SAMPLE
Small Area Methods for Poverty and Living Condition Estimates

SAMPLE DELIVERABLE 19

POOLED ESTIMATES OF INDICATORS

Grant agreement No:	SSH - CT - 2007 – 217565
Project Acronym:	SAMPLE
Project Full title:	Small Area Methods for Poverty and Living Conditions Estimates
Funding Scheme:	Collaborative Project - Small or medium scale focused research project
Deliverable n.	19
Deliverable name:	Pooled Estimates of Indicator
WP no.:	WP 1.3
Lead beneficiary:	2
Nature:	Report
Dissemination level: PU	PU
Due delivery date from Annex I:	31 October 2010
Actual delivery date:	10 November 2010
Project co-ordinator name:	Mrs. Monica Pratesi
Title:	Associate Professor of Statistics - University of Pisa
Organization:	Department of Statistics and Mathematics Applied to Economics of the University of Pisa (UNIPI-DSMAE)
Tel:	+39-050-2216252, +39-050-2216492
Fax:	+39-050-2216375
E-mail:	coordinator@sample-project.eu
Project website address:	www.sample-project.eu

POOLED ESTIMATES OF INDICATORS

Project Acronym:	SAMPLE
Project Full title:	Small Area Methods for Poverty and Living Conditions Estimates
Project/Contract No:	EU-FP7-SSH-2007-1
Grant agreement No	217565
Work Package 1:	New indicators and models for inequality and poverty with attention to social exclusion, vulnerability and deprivation
Document title:	Pooled estimates of indicators
Date:	10 November, 2010
Type of document:	Deliverable D19
Status	Final
Editors:	Achille Lemmi, Vijay Verma
Authors	Achille Lemmi (lemmi@unisi.it), CRIDIRE - UNISI Vijay Verma (verma@unisi.it), CRIDIRE - UNISI Gianni Betti (betti2@unisi.it), CRIDIRE - UNISI Laura Neri (neri@unisi.it), CRIDIRE - UNISI Francesca Gagliardi (gagliardi10@unisi.it), CRIDIRE – UNISI Giulio Tarditi (giuliotarditi@gmail.com), CRIDIRE – UNISI Caterina Ferretti (ferretti@ds.unifi.it), CRIDIRE - UNIFI

Contents

POOLED ESTIMATES OF INDICATORS

1. 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 for monitoring trends. 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 paper addresses some statistical aspects relating to improving the sampling precision of such indicators for subnational regions in EU countries (Verma *et al*., 2006), in particular through the cumulation of data over rounds of regularly repeated national surveys (Verma, Gagliardi and Ferretti, 2009). The reference data for this purpose are based on EU Statistics on Income and Living Conditions (EU-SILC), which is the major source of comparative statistics on income and living conditions in Europe. EU-SILC covers data and data sources of various types: cross-sectional and longitudinal; household-level and person-level; on income and social conditions; and from registers and interview surveys depending on the country. A standard integrated design has been adopted by nearly all EU countries. It involves a rotational panel in which a new sample of households and persons is introduced each year to replace one quarter 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 various durations. Two types of measures can be so constructed at the regional level by aggregating information on individual elementary units: average measures such as totals, means, rates and proportions constructed by aggregating or averaging individual values; and distributional measures, such as measures of variation or dispersion among households and persons in the region. Average measures are often more easily constructed or are available from alternative

sources. Distributional measures tend to be more complex and are less readily available from sources other than complex surveys; at the same time, such measures are more pertinent to the analysis of poverty and social exclusion. An important point to note is that, more than at the national level, many measures of averages can also serve as indicators of disparity and deprivation when seen in the regional context: the dispersion of regional means is of direct relevance in the identification of geographical disparity. Survey data such as from EU-SILC can be used in different forms or manners to construct regional indicators.

(1) Direct estimation from survey data – in the same way as done normally at the national level – provided that the regional sample sizes are adequate for the purpose.

(2) Constructing alternative (but with a substantively similar meaning) indicators 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 – which are larger in size but less detailed in content than survey data – in order to produce improved estimates using small area estimation (SAE) techniques.

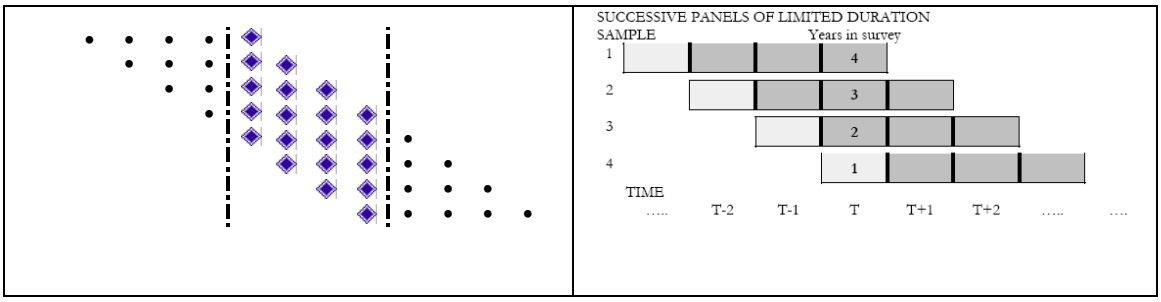
(5) Going altogether beyond the survey by exploiting administrative and other sources.

2. Cumulation over waves in a rotational panel design

The two most important regular social surveys in the EU are the Labour Force Survey (EU-LFS) and Statistics on Income and Living Conditions (EU-SILC). The EU-LFS was initiated at EU level in 1960, with a systematic common framework adopted from 1983. It is a large sample survey, conducted in all EU countries on a continuous basis, providing quarterly and annual results on labour participation along with socio-demographic and educational variables. Annually ad-hoc modules dedicated to specific topics supplement the core survey. The EU-SILC was launched starting from 2003 in some countries; it covered 27 EU and EFTA countries by 2005, and all 30 by 2008. In each country it involves an annual survey with a rotational panel design. Its content is comprehensive, focusing on income, poverty and living conditions.

Both EU-LFS and EU-SILC involve comprehensiveness in the substantive dimension (coverage of different topics), in space (coverage of different countries), and in time

(regular waves or rounds). EU-LFS involves diverse types of rotational designs; a simple and common one is illustrated below on the left hand side. In this example, a sample address stays in the survey for 5 consecutive quarters before being dropped. The subsamples contributing to a particular year have been identified in the central part of the diagram. By contrast, for EU-SILC most countries use the standard rotational household panel design shown below on the right. Here the survey is annual, and each panel stays in the survey for four consecutive years.



3. Pooling of data versus pooling of estimates

When two or more data sources contain – for the same type of units such as households or persons – a set of variables measured in a comparable way, then the information may be pooled either (a) by combining estimates from the different sources, or (b) by pooling data at the micro level. Technical details and relative efficiencies of the procedures depend on the situation. The two approaches may give numerically identical results, or the one or the other may provide more accurate estimates; in certain cases, only one of the two approaches may be appropriate or feasible in any case.

Consider for instance the common case of pooling results across countries in a multi-country survey programme such as EU-SILC or EU-LFS. For linear statistics such as totals, pooling individual country estimates say ϕ_i with some appropriate weights P_i gives the same result as pooling data at the micro level with unit weights w_{ij} rescaled as $w'_{ij} = w_{ij}.(P_i/\Sigma w_{ij})$. For ratios of the form $\phi_i = \Sigma w_{ij}.v_{ij}/\Sigma w_{ij}.u_{ij}$, the two forms give very similar but not identical results, corresponding respectively to the 'separate' and 'combined' types of ratio estimate.

This paper is concerned with a different but equally common type of problem, namely pooling of different sources pertaining to the same population or largely overlapping and similar populations. In particular, the interest is in pooling over survey waves in a

national survey in order to increase the precision of regional estimates. Estimates from samples from the same population are most efficiently pooled with weights in proportion to their variances (meaning, with similar designs, in direct proportion to their sample sizes). Alternatively, the samples may be pooled at the micro level, with unit weights inversely proportion to their probabilities of appearing in any of the samples. This latter procedure may be more efficient (e. g., O'Muircheataigh and Pedlow, 2002), but be impossible to apply as it requires information, for every unit in the pooled sample, on its probability of selection into each of the samples irrespective of whether or not the unit appears in the particular sample (Wells, 1998). Another serious difficulty in pooling samples is that, in the presence of complex sampling designs, the structure of the resulting pooled sample can become too complex or even unknown to permit proper variance estimation. In any case, different waves of a survey like EU-SILC or EU-LFS do not correspond to exactly the same population. The problem is akin to that of combining samples selected from multiple frames, for which it has been noted that micro level pooling is generally not the most efficient method (Lohr and Rao, 1996).
For the above reasons, pooling of wave-specific estimates rather than of micro data sets is generally the appropriate approach to aggregation over time from surveys such as EU-SILC.

4. Gain in precision from cumulation over survey waves

Consider that for each wave, a person's poverty status (poor or non-poor) is determined based on the income distribution of that wave separately, and the proportion poor at each 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. For this purpose, the JRR variance estimation methodology can be easily extended on the following lines. 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, from which variance is then estimated in the usual way (Betti *et al*., 2007).

Table 1: Gain from cumulation over two waves: cross-sectional and persistent poverty rates. Poland EU-SILC 2005-2006

Sample base	Poverty rate	Est	n persons	%se* actual		mean income	HCR: poverty line national	regional
CS-2006	HCR 2006	19.1	45,122	0.51	**(1)**	0.42	0.34	0.40
CS-2005	HCR 2005	20.6	49,044	0.45	**(2)**	1.31	1.18	1.18
LG 05-06	HCR 2006	18.5	32,820		**(3)**	0.55	0.40	0.47
LG 05-06	HCR 2005	20.2	32,820		**(4)**	0.60	0.48	0.56
LG 05-06	Persistent '05-06	12.5	32,820		**(5)**	14%	30%	30%

In terms of the quantities defined above, rows (1)-(5) of Table 1 are as follows.

Standard error of average HCR over two years (assuming independent samples)	$(1) \quad = \quad 1/2.(V_1 + V_2)^{1/2}$
Factor by which standard error is increased due to positive correlation between waves	$(2) = .(1 + b.(n/n_H))^{1/2}$
Standard error of average HCR over two years (given correlated samples)	$(3) = (1).(2) = .(\mathrm{V})^{1/2}$
Average standard error over a single year	$(4) = \frac{(V_1)^{1/2} + (V_2)^{1/2}}{2}$
Average gain in precision (variance reduction, or increase in effective sample size, over a single year sample)	$(5) = 1 - ((3)/(4))^2$

Gain from cumulation over two waves. Results for Italy, Poland and Czech Republic.

	Italy EU-SILC 2007-2008	**Poland** EU-SILC 2005-2006	**Czech Republic** EU-SILC 2005-2006
$(1) \quad = \quad \frac{(V_1 + V_2)^{1/2}}{2}$	0.36	0.34	0.43
$(2) = (1 + b \cdot (n/n_H))^{1/2}$	1.20	1.18	1.18
$(3) = (1) \cdot (2)$	0.43	0.40	0.51
$(4) = \frac{(V_1)^{1/2} + (V_2)^{1/2}}{2}$	0.50	0.48	0.61
$(5) = 1 - ((3)/(4))^2$	26%	30%	30%

In place of the full JRR application, it is more illuminating to provide here the following simplified procedure for quantifying the gain in precision from averaging over waves of the rotational panel. It illustrates the statistical mechanism of how the gain is achieved. Indicating by pj and p'j the (1, 0) indicators of poverty of individual j over the two adjacent waves, we have the following for the population variances:

$\text{var}(p_j) = \Sigma(p_j - p)^2 = p.(1-p) = v$; similarly, $\text{var}(p_j') = p'.(1-p') = v'$

$\text{cov}(p_j, p_j') = \Sigma(p_j - p)(p_j' - p') = a - p.p' = c_1$, say,

where 'a' is the persistent poverty rate over the two years. For the simple case where the two waves completely overlap and $p' = p$, variance v_A for the averaged measure is: $v_A = \frac{v}{2}.(1+b)$, with correlation $b = \left(\frac{c_1}{v}\right) = \left(\frac{a-p^2}{p-p^2}\right)$. The correlation between two periods is expected to decline as the two become more widely separated. Consider, for example, the case when the correlation between two points k waves apart can be approximated as $(c_k/v) = (c_1/v)^k$. In a set of K periods there are (K-k) pairs exactly k periods apart, k=1 to (K-1). It follows that variance v_K of an average over K periods relates to variance v of the estimate from a single wave as:

$$f_c = \left(\frac{v_k}{v}\right) = \frac{1}{K}.\left(1 + 2.\sum_{k=1}^{K-1}\left(\frac{K-k}{K}\right).\left(\frac{c_1}{v}\right)^k\right)$$

where a, the persistent poverty between pairs of adjacent waves, and p, the cross-sectional poverty rate, are averages over the waves involved. For application to pairs of waves in EU-SILC, it is necessary to allow for variations in cross-sectional sample sizes and partial overlaps. The result is:

$$V = (V_1 + V_2)/4.(1 + b.(n/n_H))$$

where V_1 and V_2 are the sampling variances, b the correlation coefficient over the two cross-sections, n is the overlap between the cross-sectional samples, and n_H is the harmonic mean of their sample sizes n_1 and n_2.

The methodology described above was applied to the 2005-2006 cross-sectional and longitudinal EU-SILC samples for Poland. Table 1 shows some results at the national level. Averaging the HCR over two waves leads to a variance of this averaged estimator that is 30% less than the variance of the HCR estimated from just a single wave.

Consider a rotational sample in which each unit stays in the sample for n consecutive periods, with the required estimate being the average over Q consecutive periods, such as Q=4 quarters for annual averages. The case n=1 corresponds simply to independent samples each quarter. Under the simplifying assumption of uniform variances, variance of the estimate of average over Q period is $V_a^2 = V^2/Q$.

In the general case, the total sample involved in the estimation consists of (n+Q-1) independent subsamples. These correspond to the rows in the figures below. Each

subsample is 'observed' over a certain number of consecutive periods within the interval (Q) of interest.[1] In principle, for a given subsample the sample cases involved in these 'observations' are fully overlapping. The distribution of the (n+Q-1) subsamples according to the number of observation (m) provided is:

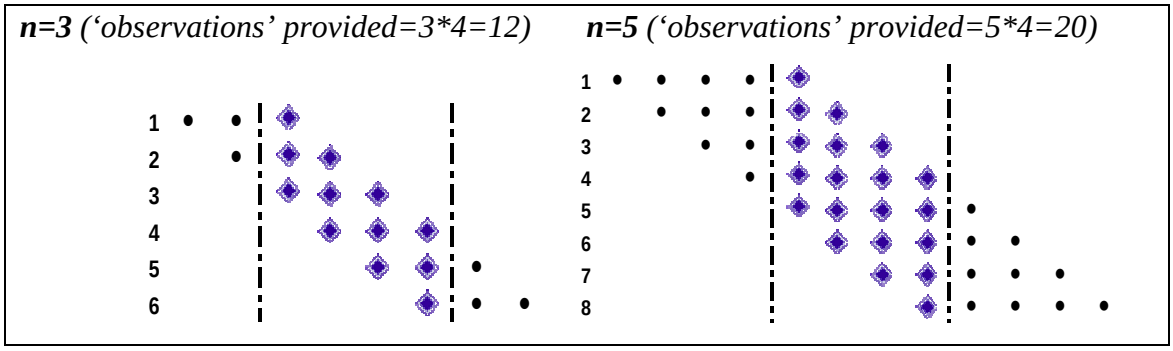
No. of observations (m)→	*provided by no. (x) of subsamples*	*Total no. of 'observations' provided by all subsamples*
m = 1, 2, …, (m_1-1)	*x = 2 for each value of m*	$\sum_{i=1}^{(m_1-1)} 2i = (m_1-1)\cdot m_1$
m = m_1	*x = m_2-(m_1-1)*	$m_1 \cdot m_2 - (m_1-1)\cdot m_1$
Total →	*no. of sublamples equal to* $2\cdot(m_1-1)+m_2-(m_1-1)=$ $=m_2+m_1-1=n+Q-1$	*no. of observations equal to* $m_1 \cdot m_2 = n.Q$

where m_1=min(n, Q) and m_2=max(n, Q).

Note that the total number of 'observations' provided by all subsamples over interval Q is $m_1 \cdot m_2 = n \cdot Q$. This is consistent with the fact that, obviously, there are n subsamples observed at each of the Q periods in the interval being considered (see diagrams below).

Q=4

n=3 *('observations' provided=3*4=12)* ***n=5*** *('observations' provided=5*4=20)*



Note: The numbers on the left side of the figures represent the number of subsamples (n+Q-1).

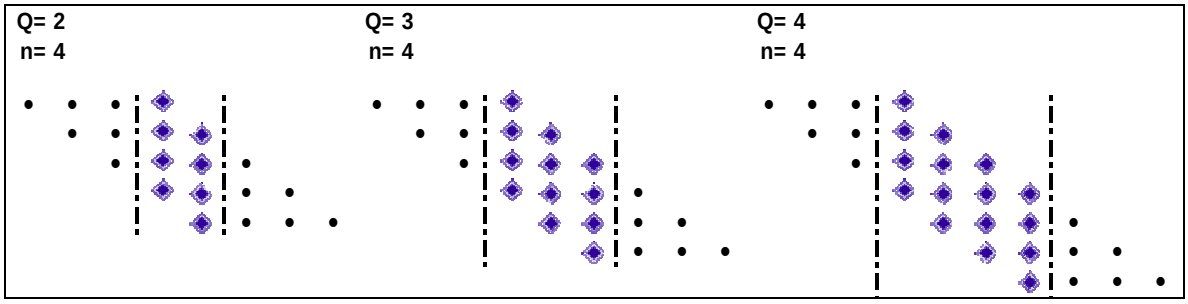
[1] For 'observation' we mean surveying one subsample on one occasion. These correspond to individual diamonds in the figures below.

For illustration, consider $Q=m_1=4$, $n=m_2=5$. There are 2 contributing subsamples for each number 1, 2 and $(m_1-1)=3$ of observations; and in addition there are $m_2-(m_1-1)=2$ subsamples each contributing $m_1=4$ observations.

Similarly, for $Q=m_2=4$, $n=m_1=3$, we have 2 contributing subsamples for each number 1 and $(m_1-1)=2$ of observations, and in addition $m_2-(m_1-1)=2$ subsamples each contributing $m_1=3$ observations.

In the EU-SILC survey in most countries, n is always equal to 4 (each survey rounds is made of 4 subsamples), and at the present stage Q could be equal to 2 (years 2003-2004), 3 (years 2003-2004-2005) and 4 (years 2003-2004-2005-2006).

So the previous figure could be adapted as follow:



In order to provide a simplified formulation of the effect of correlation arising from sample overlaps, we assume the following model. If R is the average correlation between estimates from overlapping samples in adjacent periods (as defined above), then between points one period apart (e.g. between the 1st and 3rd quarters), the average correlations is reduced to R^2, the correlation between points two periods apart (e.g. the 1st and the 4th quarters) is reduced to R^3, and so on.

Consider a subsample contributing m observations during the interval (Q) of interest with full sample overlap. Considering all the pairs of observations involved and the correlations between them under the model assumed above, variance of the average over the m observations is given by:

$$V_m^2 = \frac{V^2}{m} \cdot \left(1 + f(m)\right)$$

Where:

$$f(m)=\frac{2}{m}\cdot\left\{(m-1)\cdot R+(m-2)\cdot R^2+...+R^{m-1}\right\}$$

The term $V_m^2\Big/\left(\frac{V^2}{m}\right)=1+f(m)$ reflects the loss in efficiency in cumulation or averaging over overlapping samples, compared to cumulation over entirely independent samples. The following illustrates its values for various values of m:

m	*f(m)*
2	R
3	$\frac{2}{3}(2R+R^2)$
4	$\frac{2}{4}(3R+2R^2+R^3)$
5	$\frac{2}{5}(4R+3R^2+2R^3+R)$

Repeated observations over the same sample are less efficient in the presence of positive correlations (R). The loss depends on the number of repetitions (m) and is summarised by the factor *[1+f(m)]*.

In estimating the average using the whole available sample of $(n\cdot Q)$ subsample observations[2], we may simply give each observation the same weight. Taking into account the number of observations and the variances involved, the resulting variance of the average becomes:

$$V_a^2=\left(\frac{V^2}{n\cdot Q}\right)\cdot\left\{m_1\cdot\left[m_2-(m_1-1)\right]\cdot\left[1+f(m_1)\right]+2\sum_{m=1}^{m_1-1}m\cdot\left[1+f(m)\right]\right\}\Big/(n\cdot Q)=\left(\frac{V^2}{n\cdot Q}\right)\cdot F(R)$$

The first factor is the variance to be expected from $(n\cdot Q)$ independent observations (with no sample overlaps or correlation), each observation with variance V^2. The other terms are the effect of correlation with sample overlaps. This effect, *F(R)* disappears

[2] Obviously, we have n subsamples observed during each of Q periods in the rotational design assumed.

when *f(i)=0* for all *i=1 to m* (which will be the case of *R=0*), as can be verified in the above expression.

5. Variance and design effects

The issues addressed concern the efficiency of (3) in section 1 – cumulating information over consecutive waves of a survey such as EU-SILC, involving complex statistics based on complex sample designs. Estimates are required for the whole population and also for subpopulations of different types. Both cross-sectional and longitudinal statistics are involved. Comparisons and cumulation over correlated cross-sections, with which this paper is concerned, add another layer of complexity.

Jackknife Repeated Replication (JRR) provides a versatile and straightforward technique for variance estimation in these situations. It is one of the classes of variance estimation methods based on comparisons among replications generated through repeated re-sampling of the same parent sample. Once the set of replications has been appropriately defined for any complex design, the same variance estimation algorithm can be applied to a statistic of any complexity. 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 rotational panel design of EU-SILC (Verma and Betti, 2007). 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 overcome these limitations at least partially, and used them to produce useful estimates of sampling errors (Verma, Betti and Gagliardi, 2010). Use has been made of these results in this paper, but it is not possible here to go into detail concerning them.

A most useful concept for the computation, analysis and interpretation of sampling errors concerns 'design effect' (Kish, 1995). Design effect is the ratio of the variance (v) under the given sample design, to the variance (v_0) under a simple random sample of the same size: $d^2 = v/v_0$, $d = se/se_0$. Proceeding from estimates of sampling error to estimates of design effects is essential for understanding the patterns of variation in and the determinants of magnitude of the error, for smoothing and extrapolating the results of computations, and for evaluating the performance of the sampling design.

Analysis of design effects into components is also needed in order to understand from where inefficiencies of the sample arise, to identify patterns of variation, and through that to extend the results to other statistics, designs and situations. And most importantly, with JRR (and other replication methods) the total design effect can only be estimating by estimating (some of) its components separately (Verma, Betti, 2010). In applications for EU-SILC, there is in addition a most important and special reason for decomposing the total design effect into its components. 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. We may decompose total variance v (for the actual design) into the components or factors as $v = v_0.d^2 = v_0.(d_W.d_H.d_D.d_X)^2$, where d_W is the effect of sample weights, d_H of clustering of individual persons into households, d_D of clustering of households into dwellings, and d_X that of other complexities of the design, mainly clustering and stratification. All factors other than d_x do not involve clusters or strata, but depend only on individual elements (households, persons etc.), and the sample weight associated with each such element in the sample. Parameter d_W depends on variability of sample weights, and secondly also on the correlation between the weights and the variable being estimated; d_H is determined by the number of and correlation among relevant individuals in the household, and similarly d_D by the number of households per dwelling in a sample of the latter. By contrast, factor d_X represents

the effect on sampling error of various complexities of the design such as multiple stages and stratification. Hence unlike other components, d_X requires information on the sample structure linking elementary units to higher stage units and strata. This effect can be estimated as follows using the JRR procedures. We compute variance under two assumptions about structure of the design: variance v under the actual design, and v_R computed by assuming the design to be (weighted) simple random sampling of the ultimate units (addresses, households, persons as the case may be). This can be estimated from a 'randomised sample' created from the actual sample by completely disregarding its structure other than the weights attached to individual elements. This gives $(d_x)^2 = (v/v_R)$, with $v_R = v_0.(d_W.d_H.d_D)^2$.

Table 2 gives standard error, design effect and components of design effect for the cross-sectional 2006 EU-SILC sample for Poland. The sample was a two stage stratified sample of dwellings containing 45,122 individual persons. With "%se" (3rd and last column) we mean: for mean statistics e.g. equivalised disposable income – standard error expressed as percentage of the mean value; for proportions and rates (e.g. poverty rates) – standard error given as absolute percent points. Terms (%se actual) and (%se SRS) relate, respectively, to the variances v and v0 in the text. Parameter d_D cannot be estimated separately because of lack of information, but its effect is small and is, in any case, already incorporated into overall design effect d.

Table 2: Estimation of variance and design effects at the national level. Cross-sectional sample. Poland EU-SILC 2006

	Est.	%se actual	Design effect d_X	d_W	d_H	d	%se SRS
(1) Mean equivalised disposable income	3,704	0.57	0.94	1.22	1.74	1.99	0.29
(2) HCR – 'head count' or poverty rate, using national poverty line	19.1	0.51	1.02	1.09	1.74	1.94	0.26
(3) HCR – 'head count' or poverty rate, using regional (NUTS1) poverty line	19.0	0.61	1.05	1.09	1.74	1.99	0.30

Table 2 gives poverty rates defined with respect to two different 'levels' of poverty line: country level and NUTS1 level. By this we mean the population level to which the income distribution is pooled for the purpose of defining the poverty line. Conventionally poverty rates are defined in terms of the country poverty line (as 60% of the national median income). The income distribution is considered at the country level, in relation to which a poverty line is defined and the number (and proportion) of poor

computed. It is also useful to consider poverty lines at other levels. Especially useful for constructing regional indicators is the use of regional poverty lines, i.e. a poverty line defined for each region based only on the income distribution within that region. The numbers of poor persons identified with these lines can then be used to estimate regional poverty rates. They can also be aggregated upwards to give an alternative national poverty rate – but which still remains based on the regional poverty lines. So defined, the poverty measures are not affected by disparities in the mean levels of income among the regions. The measures are therefore more purely relative.

6. Illustrative applications of cumulation at the regional level

Table 3 shows results for the estimation of variance and design effect for the cross-sectional 2006 and 2005 Poland datasets. The results at national level for the three measures considered have been already presented in the previous section. Here we present the results at NUTS1 regional level. All the values, except "%se SRS" and d_X, are computed at regional level in the same manner as the national level. All factors other than d_X do not involve clusters or strata, but essentially depend only on individual elements and the associated sample weights. Hence normally they are well estimated, even for quite small regions. Factor $d_{X(G)}$ for a region (G) may be estimated in relation to $d_{X(C)}$ estimated at the country (C) level on the following lines. For large regions, each with a large enough number of PSUs (say over 25 or 30), we may estimate the variance and hence $d_{X(G)}$ directly at the regional level. Sometimes a region involves a SRS of elements, even if the national sample is multi-stage in other parts; here obviously, $d_{X(G)}=1$. If the sample design in the region is the same or very similar to that for the country as a whole – which is quite often the case – we can take $d_{X(G)}=d_{X(C)}$. It is common that the main difference between the regional and the total samples is the average cluster size (b). In this case we use $d^2_{X(G)}=1+\left(d^2_{X(C)}-1\right).b_{(G)}/b_{(C)}$. The last-mentioned model concerns the effect of clustering and hence is meaningful only if $d_{X(C)}\geq 1$, which is often but not always the case in actual computations. Values smaller than 1.0 may arise when the effect of stratification is stronger than that of clustering, when units within clusters are negatively correlated (which is rare, but not impossible), or simply as a

result of random variability in the empirical results. In any case, if $d_{x(C)} < 1$, the above equation should be replaced by $d_{x(G)} = d_{x(C)}$. The quantity (%se* SRS) can be directly computed at the regional level as was done for the national level in Table 1. However, very good approximation can be usually obtained very simply without involving JRR computations of variance. The following model has been used in Table 3. For means (such as equivalised income) over very similar populations, assumption of a constant coefficient of variation is reasonable. The region-to-country ratio of relative standard errors (expressed as percentage of the mean value as in Table 3) under simple random sampling is inversely proportional to the square-root of their respective sample sizes: $(\%se*SRS)^2_{(G)} = (\%se*SRS)^2_{(C)} \cdot (n_{(C)}/n_{(G)})$. For proportions (p, with q=100-p), with standard error expressed in absolute percent points as in Table 3, we can take: $(\%se*SRS)^2_{(G)} = (\%se*SRS)^2_{(C)} \cdot \left(\frac{p_{(G)} \cdot q_{(G)}}{p_{(C)} \cdot q_{(C)}}\right) \cdot (n_{(C)}/n_{(G)})$. A poverty rate may be treated as proportions for the purpose of applying the above. We see from Table 3 that the (%se*actual) at regional level is generally, for all the three measures, 2 to 3 times larger than that at the national level.

Table 3: Estimation of variance and design effects at the regional (NUTS1) level. Full cross-sectional dataset

	2006 Est.	n persons	%se* SRS	d_X	d	%se* actual	**2005** Est.	n persons	%se* actual
Mean equivalised disposable income									
Poland	3,704	45,122	**0.29**	**0.94**	1.99	0.57	3,040	49,044	0.62
PL1	4,236	8,728	0.65	0.94	2.06	1.34	3,455	9,871	1.32
PL2	3,889	9,273	0.63	0.94	1.78	1.13	3,143	10,181	1.22
PL3	3,162	9,079	0.64	0.94	2.00	1.28	2,618	9,674	1.32
PL4	3,530	6,912	0.73	0.94	1.90	1.39	2,977	7,195	1.84
PL5	3,906	4,538	0.90	0.94	1.96	1.77	3,164	5,066	1.85
PL6	3,419	6,592	0.75	0.94	1.90	1.43	2,816	7,057	1.58
At-risk-of-poverty rate, national poverty line									
Poland	19.1	45,122	**0.26**	1.02	1.94	0.51	20.6	49,044	0.45
PL1	17.1	8,728	0.57	1.02	1.85	1.06	19.1	9,871	0.92
PL2	14.7	9,273	0.52	1.02	1.86	0.97	16.4	10,181	0.87
PL3	25.2	9,079	0.64	1.02	2.09	1.34	25.2	9,674	1.13
PL4	18.7	6,912	0.66	1.02	1.98	1.32	20.2	7,195	1.19
PL5	18.6	4,538	0.82	1.02	1.91	1.56	20.2	5,066	1.43
PL6	21.4	6,592	0.71	1.02	1.95	1.40	23.7	7,057	1.26
At-risk-of-poverty rate, regional poverty lines									
Poland	19.0	45,122	**0.30**	1.05	1.99	0.61	20.5	49,044	0.51
PL1	19.8	8,728	0.70	1.04	1.90	1.34	20.9	9,871	1.07
PL2	18.5	9,273	0.67	1.04	1.91	1.27	19.0	10,181	1.05
PL3	18.6	9,079	0.68	1.06	2.14	1.45	20.8	9,674	1.21
PL4	17.5	6,912	0.76	1.05	2.04	1.54	20.1	7,195	1.35
PL5	20.9	4,538	1.00	1.04	1.97	1.96	22.2	5,066	1.68
PL6	19.1	6,592	0.80	1.05	2.00	1.60	21.3	7,057	1.37

Regional HCR estimates based on the national poverty line are quite different from those based on the regional ones. Also, while individual regional estimates of HCR using the regional poverty line are quite close to the national estimate (19.0 for 2006), the ones using the national poverty line are more variable (from 14.7 to 25.2 for 2006).

From the previous table it can be seen that generally for the HCR measures, both for country and NUTS1 level poverty lines, cumulating the estimates over two waves leads to a reduction of 30% in variance compared to that for a single wave. This reduction of the variance is smaller for mean equivalised income due to a higher correlation between incomes for the two years – generally the coefficient of correlation of the equivalised income between waves exceeds 0.70.

Table 4: Gain in precision from averaging over correlated samples. Poland NUTS1 regions

Mean equivalised income

	Country	PL1	PL2	PL3	PL4	PL5	PL6
(1)	0.42	0.94	0.83	0.92	1.15	1.28	1.07
(2)	1.31	1.33	1.30	1.31	1.27	1.32	1.32
(3)	0.55	1.26	1.08	1.20	1.47	1.70	1.41
(4)	0.60	1.33	1.17	1.30	1.62	1.81	1.51
(5)	14%	11%	15%	14%	18%	12%	12%

HCR national poverty line

	Country	PL1	PL2	PL3	PL4	PL5	PL6
(1)	0.34	0.70	0.65	0.88	0.89	1.06	0.94
(2)	1.18	1.18	1.17	1.18	1.18	1.17	1.19
(3)	0.40	0.83	0.76	1.03	1.05	1.23	1.12
(4)	0.48	0.99	0.92	1.24	1.26	1.50	1.33
(5)	30%	29%	31%	30%	30%	32%	29%

HCR regional poverty line

	Country	PL1	PL2	PL3	PL4	PL5	PL6
(1)	0.40	0.86	0.83	0.94	1.03	1.29	1.05
(2)	1.18	1.18	1.18	1.17	1.18	1.17	1.18
(3)	0.47	1.02	0.98	1.10	1.21	1.51	1.24
(4)	0.56	1.21	1.16	1.33	1.45	1.82	1.49
(5)	30%	29%	29%	31%	30%	31%	31%

Rows (1) – (5) have been defined in Table 1.

References

1. 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).
2. Kish, L.: Methods for design effects. *J. Official Statist*. 11, 55-77 (1995).
3. Lohr, S. L., Rao, J,N.K.: Inference from dual frame surveys. *Journal of American Statistical Association*, 95, 271-280 (2000).
4. O'Muircheataigh, C., Pedlow, S.: Combining samples vs. cumulating cases: a comparision of two weighting strategies in NLS97. *American Statistical Association Proceedings of the Joint Statistical Meetings*, pp. 2557-2562 (2002).
5. 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).
6. Verma, V., Betti, G.: Taylor linearization sampling errors and design effects for poverty measures and other complex statistics, *Journal of Applied Statistics* (2010).
7. Verma, V., Betti, G., Gagliardi, F.: An assessment of survey errors in EU-SILC, Eurostat Methodologies and Working Papers, Eurostat, Luxembourg (2010).
8. 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).
9. Verma, V., Gagliardi, F., Ferretti, C.: On pooling of data and measures. Working Paper no° 84/2009, DMQ, Università di Siena (2009).
10. Wells, J. E.: Oversampling through households or other clusters: comparison of methods for weighting the oversample elements. *Australian and New Zeeland Journal of Statistics*, 40, 269-277.