SAMPLE
Small Area Methods for Poverty
and Living Condition Estimates

Cumulation of Poverty measures: the theory beyond it, possible applications and software developed

(**Francesca Gagliardi** and **Giulio Tarditi**)

Siena, October 6th , 2010

Context and scope

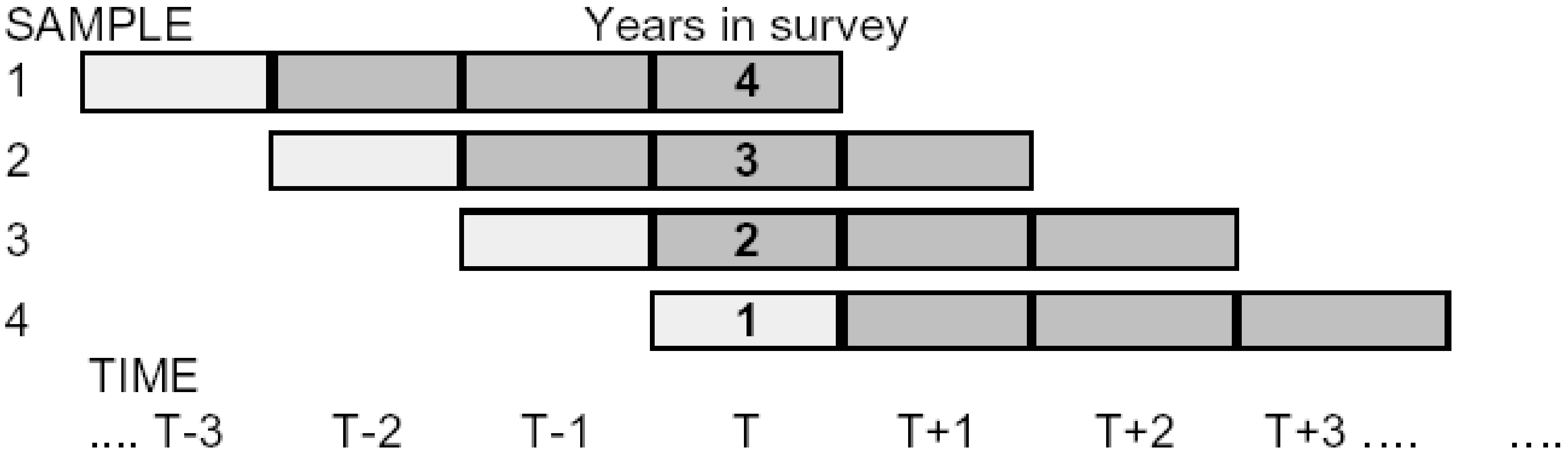
Reliable indicators of poverty and social exclusion are an essential monitoring tool. In the EU-wide context, these indicators are most useful when they are comparable across countries and over time. Furthermore, policy research and application require statistics *disaggregated to increasingly lower levels and smaller subpopulations*. Direct, one-time estimates from surveys designed primarily to meet national needs tend to be insufficiently precise for meeting these new policy needs. This is particularly true in the domain of poverty and social exclusion, the monitoring of which requires complex distributional statistics – statistics necessarily based on intensive and relatively small-scale surveys of households and persons.

This work addresses some statistical aspects relating to improving the sampling precision of such indicators in EU countries, in particular through the *cumulation of data over rounds of regularly repeated national surveys*.

EU-SILC

The reference data for this purpose are EU Statistics on Income and Living Conditions, the major source of comparative statistics on income and living conditions in Europe.

A standard integrated design has been adopted by nearly all EU countries.



It involves a rotational panel, with a new sample of households and persons introduced each year to replace one-fourth of the existing sample. Persons enumerated in each new sample are followed-up in the survey for four years. The design yields each year a cross-sectional sample, as well as longitudinal samples of 2, 3 and 4 year duration.

Different modes of using data

Survey data such as from EU-SILC can be used in different forms to construct regional indicators.

(1) Direct estimation from survey data – in the same way as at the national level – provided that the regional sample sizes are large enough.

(2) Constructing alternative indicators - but with substantively similar meaning - which utilise the available survey data more intensively.

(3) Cumulation of data over survey waves to increase precision of the direct estimates.

(4) Using survey data in conjunction with data from other - especially administrative - sources, to produce improved estimates using small area estimation (SAE) techniques.

(5) Going altogether beyond the survey by constructing indicators from alternative sources.

The issues addressed in our work concern the efficiency of (3).

Gain from cumulation over waves

Consider that a person’s poverty status (poor or non-poor) is determined from the income distribution for each wave separately, and the proportion of poor at that wave is computed. These proportions are then averaged over a number of consecutive waves.

The issue is to quantify the gain in sampling precision from such pooling, given that data from different waves of a rotational panel are highly correlated.

Full overlap, two waves, simple model:	$v_A = \frac{v}{2}.(1+b)$	$b = \left(\frac{c_1}{v}\right) = \left(\frac{a-p^2}{p-p^2}\right)$
Partial overlap, two different waves:	$v_A = \frac{1}{2}.\left(\frac{V_1+V_2}{2}\right).\left(1+b.\left(\frac{n}{n_H}\right)\right)$	
Full overlap, multiple waves, simple model:	correlation declines with increasing distance (k years) between waves, e.g. as:	$(\mathrm{c}_\mathrm{k}/\mathrm{v}) = (\mathrm{c}_1/\mathrm{v})^\mathrm{k}$
	gain in precision with averaging over K waves (compared to a single wave):	$f_c = \left(\frac{v_k}{v}\right) = \frac{1}{K}.\left(1+2.\sum_{k=1}^{K-1}\left(\frac{K-k}{K}\right).\left(\frac{c_1}{v}\right)^k\right)$

where:

a is the persistent poverty rate over the two years

n is the overlap between the cross-sectional samples

n_H is the harmonic mean of the cross-sectional sample sizes n_1 and n_2

***Assumption in our work*:**

Given that in EU-SILC data cross-sectional datasets cannot be linked, correlation between the two waves cannot be computed directly.

We computed it through the longitudinal dataset. So:

- ♦ $b = \left(\frac{a - p^2}{p^2 - p} \right)$ where *a* is persistent poverty rate of the panel (e.g. 2007-2008 for Italy) and *p* is the average of the estimates done in the longitudinal dataset for each of the two years (e.g. 2007-2008 for Italy).
- ♦ $n_h = \frac{2 \cdot n_1 \cdot n_2}{n_1 + n_2}$ where n_1 and n_2 are the sample size of the cross sectional dataset.
- ♦ $n = 0.75 \cdot \min(n_1, n_2) \cdot \frac{m}{\min(m_1, m_2)}$ where m is the sample size of the panel (e.g. 2007-2008 for Italy) and m_1 and m_2 are the sample size of the longitudinal dataset in the two years.

Gain from cumulation over two waves.

		Italy EU-SILC 2007-2008	**Poland** EU-SILC 2005-2006
Standard error of average HCR over two years (assuming independent samples)	$(1) \quad = \quad \frac{(V_1 + V_2)^{1/2}}{2}$	0.36	0.34
Factor by which standard error is increased due to positive correlation between waves	$(2) = (1 + b \cdot (n/n_H))^{1/2}$	1.20	1.18
Standard error of average HCR over two years (given correlated samples)	$(3) = (1) \cdot (2)$	0.43	0.40
Average standard error over a single year	$(4) = \frac{(V_1)^{1/2} + (V_2)^{1/2}}{2}$	0.50	0.48
Average gain in precision (variance reduction, or increase in effective sample size, over a single year sample)	$(5) = 1 - ((3)/(4))^2$	26%	30%

Variance estimation for complex statistics from complex samples

The Jackknife Repeated Replication (JRR) provides a versatile technique for variance estimation for such statistics, including cumulative and longitudinal measures.

JRR is one of the variance estimation methods based on comparisons among replications generated through repeated re-sampling of the parent sample. Once the set of replications has been defined for any complex design, the same variance estimation algorithm applies to a statistic of any complexity:

$$\mathrm{var}(\lambda) = \Sigma_k \left[(1 - f_k) . \frac{a_k - 1}{a_k} . \Sigma_j \left(\lambda_{(kj)} - \lambda_{(k)} \right)^2 \right]$$

In the standard application of JRR, a replication is formed by (i) eliminating one PSU from the sample, and (ii) appropriately increasing the weight given to the remaining units in its stratum. The number of replications generated is the same as the number of PSUs in the sample.

Extensions to longitudinal and other measures over time

We have extended and applied this method for estimating variances for subpopulations (including regions and other geographical domains), longitudinal measures such as persistent poverty rates, and measures of net changes and averages over cross-sections in the EU-SILC rotational panel design.

The total sample of interest is formed by the union of all the cross-sectional samples being compared or aggregated. Using as basis the common structure of this total sample, a set of JRR replications is defined in the usual way. Each replication is formed such that when a unit is to be excluded in its construction, it is excluded simultaneously from every wave where the unit appears. For each replication, the required measure is constructed for each of the cross-sectional samples involved, and these measures are used to obtain the required averaged measure *for the replication*. Variance of the statistic of interest is then estimated from the replication estimates in the usual way.

Software developed

SAS and R programs have been implemented for the estimation.

The logic of the programs is the following:

- ♦ Estimate from the longitudinal data set: *a* and m (for the two years panel), p_1, p_2, m_1 and m_2.
- ♦ Estimate from the two cross sectional datasets (once the appropriate structure, STRATA and PSUs, are constructed): est_1, $s.e._1$, n_1, est_2, $s.e._2$, $n_{2.}$

Concluding remark

Appropriate coding of the sample structure, in the survey micro-data and accompanying documentation, is an essential requirement in order to compute sampling errors taking into account the actual sample design.

Lack of information on the sample structure in survey data files is a long-standing and persistent problem in survey work, and unfortunately affects EU-SILC as well.

Indeed, the major problem in computing sampling errors for EU-SILC is the lack of sufficient information for this purpose in the micro-data available to researchers. We have developed approximate procedures in order to reduce these limitations, and used them to produce useful estimates of sampling errors.

Because of the limited information on sample structure included in the micro-data available to researchers, direct and complete computation of variances cannot be done in many cases.

Decomposition of variances and design effects identifies more ‘portable’ components, which may be more easily imputed (carried over) from a situation where they can be computed with the given information, to another situation where such direct computations are not possible. On this basis valid estimates of variances can be produced for a wider range of statistics, thus at least partly overcoming the problem due to lack of information on sample structure.

Statistic Software: SAS or R?

1. Two different generations
2. How to evaluate their performance
3. Difficulties with discussing code properties

How to choose?

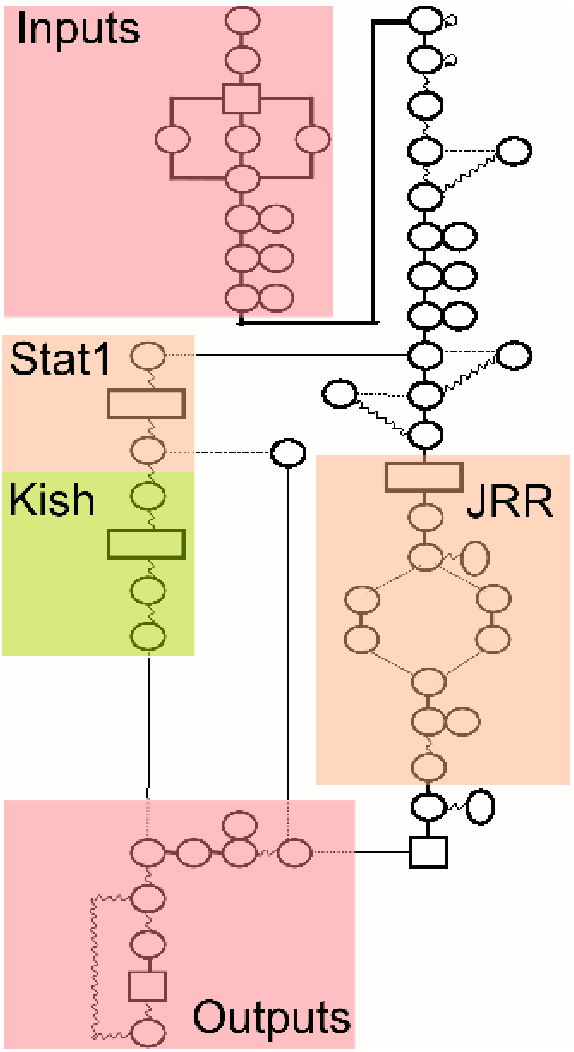
AcaStat, ADaMSoft, Analyse-it, Auguri, Autobox, BioStat, BMDP, BrightStat, Dataplot, EasyReg, Epi Info, EViews, GAUSS, Golden Helix, GraphPad Prism, gretl, JMP, JHepWork, MacAnova, Maple, Matlab, Mathematica, MedCalc, Minitab, modelQED, NCSS, NMath Stats, NumXL, OpenEpi, Origin, Partek, Primer, ProFicient, PSPP, R Commander, R, RATS, RKWard, Sage, SalStat, SAS, SHAZAM, SOCR, SOFA Statistics, SPlus, SPSS, Stata, Statgraphics, STATISTICA, Statistix, StatIt, STATPerl, StatPlus, StatsDirect, SYSTAT, Total Access Statistics, UNISTAT, The Unscrambler, VisualStat, Winpepi, WinSPC, XLStat, XploRe

Factors of interest:

1. Functionality
2. Flexibility
3. Sharing
4. Costs
5. License

The SAS Code:

1.General structure
2.DATA and PROC steps
3.Sub-structure operations
4.Comparison problems
5.Computation time issues



References

- Betti, G., Gagliardi, F., Nandi, T.: Jackknife variance estimation of differences and averages of poverty measures. Working Paper no° 68/2007, DMQ, Università di Siena (2007).

- Kish, L.: Methods for design effects. J. Official Statist. 11, 55-77 (1995).

- Verma, V., Betti, G.: Cross-sectional and Longitudinal Measures of Poverty and Inequality: Variance Estimation using Jackknife Repeated Replication. Conference 2007 ‘Statistics under one Umbrella’, Bielefeld University (2007).

- Verma, V., Betti, G., Natilli, M., Lemmi, A.: Indicators of social exclusion and poverty in Europe’s regions. Working Paper no° 59/2006, DMQ, Università di Siena (2006).

- Verma, V., Gagliardi, F., Ferretti, C.: On pooling of data and measures. Working Paper no° 84/2009, DMQ, Università di Siena (2009).

- Verma, V., Gagliardi, F., Ferretti, C.: Cumulation of poverty measures to meet new policy needs. Presented at SIS, Padova (2010).