Synthetic data sources in the spatial analysis of poverty

Wojciech Roszka
Poznań University of Economics

4^{th} Baltic-Nordic Conference on Survey Statistics
BaNoCoSS-2015
Helsinki, Finland

24-28 August 2015

The objective of the study

1. The creation of full-coverage synthetic datafile – based on EU-SILC.
2. The estimation of **poverty indicator** at NUTS 3 level.
3. The quality assessment and comparison with other studies.

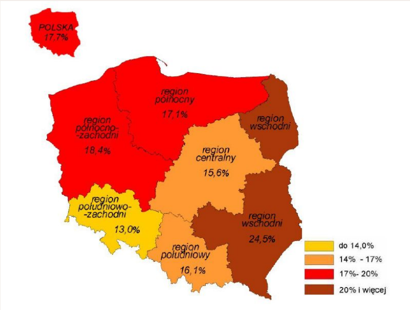
EU-SILC

1. EU-SILC (*European Union Statistics on Income and Living Conditions*) is a sample survey conducted yearly in all European Union countries.
2. The objective of EU-SILC is to obtain a primary source of comparable data at EU level on the income situation, poverty and other aspects of living conditions of the population.
3. EU-SILC units are private households and persons aged 16 years and older included in these households.
4. The survey is carried out in May-June of current year.
5. The reference period for the income variables is the last full calendar year. Reference period for other variables is the current situation.

At-risk-of-poverty-rate - definition

At-risk-of-poverty-rate after taking into account social transfers

The percentage of people with equivalent disposable income below the risk-of-poverty threshold, which is 60% of the national median of equivalent disposable income after social transfers.
The value of the indicator in 2011 in Poland was **17.7%**.



Sample size and generalization of results

1. In 2011 the effective sample size of EU-SILC was 12,871 households (which was approx. 65% of the established size).
2. The poverty indicator in Poland is published for the whole country and at NUTS 1 level.

Weightings

1. Design weights (variable *DB*080) are the inversion of inclusion probability of apartment in h layer:

$$f_h = \frac{n_h * m_h^{'}}{M_h} \tag{1}$$

where:
n_h - the number of areas drawn from h-layer;
$m_h^{'}$ - the number of apartments drawn in h-layer;
M_h - the total number of apartments in h-layer.

Completeness factor

Design weights DB080 were corrected using so called „completeness ration" computed for each class of place of residence separately using formula:

$$DB080_p^{cor} = \frac{DB080_p}{cr_p} \tag{2}$$

where:
cr_p - completeness ratio in class p.
On the basis of $DB080_p^{cor}$, the final weights were computed - DB090.

The symbol of class of place of residence (p)	**Class of place**	**Completeness ratio (cr_p)**
	Poland	0,649
1	Warsaw	0.411
2	cities 500 k - 1 mln	0.473
3	cities 100 k - 500 k	0.625
4	cities 20 k - 100 k	0.669
5	cities below 20 k	0.684
6	village	0.747

Weight calibration

Using external - Census - information, the DB090 weights were calibrated with IPF algorithm. The correction was based on the joint distribution of households in subregions (NUTS 3) and household size cross-sections.

Raking was made using loglinear model:

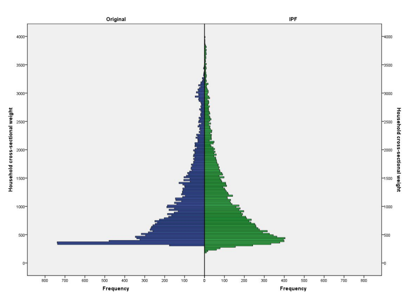
$$N_{ij} = a_i b_i n_{ij} \tag{3}$$

written as probabilities: $\pi_{ij} = a_i b_i p_{ij}$ where:
π_{ij} and p_{ij} are population and sample probabilities respectively

$$log(\frac{\pi_{ij}}{p_{ij}}) = log(a_i) + log(b_i) + \epsilon_{ij} \tag{4}$$

- loglinear models are fit by IPF
- observed counts are assumed to be independent Poisson variables
- fit by MLE using Newton-Raphson algorithm

Weight calibration



	Original weight	IPF weight
Mean	1054.15	1054.15
StDev	718.14	744.12
Median	810.82	813.65
Min	292.37	181.97
Max	3584.68	9718.35

Records replication

- The records were replicated basing on rounded values of calibrated weight.

NUTS 3	Sex	Age	Mar. status	Place of res.	Educ.	weight	
01	M	15-19	single	city	primary	1000	replicated 1000x
...	...	...	...	...	...	...	
16	F	60-64	widow	countryside	secondary	100	replicated 100x
...	...	...	...	...	...	...	
66	M	40-49	married	city	tertiary	2000	replicated 2000x

- The datafile containing 13,568,068 synthetic units was created.
- The values of at-risk-of-poverty variable (*Poverty Indicator*, HX080) for replicated records was deleted (it was contained for original, sample, records).
- In such datafile multiple imputation method was performed.

Multiple imputation

- Each missing data is imputed by multiple (m) values.
  - theoretical values are imputed from estimated model:

$$\tilde{y}_i = \hat{y}_i + e_i = \hat{\alpha}_Y + \hat{\beta}_{YX} x_i + e_i, e_i \sim N(0, \hat{\sigma}_{Y|X}) \tag{5}$$

- These m values are ordered in such a way that the first set of values forming a first dataset, etc.
- It means that for m values, m complete (synthetic) datasets are being created.
- Each of these sets are analyzed using standard procedures using the full information in such a way as if the imputed values were true.

The imputation estimator for each of t ($t = 1, 2, \ldots, m$) models is $\hat{\theta}^{(t)} = \theta(U_{obs}, U_{mis}^{(t)})$, where U_{obs} are observed values, and $U_{mis}^{(t)}$ are imputed missing data. The variance of the estimator is formulated as $\widehat{var}(\theta^{(t)}) = \widehat{var}(\hat{\theta}(U_{obs}, U_{mis}^{(t)}))$.

Estimation

The point estimate of the multiple imputations is an arithmetic mean:

$$\hat{\theta}_{MI} = \frac{1}{m}\sum_{t=1}^{m}\hat{\theta}^{(t)} \tag{6}$$

"Between-imputation" variance is estimated by formula:

$$B = \frac{1}{m-1}\sum_{t=1}^{m}(\hat{\theta}^{(t)} - \hat{\theta}_{MI})^2 \tag{7}$$

and "within-imputation" variance is estimated by:

$$W = \frac{1}{m}\sum_{t=1}^{m}\widehat{var}(\hat{\theta}^{(t)}) \tag{8}$$

Total variance is a sum of between- and within-variance **modified by** $\frac{m+1}{m}$, to reflect the **uncertainty** about the true values of imputed missing data:

$$T = W + \frac{m+1}{m}B \tag{9}$$

Estimation

Interval estimates are based on t-distribution:

$$\hat{\theta}_{MI} - t_{v,\frac{\alpha}{2}}\sqrt{T} < \theta < \hat{\theta}_{MI} + t_{v,\frac{\alpha}{2}}\sqrt{T} \tag{10}$$

with degrees of freedom:

$$v = (m-1)(1 + \frac{W}{(1+\frac{1}{m})B})^2 \tag{11}$$

Imputation model

- Poverty indicator (HX080) had 2 categories:
  - at risk of poverty (1)
  - not at risk of poverty (0)
- Logistic regression model was used.
- m=10 imputations were performed.

Logistic model

Variables used

Head of HH characteristics

- gender
- is he/she is still in education
- level of education
- marital status
- health condition
- age

HH characteristics

- Capacity to afford paying for one week annual holiday away from home
- Ability to make ends meet
- Class of place of residence
- Total disposable household income
- Voievodship (region)

HH composition characteristics

- number of minors
- number of unemployed
- number of inactive
- number of disabled

Logistic model

Selected model statistics – sample

- Nagelkerke's R^2 – 0.782
- The percentage of values correctly classified – 94.4%

The analysis

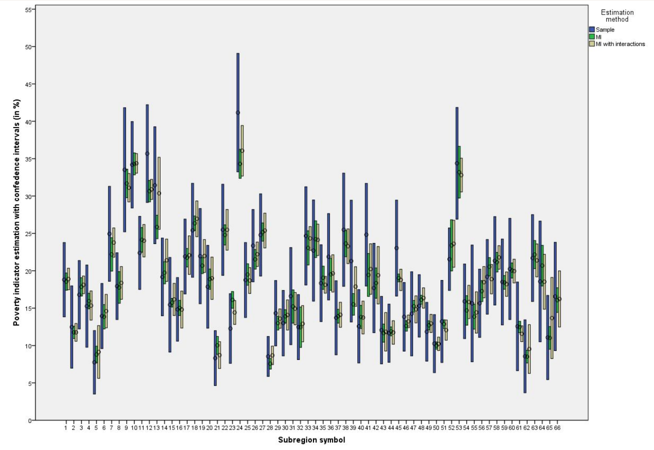
- Model without interactions
  - Computation time: 20 hours ☹
- Model with two-way interactions
  - Computation time: 6.5 days ☹ ☹ ☹

Comparison of sample and MI estimation

Spatial unit	Point estimate		
	Sample	MI	MI int
Country	17.7	17.3	17.6
NUTS 1 level			
CENTRAL	15.6	15.7	16.2
SOUTH	16.1	15.7	15.6
EAST	24.5	24.3	24.4
NORTH-WEST	18.4	17.8	18.2
SOUTH-WEST	13.0	14.3	14.4
NORTH	17.1	16.0	16.8

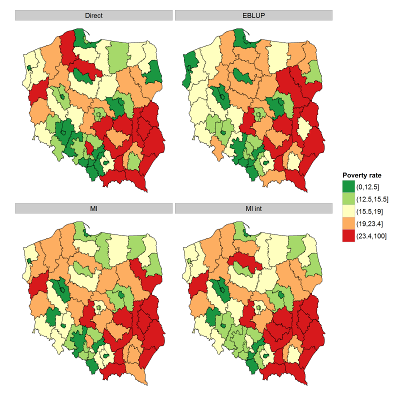
Comparison of sample and MI estimation

NUTS3 - at-risk-of-poverty-rate interval estimation



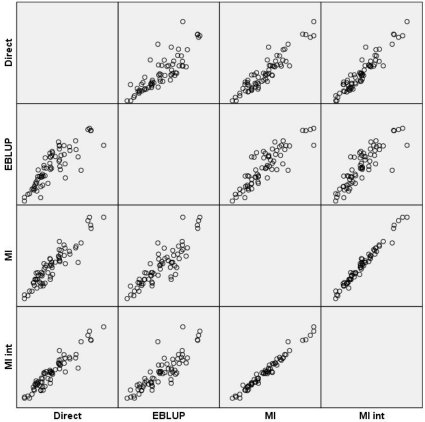
Comparison of EBLUP and MI estimation

Poverty Mapping - World Bank project, Statistical Office in Poznań



Comparison of EBLUP and MI estimation

Poverty Mapping - World Bank project, Statistical Office in Poznań



Conclusions

- The syntetic multivariate dataset containing basic characteristics on households was created.
- The information on at-risk-of-poverty was added.
- The estimation results were consistent with those obtained in other studies.

Drawbacks

- The quality assessment was based on the number of artificial records.
- The computation issues.
- The danger of model misspecification.
- The quality of results is directly dependent on sample quality.

Discussion

- The integration of other sample surveys like HBS and LFS using data fusion.
  - Increasing the effective sample size.
  - Matching new variables.
- The use of spatial microsimulation methods.

Literature

- Alfonso A., Filzmoser P., Hulliger B., Kolb J-P., Kraft S., Munnich R., Templ M., 2011, *Synthetic Data Generation of SILC Data*, European Commision, Community Research, AMELI Project
- Ballas D., Rossiter D., Thomas B., Clarke G.P., Dorling, D. 2005, *Geography Matters: Simulating the Local Impacts of National Social Policies*, York, Joseph Rowntree Foundation, UK
- Cohen M.L. 1991, *Statistical matching and microsimulation models* [in:] Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling, Vol. II: Technical Papers. Washington, DC: National Academy
- Haslett S., Jones G., Noble A., Ballas D., 2010, *More or Less? Comparing small area estimation, spatial microsimulation, and mass imputation*, Section on Survey Research Methods – JSM, American Statistical Association
- Raessler S. 2002, *Statistical Matching. A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*, Springer, New York, USA
- Rahman A. 2008, *A Review of Small Area Estimation Problems and Methodological Developments*, Discussion paper 66, NATSEM, University of Canberra
- Tanton R., Edwards K. L. (ed.) 2013, „*Spatial Microsimulation: A Reference Guide for Users*", J. Bus. Econ. Stat. 4, 87–94
- Wawrowski Ł., 2014, *Wykorzystanie metod statystyki małych obszarów do tworzenia map ubóstwa w Polsce* [*The use of small area estimation methods for poverty mapping in Poland*], Wiadomości Statystyczne 9/2014, Polskie Towarzystwo Statystyczne [Polish Statistical Association]