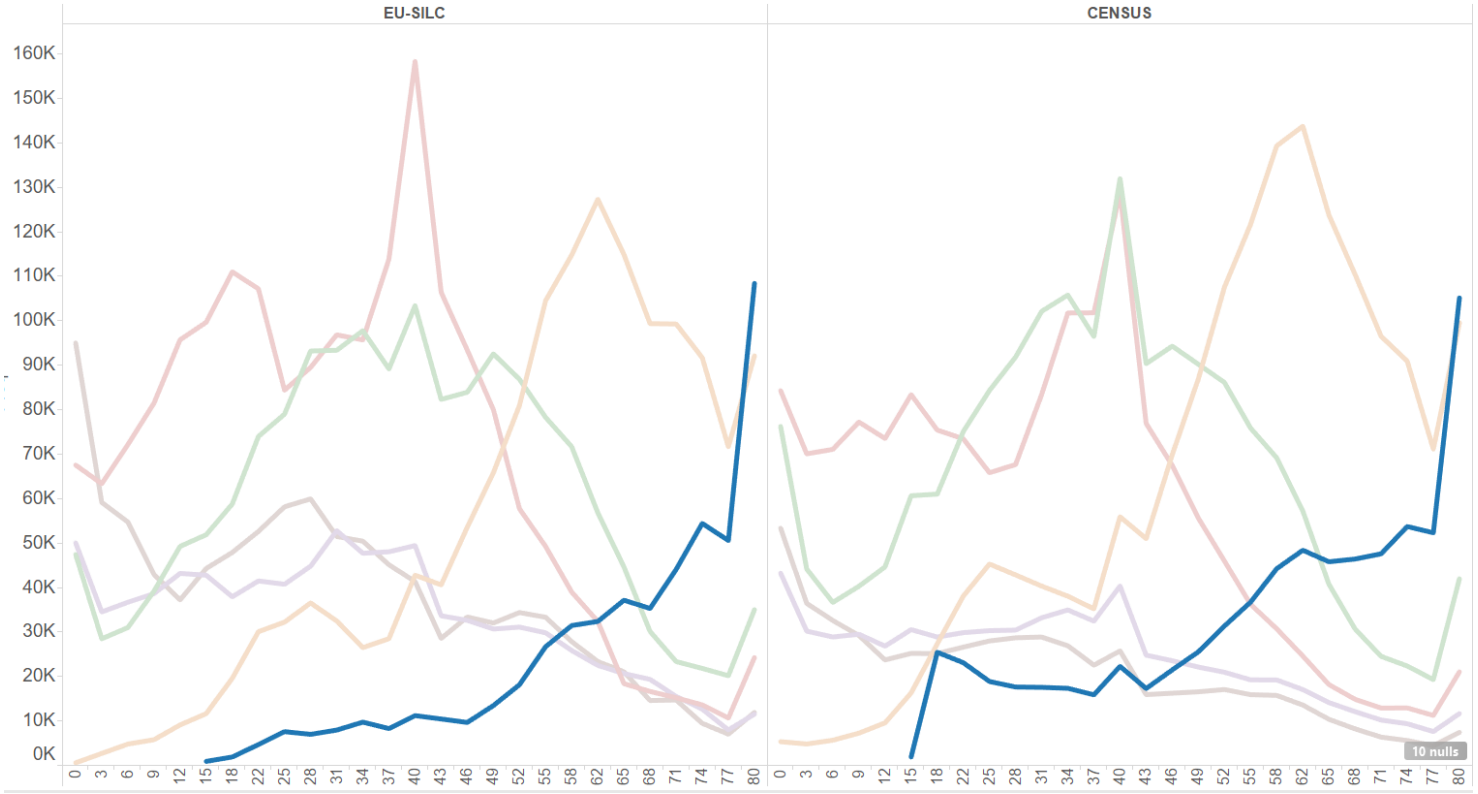
	1946	1965	1975	1985	1992	2001	2011
				(%)			
One member	10.4	17	16.8	18.2	19.7	22.7	30.8
Two members	13.6	20.7	23.3	26.7	28	28.4	28.4
Three members	19.2	21.6	21	20.3	20.4	21.6	20.18
Four members	21.9	21.1	21.1	21.5	20.4	18	13.4
Five or more	34.8	19.5	17.7	13.3	11.5	9.4	7.26
Total	**100**	**100**	**100**	**100**	**100**	**100**	**100**

Validation
Household size (SILC)

Table 2: EU-SILC household composition (share of households)

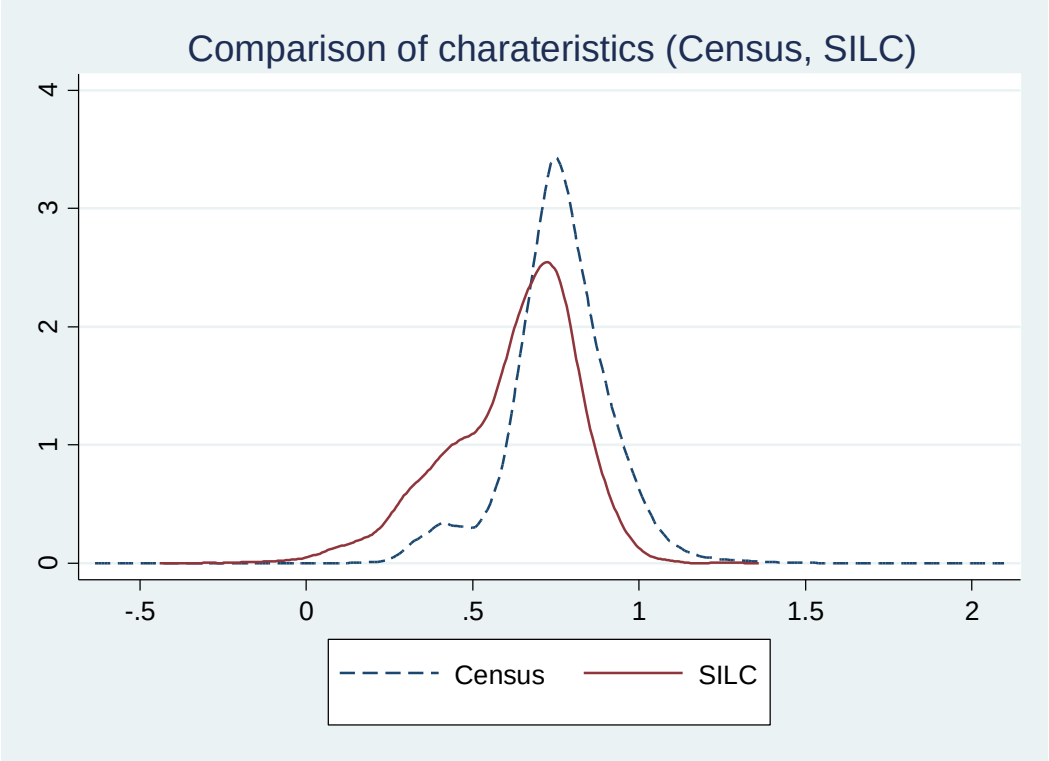
	2007	2008	2009	2010	2011	2012	2013	2014
					(%)			
One member	20.5	18.4	19.1	19.5	19.9	21.5	22.8	24.3
Two members	25.9	27.2	27.4	27.8	26.7	28.7	28.8	28.9
Three members	21.5	23	21.7	20.3	20.4	20.4	21.7	21.3
Four members	17.3	17.1	18.6	19.3	19.2	18.3	17.3	16.8
Five members	14.8	14.3	13.2	13.2	13.8	11.1	9.4	8.7
Total	100	100	100	100	100	100	100	100

Validation
Household size (SILC)



WORLD BANK GROUP

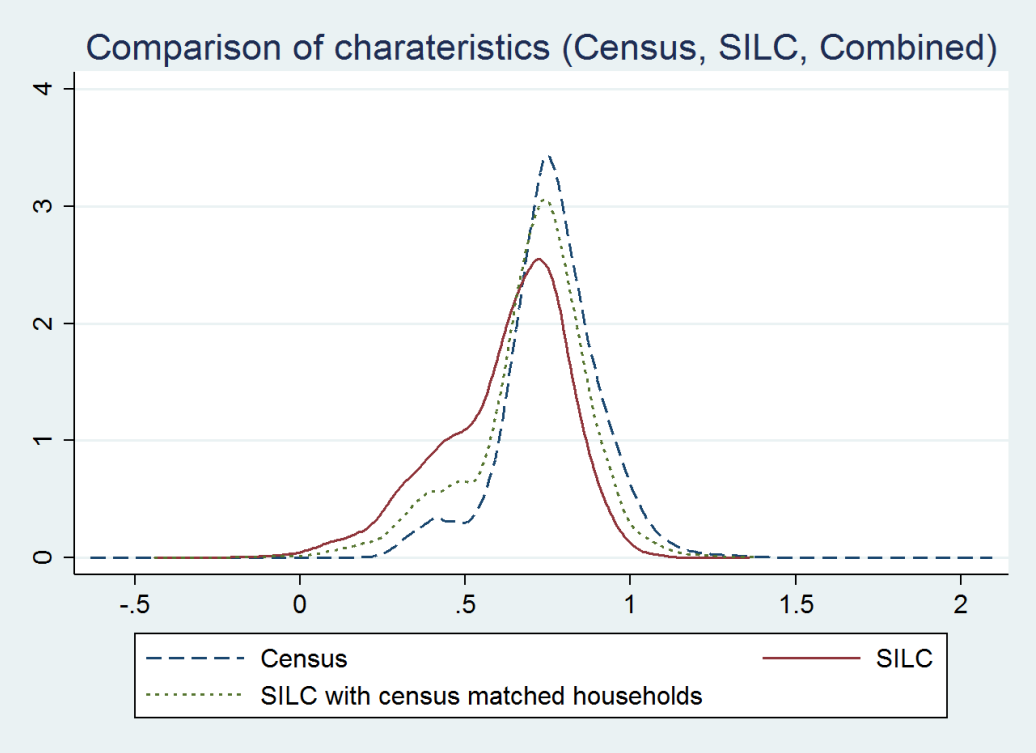
Validation
How different are they?



WORLD BANK GROUP

Validation
How to solve it?

- It is possible to link SILC households to the Census, however only about two thirds of the SILC can be matched



WORLD BANK GROUP

And External Validation of the Results is Critical,

Latvia Poverty Map Validation

	SILC		Povmap With Empirical Best	
	AROP	Std. Err.	AROP	Std. Err.
Riga	12.9%	0.007	14.2%	0.006
Pieriga	15.5%	0.012	16.2%	0.025
Kurzeme	19.8%	0.014	21.3%	0.028
Zemgale	22.2%	0.016	23.8%	0.032
Vidzeme	27.8%	0.020	25.5%	0.039
Latgale	28.6%	0.017	30.9%	0.038
National	19.2%	0.005	20.28%	0.024

Note: Direct estimates come from the 2011 SILC, Small area estimates are the outcome of 100 simulated Census' incomes.

…and in our experience encouraging

Bulgaria Poverty Map Validation

	Direct estimates			Small Area Estimates		
	Estimate	95% CI		Estimate	95% CI	
Severozapaden	29.9%	25.1%	34.7%	31.8%	29.2%	34.4%
Severen Tsentralen	24.4%	19.7%	29.0%	26.8%	24.4%	29.3%
Severoiztochen	25.0%	20.9%	29.1%	25.9%	23.2%	28.6%
Yugoiztochen	28.9%	20.5%	37.3%	26.8%	24.3%	29.3%
Yugozapaden	11.6%	9.3%	14.0%	12.7%	11.0%	14.4%
Yuzhen Tsentralen	27.3%	21.9%	32.7%	26.0%	24.1%	28.0%
Bulgaria	**22.7%**	**20.6%**	**24.7%**	**23.1%**	**21.9%**	**24.3%**

Note: Direct estimates come from the 2011 SILC, Small area estimates are the outcome of 100 simulated Census' incomes. Threshold at 3,236 Lev.

And hold for both consumption and income based measures of poverty as well as different sub groups.

Croatia Poverty Map Validation

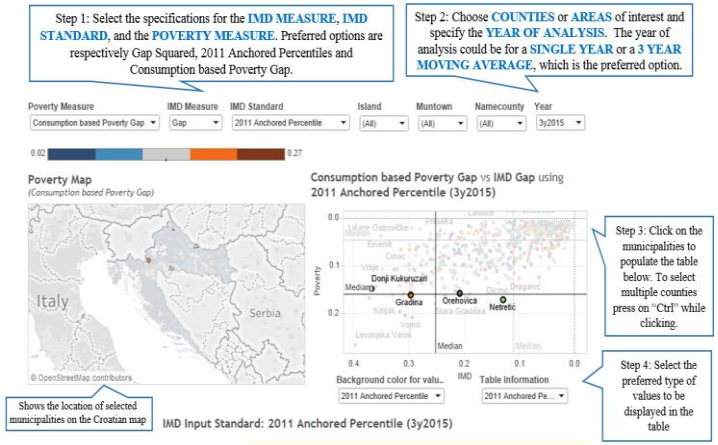
	Direct estimates		SAE	
	Income	Consumption	Income	Consumption
Total	20.4	16.3	19.2	17.1
Among children	23.3	41.3	20.3	38.7
Among working age adults	19.3	13	17.1	13.7
Among the elderly	25.6	10.9	26.3	11.8
Among those living alone	35.8	8.7	31.5	6
Among those living with another person	22.3	6.4	19.9	5.3
Among those living with two people	15.3	9.1	14.6	8.1
Among those living with 3 or more people	19	24.2	18.7	26.5
Among those who work	7.6	9.9	6.1	11.5

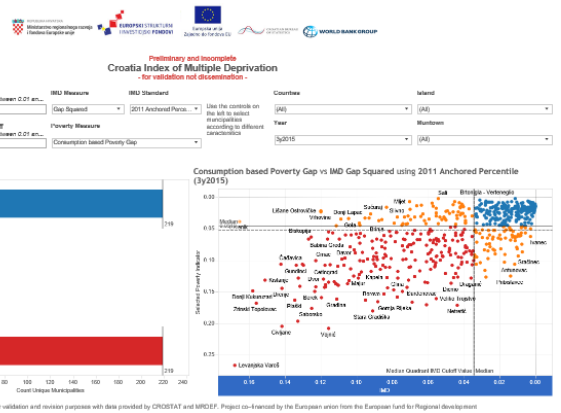
Source: Croatian HBS 2011 (threshold at 23,919 HRK), and Croatian SILC 2012 (threshold at 24,000 HRK)

Note: Croatian Bureau of Statistics and MRDEUF

The possibility of linking administrative records across the territory can also be of extreme value to improve the alignment of policies to tackle poverty and deprivation, and NSOs can play a critical role coordinating and integrating those different sources of information

By combining the information for the **poverty map** with geocoded **administrative records** from line ministries it is possible to develop a system of indicators to **identify the bottlenecks** and **monitor the impact** of EU regional development funds.





Domain	Subdomain	Indlabel	Donji Kukuruzari	Orehovica	Gradina	Netretić
Economic	Economic development	Net income of the population	0.94	0.96	0.93	0.34
		Number of active business ent	0.84	0.60	0.78	0.44
		Number of active crafts per ca	0.98	0.77	0.87	0.63
		Number of registered personal	0.94	0.80	0.78	0.12
		Share of employed in agricultu	0.81	0.83	0.97	0.56
	Fiscal capacity	Average taxable income per c	0.96	0.94	0.93	0.36
		Budget revenues (w/o grants,	0.98	0.75	0.68	0.64
		Share of taxpayers in populati	0.51	0.76	0.60	0.00
		Total budget expenditure (incl	0.67	0.79	0.73	0.73
	Labor Market	Employment rate	0.98	0.87	0.91	0.31
		Participation rate	0.28	0.88	0.57	0.21
		Pension system dependency r	0.98	0.23	0.79	0.80
		Unemployment rate	1.00	0.79	0.97	0.55
Physical	Physical infrastructure	Share of HHs with Internet co	0.90	0.66	0.84	0.83
		Share of HHs with access to p	0.99	0.99	0.99	0.99
		Share of HHs with access to p	0.74	0.55	0.60	0.43
		Share of HHs without central	0.98	0.08	0.78	0.27
	Social services	Distance to primary health cen	0.95	0.42	0.55	0.27
		Enrollment rate in kindergarte	0.99	0.38	0.99	0.23
		Transparency of local govern	0.33	0.47	0.47	0.47
Social	Demography	Dependency ratio	0.81	0.86	0.68	0.77
		Mortality rate	0.93	0.40	0.79	0.96
		Population change (year-on-y	0.97	0.19	0.94	0.88
		Population density	0.92	0.18	0.72	0.78
	Health and education	Proportion of student failing M	0.82	0.77	0.95	0.95
		Share of people with secondar	0.92	0.96	0.93	0.73
		Share of persons using the as	0.78	0.43	0.92	0.14
	Social protection	Child allowance benefit per ca	0.87	0.99	0.81	0.15
		GMB per capita per month	0.99	0.97	0.90	0.56
		Share of GMB beneficiaries in	0.95	0.75	0.74	0.62

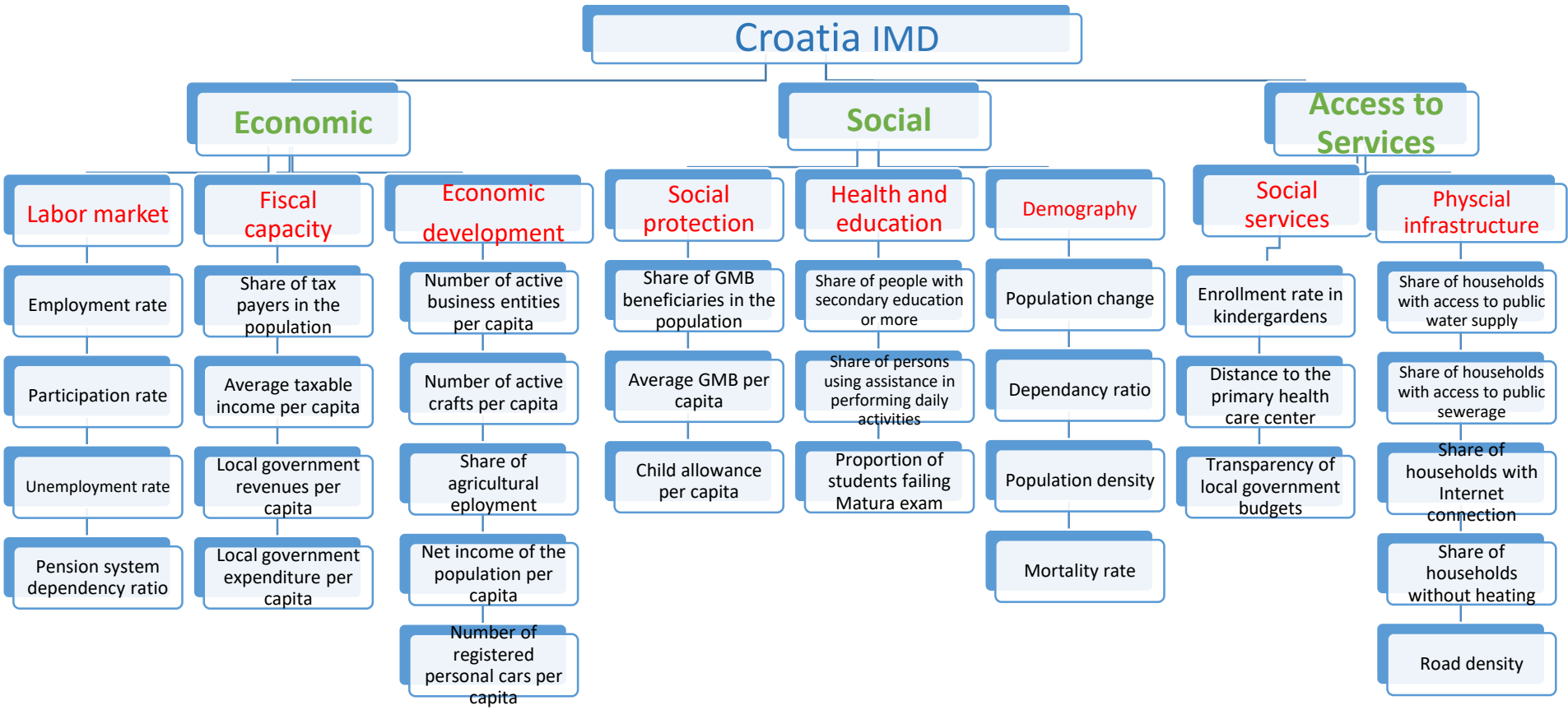
Displays the percentile ranking of each selected municipality and colors each cell based on the performance. Green indicates higher ranking while red signifies lower ranking.

WORLD BANK GROUP

Increasing the policy relevance of Poverty Maps

Croatian Index of Multiple Deprivation

IMD Dashboards



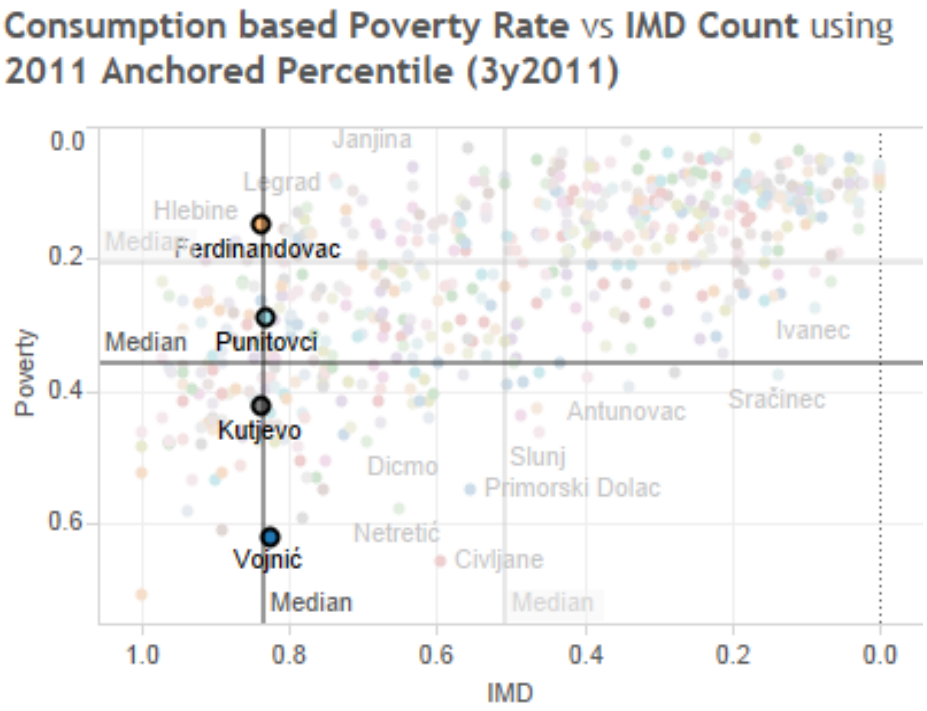
WORLD BANK GROUP

Increasing the policy relevance of Poverty Maps

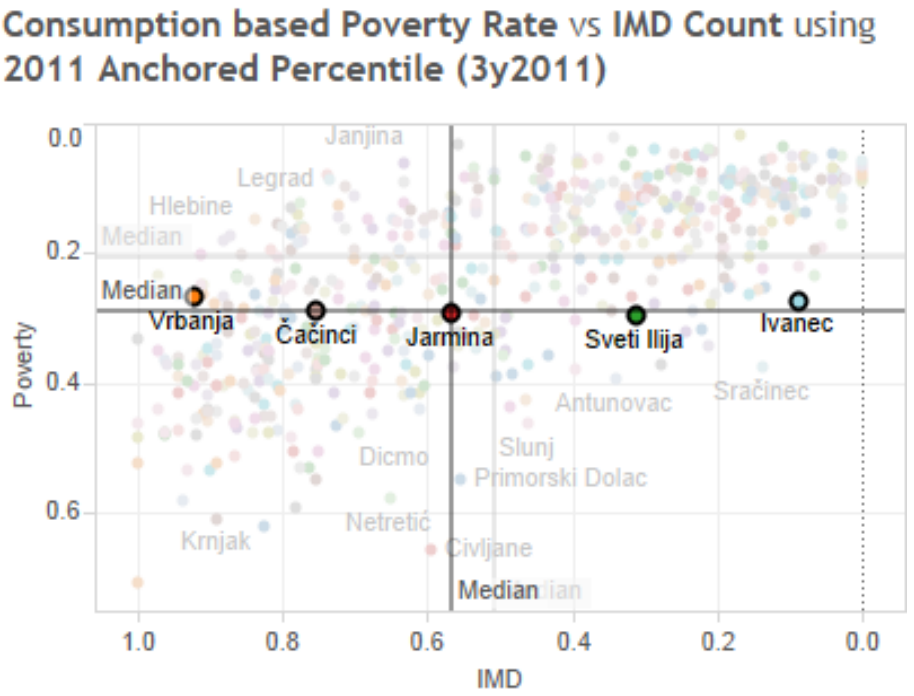
Croatian Index of Multiple Deprivation

IMD Dashboards

Municipalities at the same IMD can have very different monetary poverty levels….



Municipalities at the Poverty level can have very different IMD.



WORLD BANK GROUP

Potential further policy applications

Presentation matters: Much more than a pretty picture (but often not more than a busy picture)

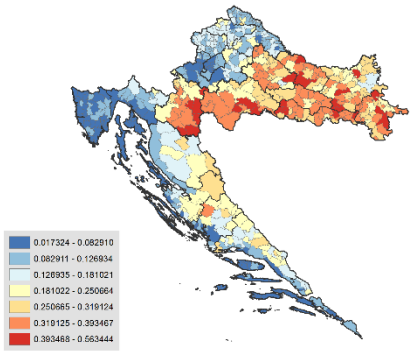
To be policy relevant poverty maps need to be presented in a simple and meaningful manner.

Technology has simplified the production of dashboards, and they are promising.

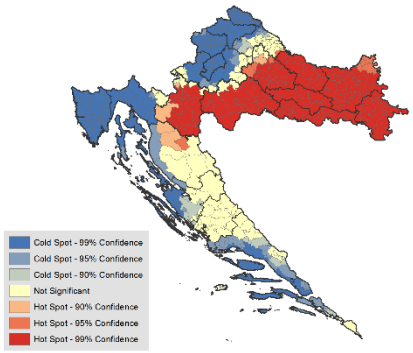
Local Indicators of Spatial Association (LISA), can be quite powerful as the bring statistical rigor (when properly done) and simplicity.

Much work still needed in this front.

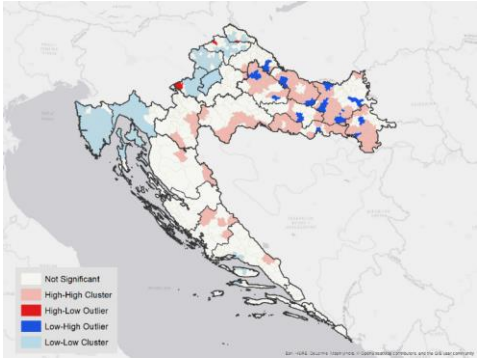
Croatia Poverty Rate (HBS)



Croatia Poverty Rate Hot Spot Analysis



***Red indicates a cluster of high values (Hot spot)**
Blue indicates a cluster of low values (Cold spot)



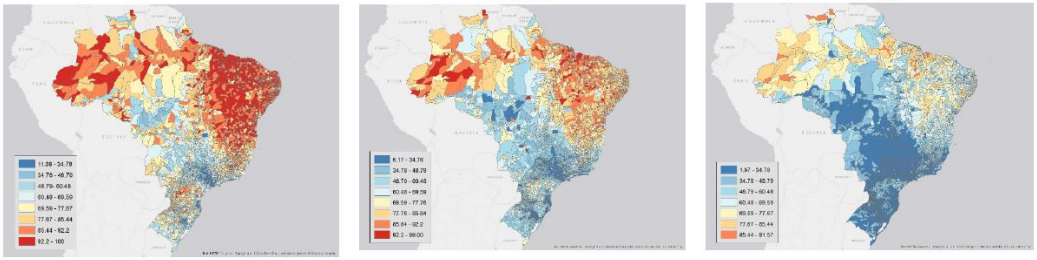
WORLD BANK GROUP

And expectations need to be managed….
How often poverty maps need to be updated?

And although the poverty levels might change over time the spatial distribution of poverty and where the poor lives tends to be quite persistent

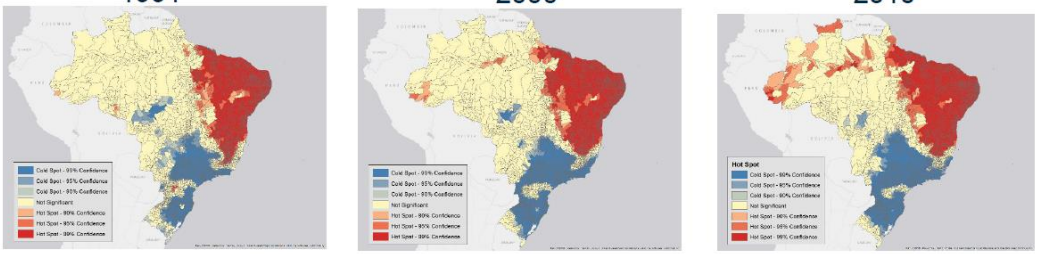
Poverty Maps

1991 2000 2010



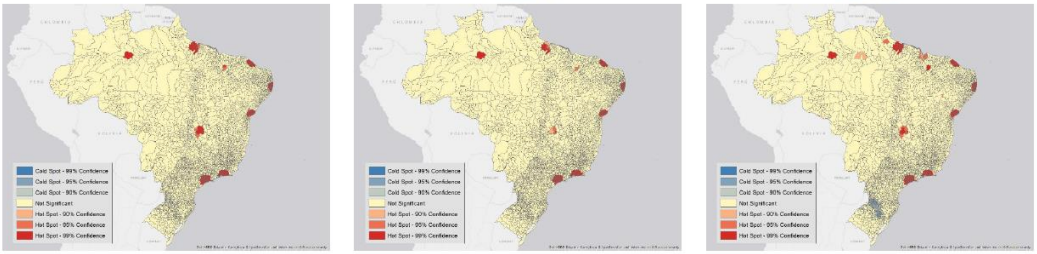
Hot and Cold Spots of the Poverty Rate

1991 2000 2010



Hot and Cold Spots of the Poor

1991 2000 2010

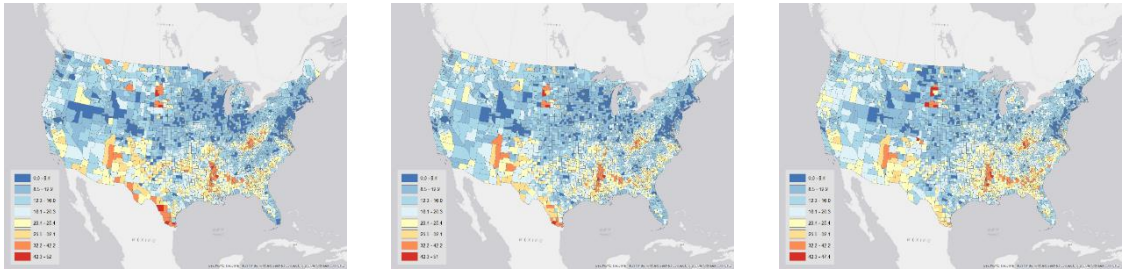


Note: Hot and Cold spots produced by author using poverty map from IBGE/IPEA/UNDP.

Regardless if the country has experience nor not a significant changes in poverty during the period

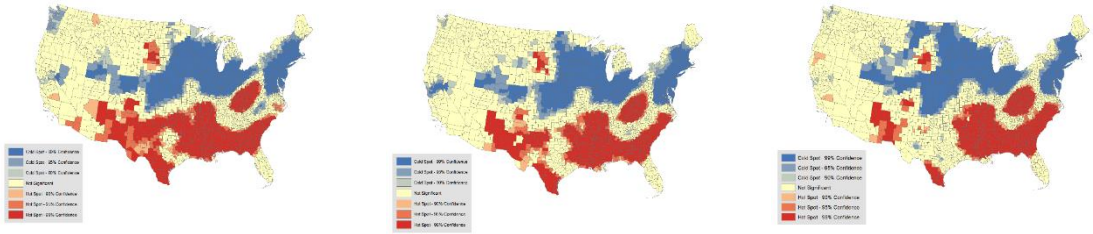
Poverty Maps

1995 2005 2015



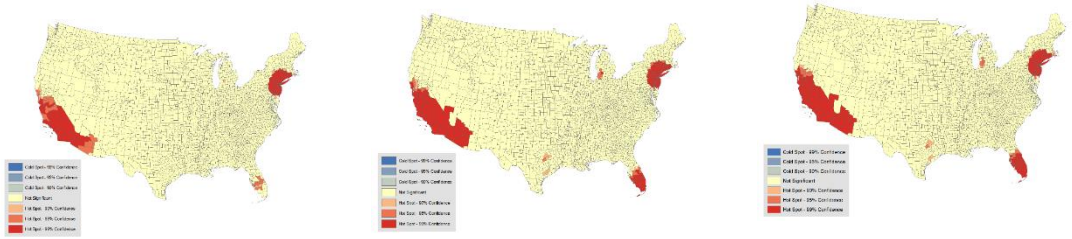
Hot and Cold Spots of the Poverty Rate

1995 2005 2015



Hot and Cold Spots of the Poor

1995 2005 2015

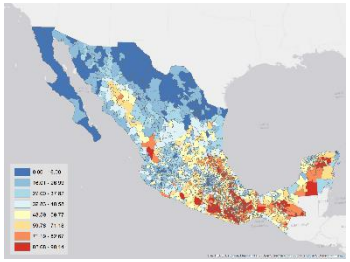


Note: Hot and Cold spots produced by author using poverty data produced using the American Community Survey, US Census Bureau.

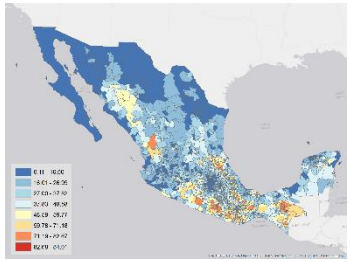
Suggesting that the shelf life of poverty maps is substantially higher than what is often perceived by policy makers, if and when the questions is what are the poor regions of the countries and where the poor lives

Poverty Maps

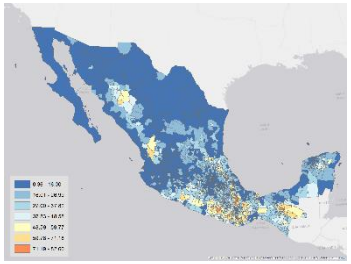
2000



2005

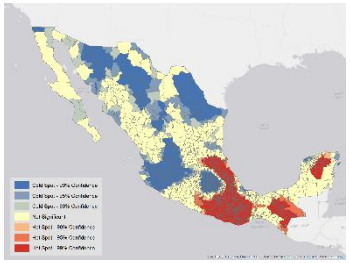


2010

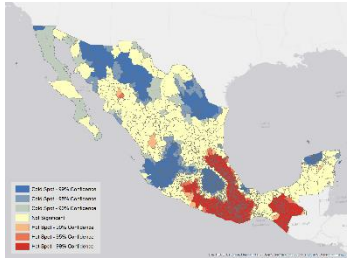


Hot and Cold Spots of the Poverty Rate

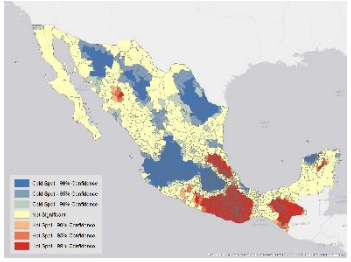
1995



2005

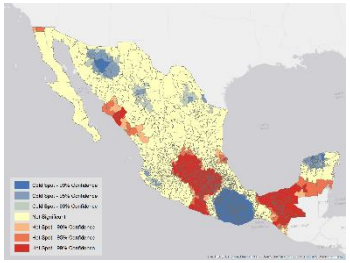


2015

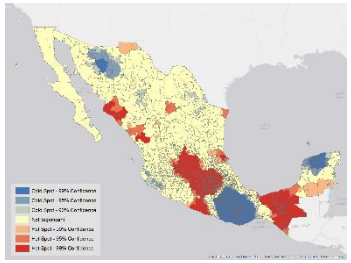


Hot and Cold Spots of the Poor

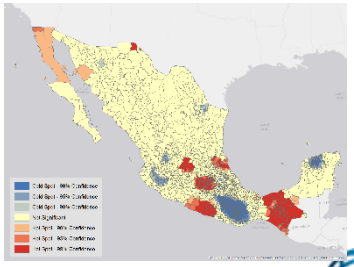
1995



2005



2015

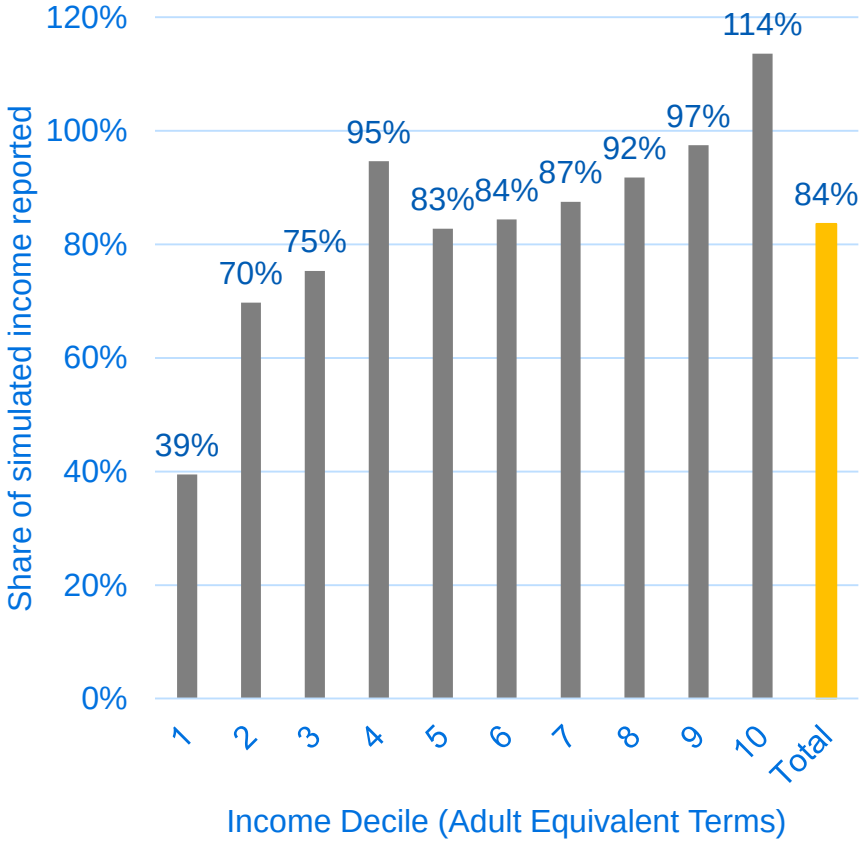


Note: Hot and Cold spots produced by author using poverty data produced by CONEVAL using data from INEGI.

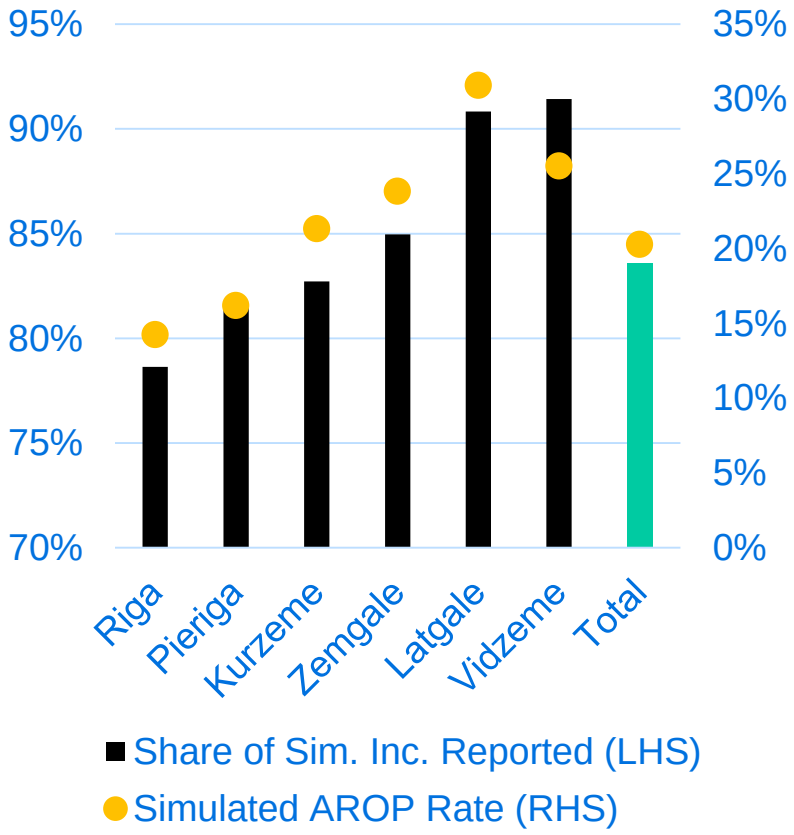
Potential further applications of the method: Inequality

Survey reported income and tax based income have different reporting biases, and SAE can help us explore the complementarity of both

Shared of Simulated Income Reported in the Tax Records



Share of Simulated Income and Poverty Rate by Regions



WORLD BANK GROUP

Potential further applications of the method: Short Welfare Modules

How good are our results?

- Household consumption patterns can be a great predictor of overall consumption level
- A limited set of 35 questions if a household has consumed certain COICOP level 4 items can go a long way towards obtaining a comparable consumption aggregate

Table 1: ELL simulation of adult equivalent consumption for Greece

	Sim (1)	Sim (2)	Sub sample 2 sim (1)	Sub sample 2 sim (2)
Model details:				
Observations	5,857	5,857	2,860	2,860
Regressors	35	35	35	35
Adjusted R2	0.67	0.67	0.66	0.66
Max. VIF	3.63	3.63	3.59	3.59
F-Stat	345.80	345.80	249.40	161.51
Het. Adj. R2		2.E-03		1.E-03
Het. F-Stat		6.50		2.52

Source: Greek 2014 HBS, own estimates

Table 2: Simulation exercise for Greece

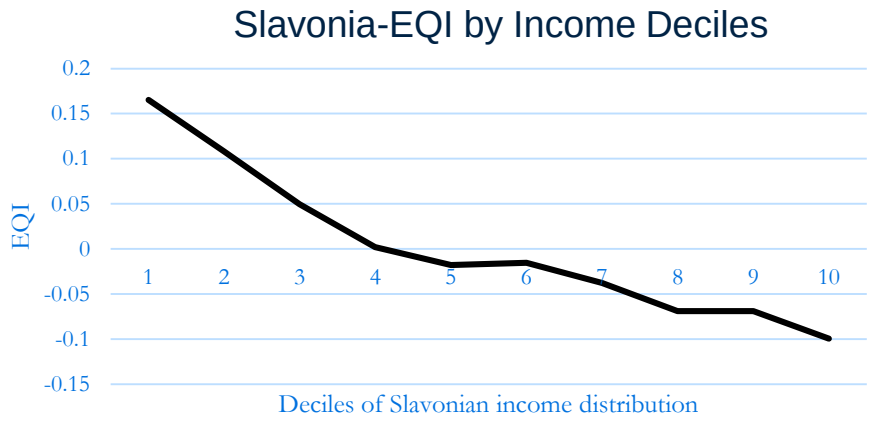
	Direct estimate	Full sample Sim (1)	Full sample Sim (2)	Sub-sample (50%) Sim (1)	Sub-sample (50%) Sim (2)
Children					
Headcount	21.5	25.1	20.3	25.1	20.0
Gap	4.1	8.4	5.2	8.3	5.2
Severity	1.4	3.9	1.9	3.9	2.0
Elderly					
Headcount	27.3	31.9	26.8	32.3	27.4
Gap	6.0	10.6	6.5	10.7	6.6
Severity	2.1	4.9	2.3	4.9	2.3
National					
Headcount	20.9	25.2	19.9	25.5	20.2
Gap	4.8	8.3	4.9	8.4	5.0
Severity	1.8	3.8	1.8	3.8	1.8
Gini	34.6	42.7	34.7	42.7	34.6
GE 0	20.0	31.5	19.8	31.6	19.7
GE1 (Theil)	21.1	32.1	20.3	32.0	20.2
GE2	28.2	45.7	25.6	45.3	25.2

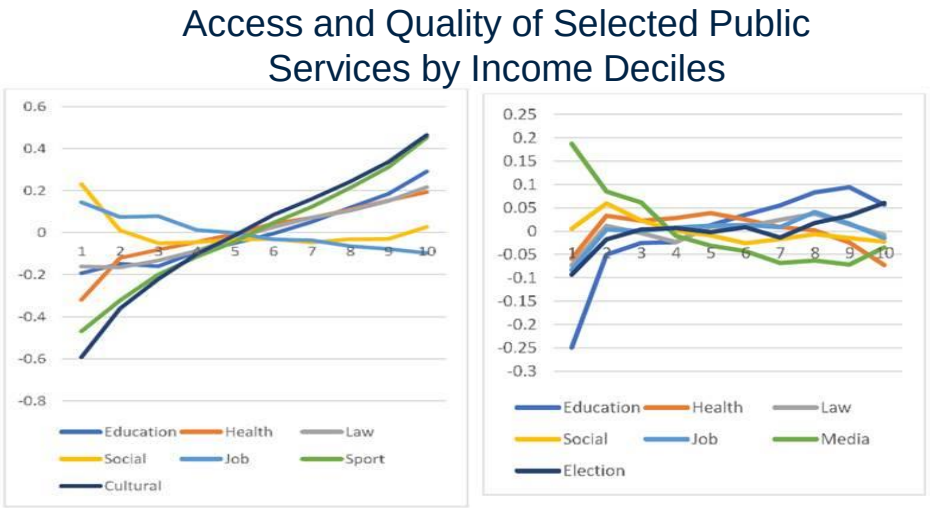
Source: Greek HBS 2014, Own estimates.

Note: Relative line (60% of median adult equivalent consumption). Sim 1: Only consumption dummies, and nat log. of hosuehold size. Sim 2: Only consumption dummies, **heteroskedasticity** modeled with share children and elderly. Simulations done using ELL methodology. Full sample (5,888 hh).

Potential further applications of the method: Short Welfare Modules

How it can be used?





- In many cases it is not feasible to include consumption in one survey for every respondent in one survey
  - Security concerns
  - Cost
  - Time
  - Increases respondent fatigue
- A full aggregate is necessary for monitoring poverty, or program evaluation
- The need for an aggregate database which includes income and consumption
  - EUROMOD uses this for simulating direct and indirect tax policies

WORLD BANK GROUP

Potential further applications of the method

Spatial Price differences matter, and we are piloting two approaches

Method 1: A spatial approach using food prices

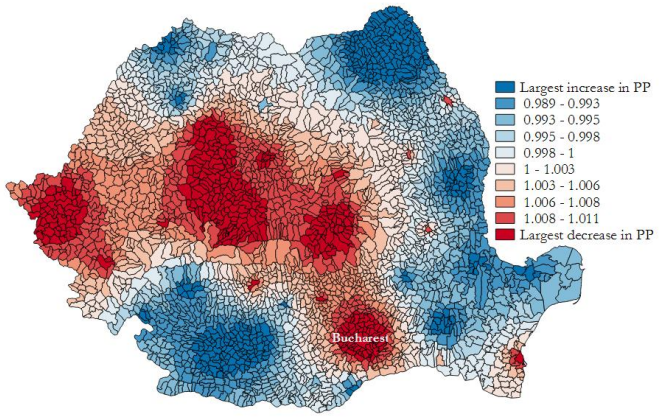
use the 2011 Household Budget Survey (HBS) for Romania to estimate local food prices. We then use the county level Paasche to obtain an index for each of the country's 3,181 municipalities.

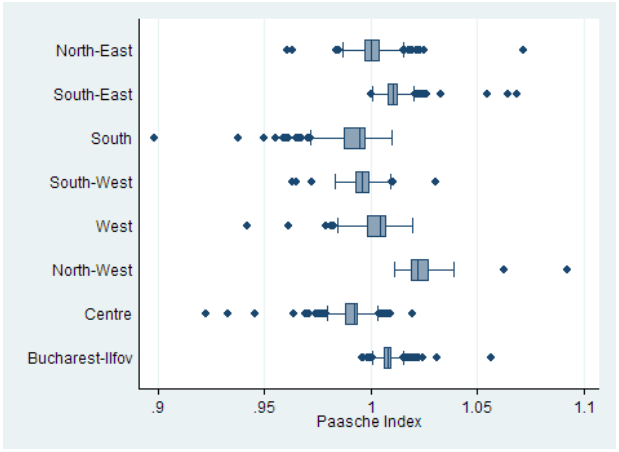
Method 2: A hedonic approach using rent data

A hedonic regression model is typically used to account for housing price differences across space (e.g. Atamanov et al., 2016). However, this method requires housing price data for each region included in the estimation. We use imputed rent data from the EU-SILC, which only covers 650 LAUs out of the 3,181 in the country. To estimate rent for all LAUs, we combine the EU-SILC data with Census data using the same approach we used to estimate income at the LAU level. The 650 LAUs are covered by a total of 7670 households for which rent is imputed.

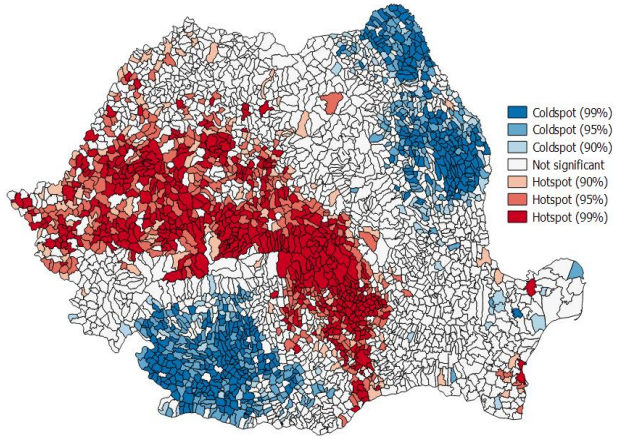
Potential further applications of the method: Spatial differences on cost of living can be quite remarkable, and it is important that going forward we experiment more systematically with methods to better capture this heterogeneity in our measures of poverty and deprivation

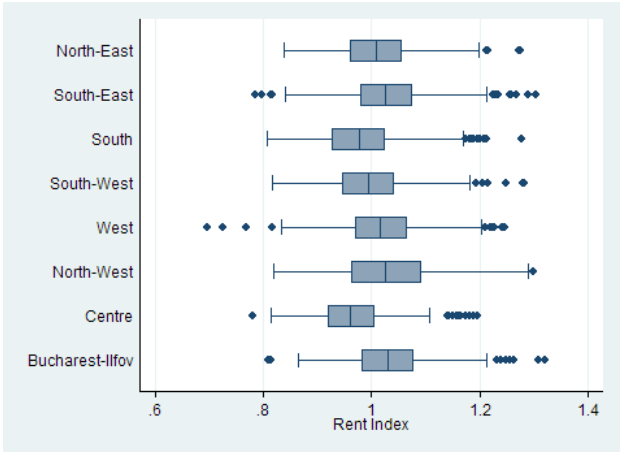
Food Price Index





Getis-Ord Hotspot Analysis of Rent Index





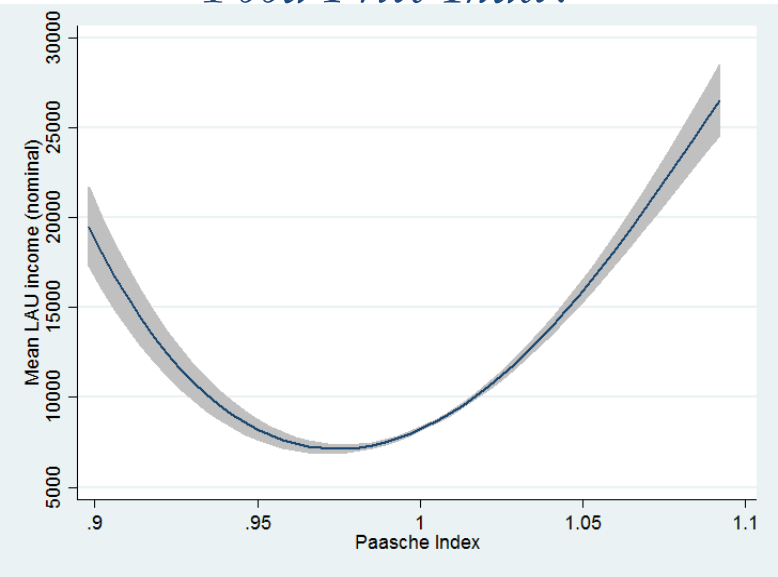
Source: Romania NSO / World Bank Poverty Map

WORLD BANK GROUP

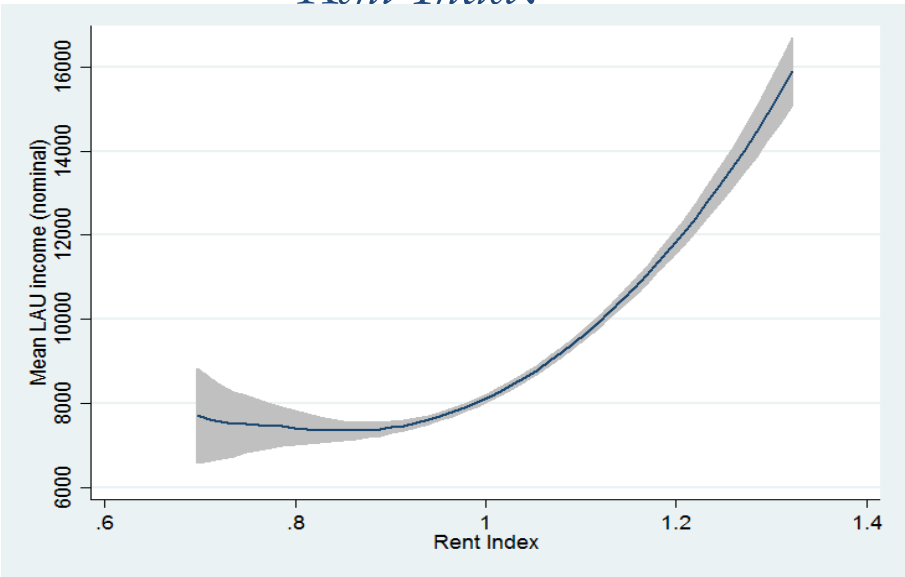
Potential further applications of the method

To test for the relationship between LAU level deflators and income, we use mean nominal income by LAU from the Romanian poverty maps.

Food Price Index



Rent Index



Source: Romania NSO / World Bank Poverty Map

The local polynomial regressions presented in Figure 6 illustrate the positive relationship between income and the Paasche, as well as with the rent index. Despite the U-shaped parabola for the Paasche, the results provide evidence of the Penn Effect; prices are higher in better off localities (Samuelson, 1994).

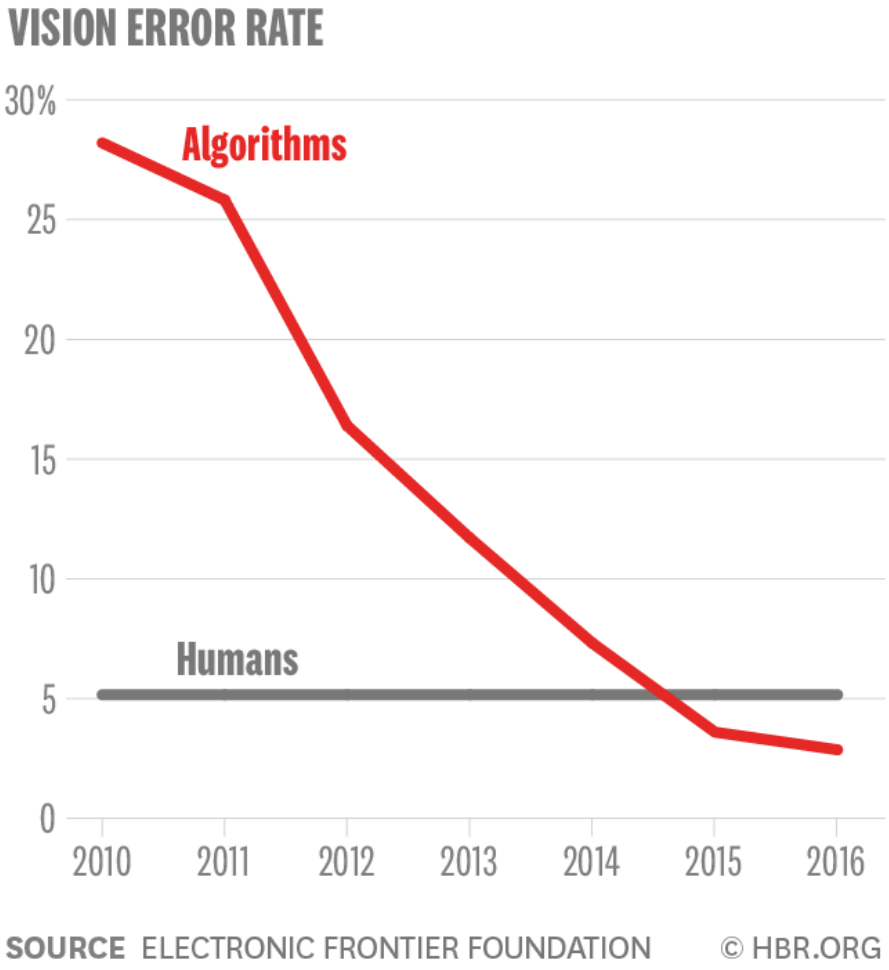
How to improve the re-purpose of SAE estimates?

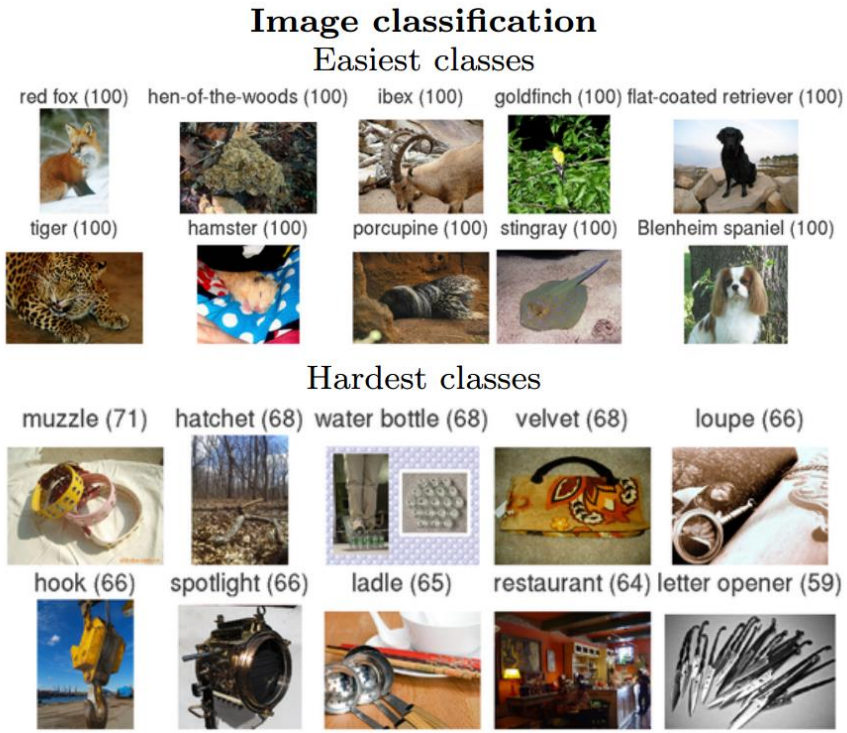
- **Documentation**: in several cases the documentation is insufficient to allow the full replicability and understanding of the quality of the exercise

- **Access**: often producers do not report standard errors associated with each point estimate; and omit the reporting of other derived indicators (such as the FGT1 and FGT2)

- **Methods and Tools**: it is important to develop methods and tools that facilitate the consumption of SAE estimates as a secondary data source

A poor map or a map of the poor?

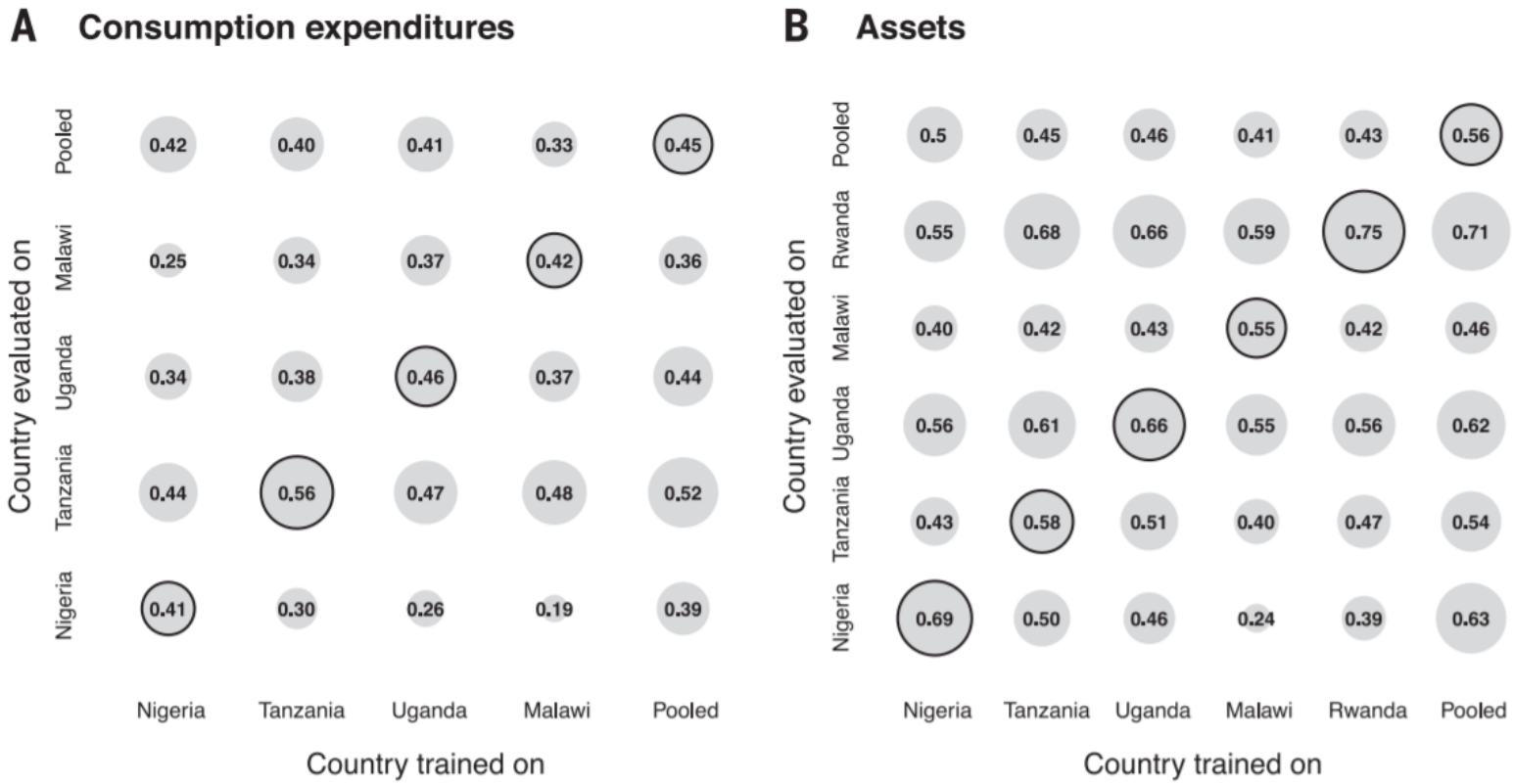




WORLD BANK GROUP

But we are not quite yet ready for prime time when it comes to social and welfare layers…

Cross-border model generalization. (A) Cross-validated r 2 values for consumption predictions for models trained in one country and applied in other countries. Countries on x axis indicate where model was trained, countries on y axis where model was evaluated. Reported r 2 values are averaged over 100 folds (10 trials, 10 folds each). (B) Same as in (A), but for assets.



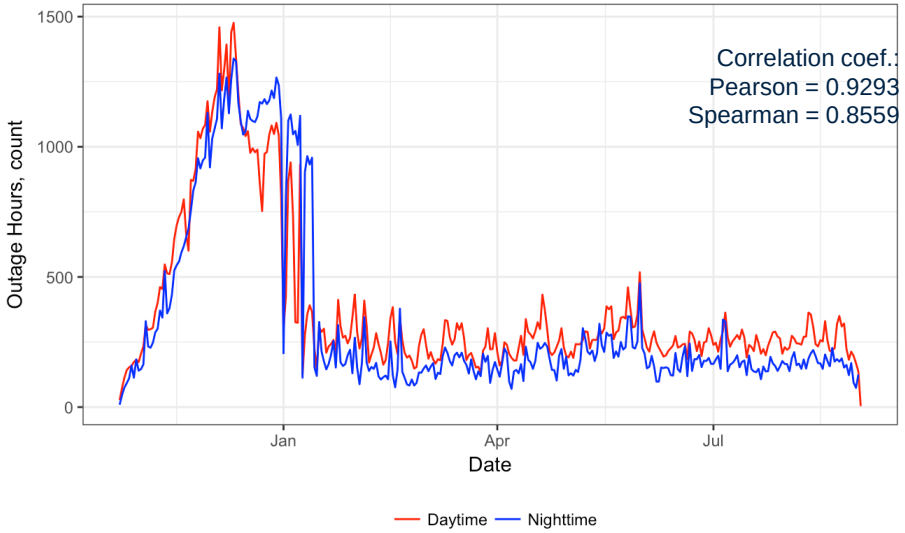
Neal et al (2016) Combining satellite imagery and machine learning to predict poverty. Science, 19 Aug 2016: Vol. 353, Issue 6301, pp. 790-794

WORLD BANK GROUP

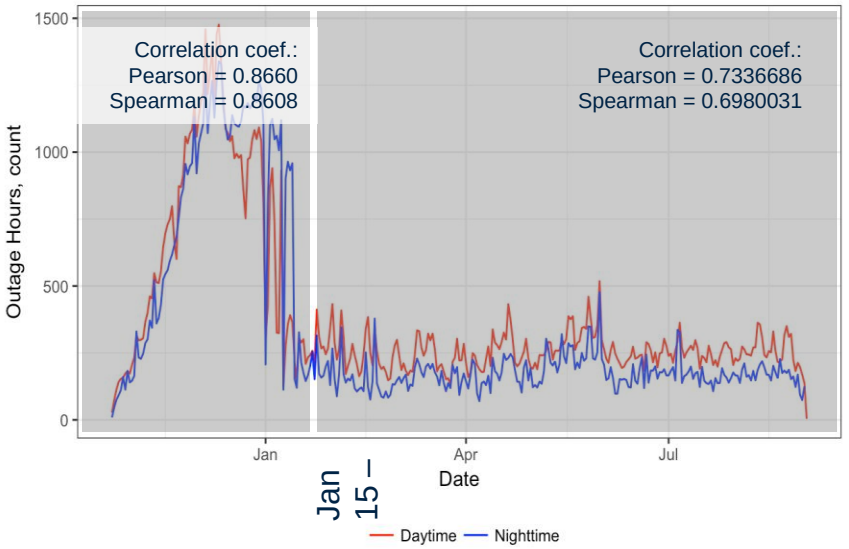
…. since not everything we measure from space has a robust relationship with what happens on the ground over time and space.

(correlations from nightlight data and IoT measures of power outages in Tajikistan)

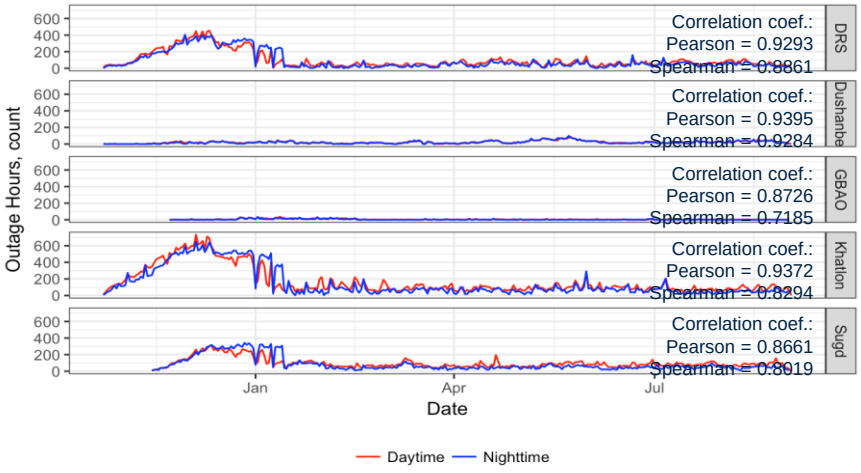
Temporal correlation of daytime and night time power outages



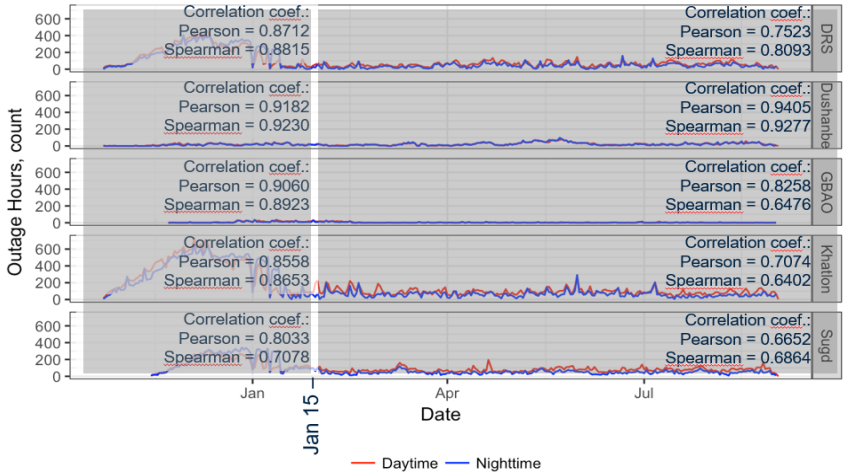
And the temporal correlation is not stable overtime…



And this relationship is not homogenous across space



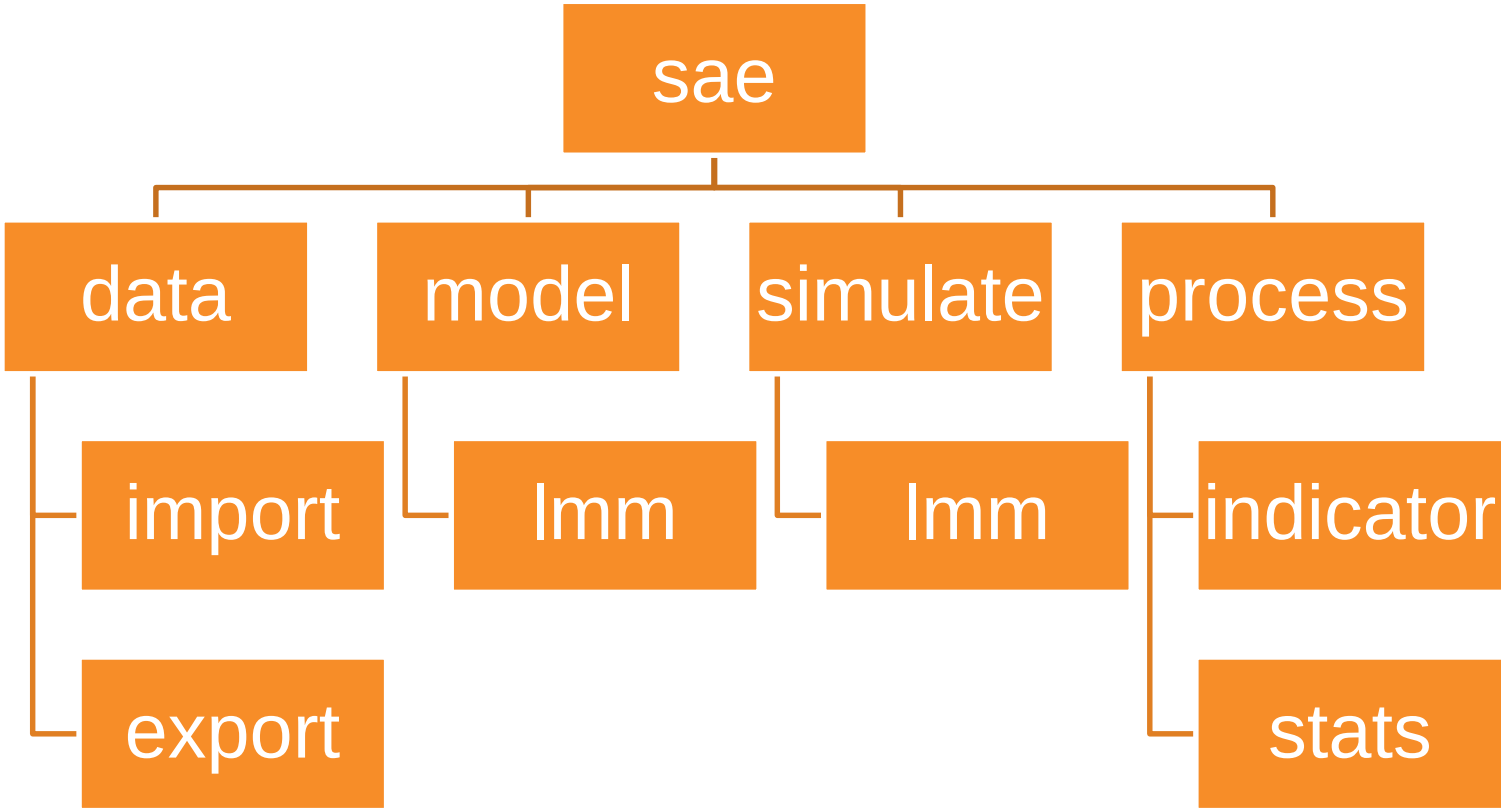
And clearly not by space and time



The World Bank team has also invested on the creation of tools to make the replicability of SAE easier….

- 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.

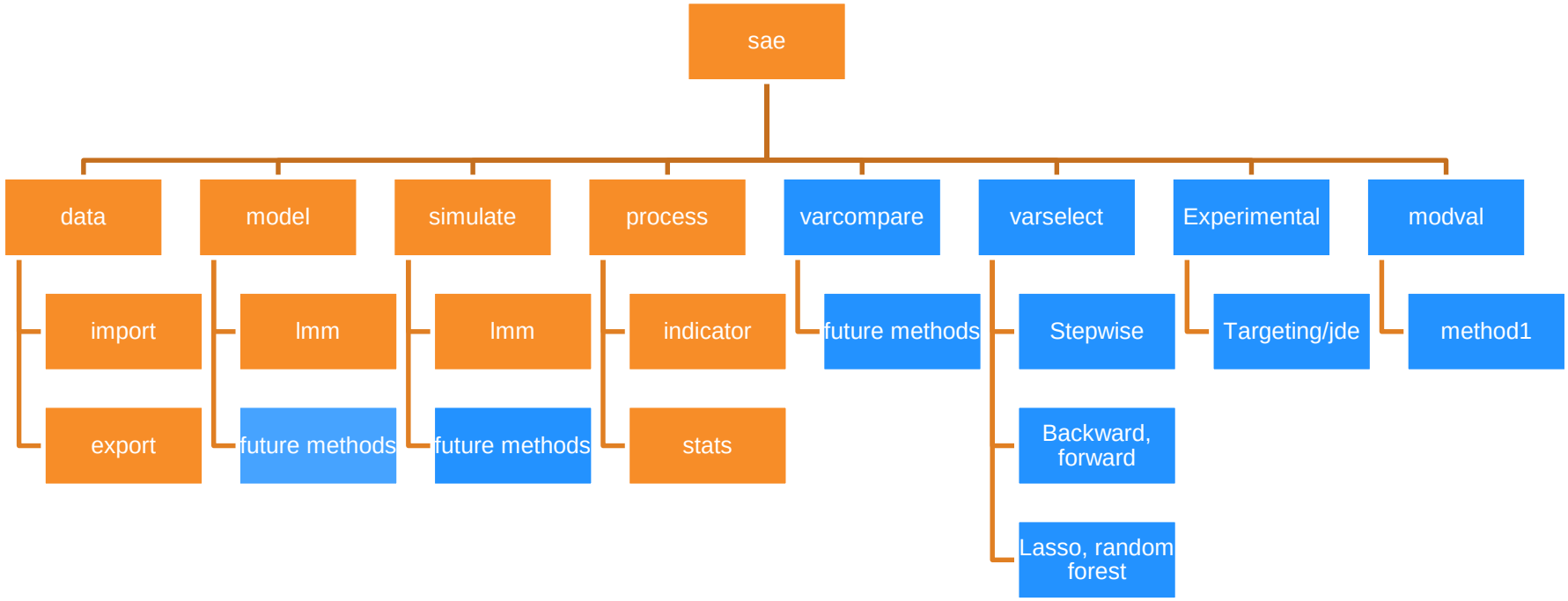
Framework of the Stata sae syntax [*current*]



lmm: linear mixed model

WORLD BANK GROUP

Framework of the Stata SAE syntax [future/plan]



WORLD BANK GROUP

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

Going forward it is also important to learn how to improve the usage of the microdata produced by teams, since there seem to be significant heterogeneity across countries and statistical operations

Citations on reports, articles, theses, books, abstracts, from government, academic publishers, professional societies, online repositories, universities and other web sites

Country	Keyword	Any Time	since 2017	since 2016	since 2013
Brazil	PNAD+BRASIl	35,200	1,090	4,170	13,100
Brazil	PNAD+BRAZIL	25,500	726	3,120	9,350
Mexico	ENIGH+MEXICO	21,300	5,460	17,400	17,300
Colombia	DANE GEIH	1,470	90	290	816
Colombia	DANE ECV	1,910	51	185	667
USA	"American Community Survey"	44,000	1,630	9,430	17,900
USA	"Current Population Survey"	127,000	1,490	6,410	15,500
UK	"British Household Panel Survey"	16,300	472	1,550	4,920
Multiple	"Demographic and Health Survey" OR DHS	391,000	2,230	22,800	17,800
Multiple	"Living Standards Measurement Survey" OR LSMS	25,600	758	2,290	7,350
Multiple	"European Community Household Panel" OR ECHP	16,800	213	818	3,340
Multiple	"Survey of Income and Living Conditions" OR SILC	26,000	1,040	3,120	10,500
Multiple	"Survey of Income and Living Conditions" OR SILC AND UDB	1,060	51	116	418
Multiple	EUROMOD	3,070	104	359	1,160
Multiple	EUROMOD + SILC	919	46	146	467

Source: Google Scholar as of June 16th 2017

Final Remarks and main lessons learned

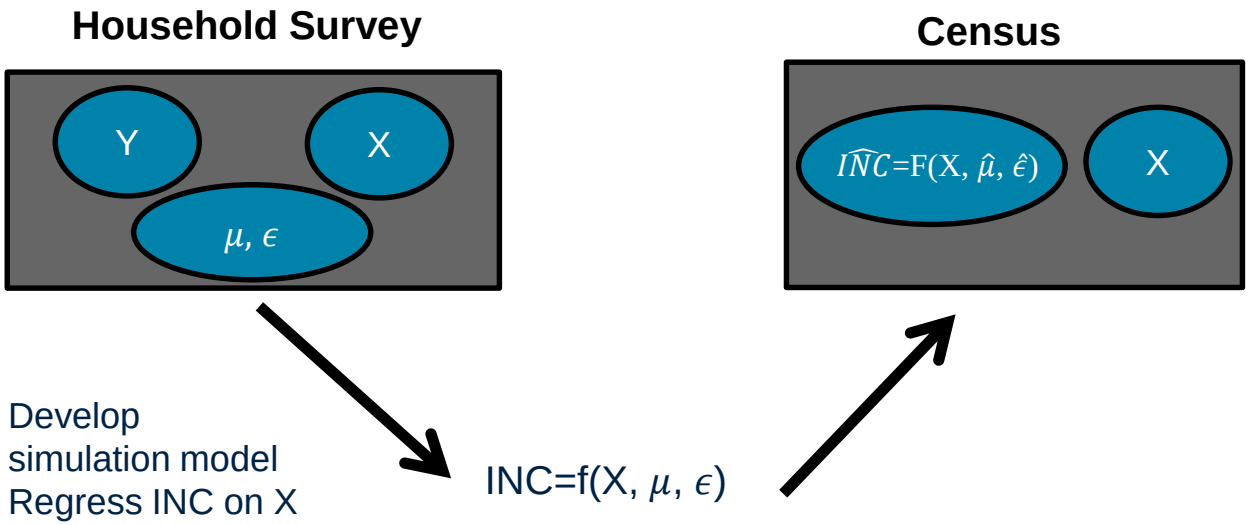
- The level of spatial disaggregation matters for the policy relevance of our data and our understanding of the phenomena
- There is significant payoff to be able to link Population Census and the SILC (i.e. Bulgaria and Latvia examples)
- SAE methods and SILC can significantly complement each other, specially given that response biases are quite different (i.e. Latvia)
- We need to do better on the presentation, documentation and re-use of this estimates (for policy purposes)
- It takes time, and automated results are unlikely to produce reliable results

Obrigado

jazevedo@worldbank.org

WORLD BANK GROUP

The Basics of ELL (Elbers, Lanjouw, and Lanjouw 2003)



$\hat{Y}$: **Simulated** welfare (replicated many times)
X: Poverty correlates like employment, education
μ, ϵ: Terms relating to the area and model error (replicated many times)

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: $\boldsymbol{\hat{u}_{ch} = \hat{u}_{c.} + (\hat{u}_{ch} -}$

- We estimate the following: $y_{ch} = x_{ch}^T \beta + \eta_c + e_{ch}$
- Obtain GLS estimates, where $\mathrm{Var}(\hat{\boldsymbol{\eta}}_{\boldsymbol{c}}) = \sigma_\eta^2$ and $\mathrm{Var}(\hat{\boldsymbol{e}}_{\boldsymbol{c}}) = \sigma_{e_{ch}}^2$

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

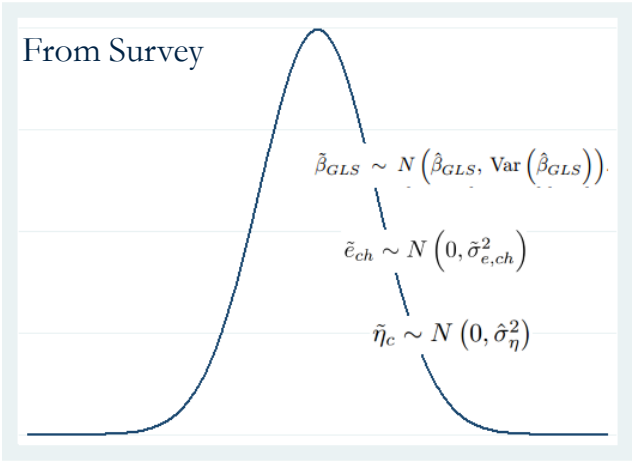
In practice, the adjusted R-squared of the alpha-model is often very modest

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

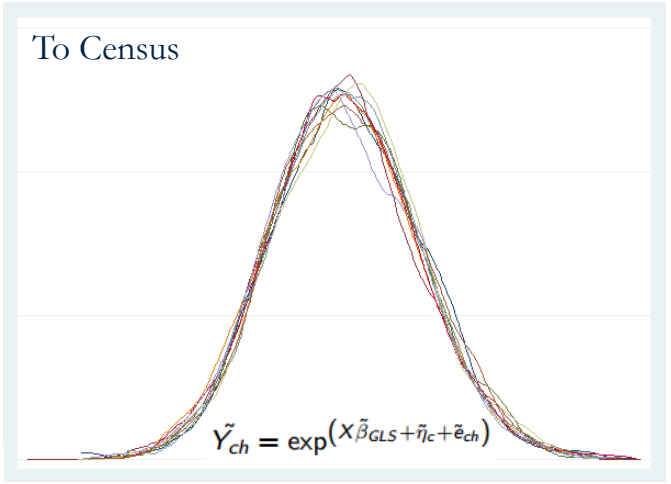
By defining $exp^{z'\alpha} \equiv D$ and using the Delta Method (Taylor expansion for $E[\sigma_{ch}^2]$) we get:

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

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

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.