Approximate variance estimates of poverty measures for the application of EBLUP for small-area estimation

Gianni Betti and Vijay Verma

Working Paper n. 57, April 2006

Approximate variance estimates of poverty measures for the application of EBLUP for small-area estimation

Gianni Betti and Vijay Verma

Abstract

This paper reports a piece of research undertaken during the project “Regional Indicators to reflect social exclusion and poverty VT/2003/43” funded by the Employment and Social Affairs DG of the European Commission.
This study takes as its point of departure the methodological framework used for defining the indicators of poverty and social exclusion endorsed at Laeken, and more generally, existing methodological research and data in the area of indicators of poverty and social exclusion as well as in the area of regional indicators.
The strategy recommend for the construction of regional (subnational) indicators of poverty and deprivation has three fundamental aspects: (a) making the best use of available sample survey data, such as by cumulating and consolidating the information so as to obtain more robust measures which permit greater spatial disaggregation; (b) exploiting to the maximum ‘meso’ data - such as the highly disaggregated tabulations available in Eurostat Free Data Dissemination (NewCronos) - for the purpose of constructing regional indicators; and (c) using the two sources in combination to produce more precise estimates for regions using appropriate small area estimation (SAE) techniques.
The present work lies within the third aspect described above; in fact, when estimating small area models, the dependent variables are direct estimates of the measure concerned on the basis of sample data from the small area concerned. These direct estimates are normally affected by large variation mainly due to small sub-sample sizes; in this paper we propose how to calculate these variances in the case of EBLUP (Empirical Best Linear Unbiased Prediction) estimators. These variances are needed in order to combine the sample and model-based estimates to produce more precise SAEs.

Table of Contents

1. Introduction

This paper reports a piece of research undertaken during the project "Regional Indicators to reflect social exclusion and poverty VT/2003/43" funded by the Employment and Social Affairs DG of the European Commission.

As provided in the project Terms of Reference, the general aim of the project was to identify the appropriate methodologies and strategies for the development of indicators of poverty and social exclusion at regional level, the ultimate goal being the development of a coherent and integrated strategy for the incorporation of the regional dimension into the Member States' NAP/incl.[1] This study takes as its point of departure the methodological framework used for defining the indicators of poverty and social exclusion endorsed at Laeken, and more generally, existing methodological research and data in the area of indicators of poverty and social exclusion as well as in the area of regional indicators.

The strategy recommend for the construction of regional (subnational) indicators of poverty and deprivation has been summarised in Betti *et al.* (2006); it has three fundamental aspects: (a) making the best use of available sample survey data, such as by cumulating and consolidating the information so as to obtain more robust measures which permit greater spatial disaggregation; (b) exploiting to the maximum 'meso' data - such as the highly disaggregated tabulations available in Eurostat Free Data Dissemination (NewCronos) - for the purpose of constructing regional indicators; and (c) using the two sources in combination to produce more precise estimates for regions using appropriate small area estimation (SAE) techniques.

The present work lies within the third aspect described above; in fact, when estimating small area models, the dependent variables are direct estimates of the measure concerned on the basis of sample data from the small area concerned. These direct estimates are normally affected by large variation mainly due to small sub-sample sizes; in this paper we propose how to calculate these variances in the case of EBLUP (Empirical Best Linear Unbiased Prediction) estimators. These variances are needed in order to combine the sample and model-based estimates to produce more precise SAEs.

The paper is composed of six sections; after this introduction, Section 2 briefly describes the SAE models taken into account in the project and fully described in Neri and Verma (2006); Section 3, which is the bulk of the paper, presents methodological notes on the estimation of sampling errors for the disaggregated direct estimates; the methodology is implemented for the case of the head count ratio (HCR) poverty measure in Section 4 and for other related poverty statistics in Section 5. Finally, Section 6 presents an empirical illustration.

2. Description of the SAE models

There is a wide variety of SAE techniques available, and the field is rapidly expanding. The suitability and efficiency of a particular technique depends on the specific situation and on the nature of the statistical data available for the purpose.

[1] National Action Plans on Social Inclusion.

Reference must be made to the work of the 'Eurarea' project on 'Enhancing Small Area Estimation Techniques to Meet European Needs', the final report of which became available in August 2004. As stated, the aim of that project was "to provide European Statisticians, particularly government statisticians, with the information they needed to assess and use a range of small area estimation techniques, including techniques incorporating recent theoretical advances." Several classes of small area estimators were investigated and evaluated under that project, in particular: (1) direct estimators; (2) area level synthetic estimators; (3) generalised regression estimators (GREG); and (4) composite (EBLUB) estimators.

The Eurarea project did not consider, perhaps in view of practical considerations in a multi-country exercise, more complex estimation procedures, such as 'Empirical Bayes' and 'Hierarchical Bayes' approaches. All the methods considered in the Eurarea project assumed a situation in which unit level information on the target variables (in our case, various measures of poverty, deprivation and social exclusion) is provided by data from a sample survey; then there are auxiliary variables (covariates) which are known for the target areas (such as all NUTS2 areas in each EU country in our case).

2.1 The ratio approach

Returning to the present application, it should be noted that – compared to the scope of modelling considered in a large project such as Eurarea, for instance – the present application is somewhat simplistic in that it does not attempt to incorporate temporal or spatial autocorrelations. On the other hand, however, *a major positive feature of the present approach is that the modelling strategy is designed to be hierarchical*. We begin with poverty rates and other target variables at the national level, using essentially direct survey estimates without involving any modelling.[2] Modelling at the level of countries is problematic in any case because most pertinent explanatory variables able to distinguish among national patterns are likely to be institutional and historical – variables which are often too complex and almost impossible to quantify. Sometimes countries or national systems are classified into types in an attempt to capture these aspects, such as 'social democratic', 'liberal, 'corporatist', 'residual', and so on (e.g., Berthoud, 2004). For some purposes categorisation such as the above might be of some use. But generally such schemes are too simplistic to be illuminating. And we suspect that not too infrequently, such 'ideal types' are constructed merely to express or promote ideological prejudice.

We can expect the predictive power of the model at the regional level to be substantially improved when the target variables as well as the covariates are expressed in terms of their values at the preceding higher level. Thus for NUTS1 region i, all target variables and all covariates in the model are expressed in the form of the ratio $R_i = Y_i / Y_0$, where (Y_i, Y_0) refer to the actual values of the variables, respectively, for NUTS1 i and its country. In this way the effect of the difficult-to-qualify institutional and historical factors, common to the country and its regions, is abstracted. Similarly, in going from NUTS1 region i to its NUTS2 region j, we

[2] The only exception to using survey estimates directly at the national level is the consolidation we have used to reduce sampling variability by 'benchmarking' the results to certain aspects of the pattern averaged over group of countries, as explained in Neri and Verma (2006).

express the model variables in the form $R_{ij} = Y_{ij}/Y_i$; and similarly from NUTS2 to NUTS3 in the form $R_{ijk} = Y_{ijk}/Y_{ij}$.

The resulting estimates of the R values can be 'raked' for consistency across levels by ensuring $\Sigma_i W_i.R_i = 1$, $\Sigma_j W_{ij}.R_j = 1$, etc., where W_i etc. are the appropriate population weights for the regions, scaled to give $\Sigma_i W_i = 1$.

Occasionally, it may be efficient to specify this type of modelling separately for different parts of a large or exceptionally heterogeneous country, examples being eastern and western parts of Germany, or the northern and southern parts of Italy. The same may apply to metropolitan versus other areas in some countries, such as the UK and France.

The same ideas are extended to the modelling of subpopulations, such as children, old persons, single person households, etc. Consider for instance child poverty rate, say Z_0 at the national level, Z_i at NUTS1 level, and Z_{ij} at NUTS2 level, with (Y_0, Y_i, Y_{ij}) as the corresponding poverty rates for the total population. Then for NUTS1 regions, we can first model the proportion (Y_i/Y_0) as above, and then using those results to model the ratio:

$$r_i = \frac{(Z_i/Z_0)}{(Y_i/Y_0)} = \frac{(Z_i/Y_i)}{(Z_0/Y_0)}$$

to obtain the required child poverty rate Z_i for region i. Factor r_i indicates how the ratio of child to all-person poverty rate varies across regions. Similarly for NUTS2 regions we can first model the variation in all-person poverty rates (Y_{ij}/Y_i) as before, and then using those results model the ratio:

$$r_{ij} = \frac{(Z_{ij}/Z_i)}{(Y_{ij}/Y_i)} = \frac{(Z_{ij}/Y_{ij})}{(Z_i/Y_i)}.$$

2.2 The target variables

The poverty and deprivation indicators listed in Table 1 are taken as the target variables in the SAE model estimated.

Head count ratios have been computed for two poverty line levels: poverty lines defined with respect to income distribution at the country level, and with respect to income distribution separately within each NUTS2 region. All other poverty or deprivation rates (measures 5-13) have been computed with reference to HCR_C, i.e., using only country-level poverty lines. (See Betti and Verma (1999) for an exhaustive description of these measures).

Country-specific details on the availability of these variables in EU25 and Candidate countries will be provided in Section 6 (Table 5). The main point to note is that for countries other than EU15, Poland and Romania, we have no micro data available and only two of the target variables could be constructed from published data: head count ratio with country poverty line, and median equivalised income.

2.3 Models used

According to the availability of data for the target variables and the access to area-coded survey data for each country, three different types of SAE models have been estimated:

- o SAE Model 1: estimated on the ratio NUTS1/Country;
- o SAE Model 2: estimated on the ratio NUTS2/ NUTS1;
- o SAE Model 3: estimated on the ratio NUTS3/ NUTS2.

One such model has been estimated for each target variable at each NUTS level; all countries with area-coded survey data and the particular target variable available are pooled together for the estimation of model parameters at the level concerned.

Such pooling across countries is clearly an over-simplification, and has been introduced here primarily for practical reasons. Nevertheless, the 'ratio approach' described above makes this procedure quite reasonable, we believe. This is because the approach removes the effect of factors common to an area and its components at the next level of the NUTS hierarchy.

Model 3 has been estimated for Italy only, as no NUTS3 codes are available in any other survey.

In countries where no area-coded survey data are available, we have had to resolve to much simpler and cruder regression-prediction models. This procedure involves using the regression coefficients determined from the corresponding EBLUP model (for the same target variable and the same NUTS level) to predict the target variables on the basis of available predictors.

Model results need to be evaluated with reference to external criteria, as well as internally for consistency. For internal evaluation of the models, the following features should be examined: (a) linearity of the regression; (b) choice of prediction variables; (c) normality of standardised residuals; (d) homogeneity of the variance for standardised residuals; and (e) residual analysis to detect outliers.

On these diagnostic aspects, only preliminary analysis could be done within the framework and resources of the present research. Our aim has been primarily illustrative; some deeper analysis must of course be performed in real life application.

Table 1 Poverty indicators (Target Variables for SAE models)

1	HCR_C	Head Count Ratio, using country poverty lines (consolidated over computations using 50, 60 and 70% of median equivalised income);
2	HCR_N2	Head Count Ratio, using nuts2 poverty lines (consolidated over computations using 50, 60 and 70% of median equivalised income);
3	LogIncPC	Mean of logarithm of the per capita income;
4	logEqInc	Mean of logarithm of equivalised income;
5	FM_C	Fuzzy monetary poverty rate (scaled to equal HCR_C at EU15 level);
6	FS_C	Fuzzy supplementary (non-monetary) deprivation rate (scaled to equal HCR_C at EU15 level);
7	LAT_C	Latent deprivation rate;
8	MAN_C	Manifest deprivation rate.
9	FSUP-1	Fuzzy supplementary deprivation rate: dimension 1 (basic life-style);
10	FSUP-2	Fuzzy supplementary deprivation rate: dimension 2 (secondary life-style);
11	FSUP-3	Fuzzy supplementary deprivation rate: dimension 3 (housing facilities);
12	FSUP-4	Fuzzy supplementary deprivation rate: dimension 4 (housing deterioration);
13	FSUP-5	Fuzzy supplementary deprivation rate: dimension 5 (environmental problems);

3. Methodological note on estimating sampling error for disaggregated direct (survey) estimates

3.1 Introduction

In the production of small area estimates (SAE) using a procedure such as EBLUP, a major technical requirement is the production of sampling error estimates for the disaggregated estimates produced directly from the survey. *Direct estimates* refer to the estimates derived from the survey data for the small areas concerned, taking into account the sampling design. *Synthetic estimates* are those derived by fitting an appropriate small area model. A weighted combination of these two types of estimates is then taken to produce the final *composite estimates* (SAEs). The weights in the combination depend on the relative magnitudes of the *design variance* pertaining to the direct estimates, and the *model variance* of the synthetic estimates.

Model variance or error is a measure of the disparity (variability) between the direct survey estimates (assuming those to be based on 100% coverage of population) of the target variables of interest, and the model estimates based on the predictor variables (regressors); its primary determinant is how well the model fits the data.

Sampling variance (or its square-root, standard error) is a measure of the variability in the direct estimates as a result of those being based only on a sample of the population. Apart from the design, the primary determinant of the magnitude of the sampling error is sample size; hence this component of error increasingly predominates as we move to small areas and domains.

Firstly, sampling error estimates in this context are doubly complex: because the statistics of interest in the study of poverty and deprivation are generally complex, much more so than for instance ordinary proportions, means and ratios; and also because the sample designs on which they are based are complex, involving unequal selection probabilities, stratification, multi-stage selections, aggregation over different samples and times, etc.

Secondly, typically very large numbers of estimates are required. This may be because of the need to include different types of measures, possibly over different subpopulations, but primarily this arises because of the large number of small domains for which the estimates must be produced.

The third difficulty arises from the fact that the estimates of sampling error are themselves subject to variability, which increases with the degree of disaggregation of results as the sample size is reduced. Results of individual computations – even if computationally possible – cannot be always trusted or directly used; this can apply not only to the estimates of variances but also to the estimated statistics themselves.

Fourthly, samples are not always designed in practice so as to permit rigorous estimation of sampling errors from the sample itself. Approximations are often required in making these estimates.

Finally, there is often a problem of insufficient or incomplete documentation and coding in the micro-data of the structure of the sample so as to permit valid estimates of sampling error taking into account the complex sample structure (see Verma, 1993).

Practical approaches and procedures are required to overcome such common difficulties. These involve using approximate procedures and modelling and averaging of individual computations as necessary and appropriate. By ‘appropriate’ we mean procedures which – while not exact or perfect – nevertheless provide estimates which can be considered sufficiently valid and usable for the *purpose for which they are produced*. In this context it is important to note that the requirement of accuracy of the sampling error estimates for the purpose of SAE is somewhat *less stringent* than, for instance, the situation when such estimates are required for constructing confidence interval and the like for individual statistics produced from the survey. This is because in the context of SAE, the role of sampling error is, in the first instance, only to determine the relative weight of the direct survey estimate in the final composite estimate. Of course error in the final estimates does depend on the sampling error, but increasingly less so as the domain sample size goes down. Approximations in the sampling error estimates can be accepted to the extent the final results from the SAE process are not sensitive to those.

3.2 Modelling of sampling errors

A common practical procedure for estimating sampling errors for a set of related statistics is to seek a so-called *generalised variance function* (GVF) which relates the required error of a statistic to some simple and known characteristics of the statistic, such as its value and the sample size. Different functional forms may be required for different types of statistics to produce reasonable approximations of sampling errors; the functional relationships have to be established and validated empirically. There are many well-known examples of the use of such functions in official statistics, for instance US Bureau of the Census (1978). In the specific SAE context, an important example of use of GVFs is National Research Council (1998), reporting the work of the Panel on Estimates of Poverty for Small Geographic Areas in the United States.

Any GVF implies, implicitly or explicitly, constancy of certain parameters (the population variance, coefficient of variation, the design effect, etc.) determining the magnitude of the sampling error of statistics in the group to which it applies. At least the statistics must be similar, and based on the same or a similar design. *This restriction makes such an approach unsuitable in our multi-country, EU-wide context*. The survey statistics which we must use are based on national samples with different designs and structures, even for standardised surveys such as the ECHP. (With the replacement of ECHP by EU-SILC, this diversity is likely to be substantially greater.) For this reason, a different and more flexible approach is required.

The following describes the procedures we have adopted for calculating standard errors for the poverty and related measures estimated at regional level, going down from the country level to NUTS1, and then to NUTS2, and even to NUTS3 where the necessary information for the purpose is available in the survey. On the basis of experience with analysis of patterns of variation of sampling errors, and taking into account our specific multi-country EU-wide context, the approach we recommend and have used has the following features.

1. The standard error of any statistic is broken down into a number of factors which together account for its magnitude. Each factor represents some aspect(s) of the complexity of the sampling design and the estimation procedure (stratification and clustering, weighting, aggregation over surveys, etc.).
2. There is a considerable body of empirical evidence suggesting that many of these factors act more or less independently of each other, so that the factor effects can

be taken as multiplicative (see for instance Verma *et al.* 1980 and Verma and Lê, 1996). In any case, such a simplifying assumption is usually unavoidable in practice.

3. Each factor depends on parameters corresponding to a number of dimensions, in our specific context from the following set: the statistic or variable concerned (v) and the population (u) over which it is defined; country (c) and its particular domain (i); and for a panel survey such as ECHP, the survey wave (w). Reliable estimates of the factors taking into account all these parameters simultaneously are not possible, or even necessary in practice. On the basis of theoretical and empirical considerations, we simplify - and also make the result more robust - by *averaging over dimensions* as appropriate for each parameter. The most obvious and common example is averaging over waves in a panel.
4. Further simplifying assumptions are often required, whether because of a lack of sufficient information (such as on aspects of the sampling design), or because the statistic involved is too complex to permit more precise treatment, or simply to make the task manageable.
5. Specifically, we often have to borrow parameters from simpler statistics for use with more complex, less traceable statistics.

These features will be illustrated in the following, starting with the most important and basic statistic – estimated poverty rate or head count ratio (HCR).

4. Domain sampling error for HCR

For the head count ratio (HCR), we may factorise the standard error estimate (se) into components as follows:

$$se_V = (ser_V).k_V.d_V.s_V.f_V.g_V.r_V$$

Subscript V is the general notation for parameters corresponding to various dimensions, such as the statistic or variable concerned (v), the population (u) over which it is defined, country (c) and its particular domain (i), and survey wave (w). Each of the factors are described below in turn.

4.1 Simple random sample standard error (ser_V)

The first factor in the equation above stands for standard error which would be obtained in a simple random sample of the same size (n_V), without complexities which the other factors represent. Neglecting minor factors such as the 'finite population correction', this factor depends on the sample size in a simple way as follows:

$$ser_V = \left(sd_V / \sqrt{n_V}\right),$$

where sd_V is the standard deviation, a measure of dispersion of the variable in the population, independent of the sample design or size. For a simple proportion p, $sd_V = \sqrt{p.(1-p)}$, which is insensitive to variations in p values over a wide range such as 0.25-0.75, and is well estimated even from samples of small size. The statistic HCR is more complex than a simple proportion, as it is defined in terms of a poverty line which is itself subject to sampling variability. However, empirical results indicate that sd_V defined as above still provides a reasonable approximation for it (Berger and Skinner, 2003; Verma and Betti, 2005). In any case, it is reasonable to average the results over waves and even domains within a country so as to obtain more reliable and stable estimates. In other words, for HCR=p and domain sample size n_i, we can approximate its simple random sample standard error as:

$$ser_{HCR} = \sqrt{p.(1-p)/n_i}\ .$$

4.2 Effect of sample weights (Kish factor, k_V)

Often variations in sampling rates and hence in the sample weights are determined by reporting requirements and other 'external' considerations largely independent of statistical characteristics of the domains of interest. In this sense the weights may be considered arbitrary or haphazard, the effect of which is to inflate the variance of overall estimates. The important thing is that such unequal weights tend to *affect (inflate) the variance of all estimates for different variables in a rather uniform way*, independently of the structure of the sample except for the weighting itself. Herein lies the practical utility of isolating this effect. It is well approximated by the following simple expression (Kish 1965):

$$k_i = \sqrt{n_i.\sum w_j^2 / \left(\sum w_j\right)^2} = \sqrt{1 + cv_i^2\left(w_j\right)}$$

where the sum is over the n_i sample cases, and cv_i is the coefficient of variation of individual weights w_j in domain i. Note that the factor has been taken to depend only on domain i. Some variation can be expected to occur over waves because of changes in the panel sample, but these are normally minor and the results can be averaged over waves.

4.3 Design factor (d_V)

Design factor (or its square, design effect) is a comprehensive summary measure of the effect on sampling error of various complexities in the design. It is the factor by which the actual standard error is different from the error in a simple random sample of the same size. Here this factor represents primarily the effect of stratification and clustering, in so far as the effect of sample weights has already been isolated in terms of (k_i) above. The design effect depends on the structure of the sample as well as the variable being estimated. In the ECHP-UDB data available for the present research, codes for the identification of the sample structure have not been provided generally; consequently, full computation of design effects is not possible at present. However, in Eurostat PAN doc.138 (2000), the information shown in Table 2 below is provided on design effects averaged over household income related variables. Note that with the exception of Portugal and Italy, the design effects are quite small, all within the range 1.0-1.2. In Denmark, Luxembourg and the Netherlands, practically simple random samples were used so that d_c=1.0. For Finland, for Sweden (register data), as well as for the survey data from Poland and Romania, we have assumed similar values in the absence of better information at hand. In view of the generally small range within which the design effects vary in the present case, it is sufficient to assume that, within each country, a common design effect value can be used for the set of income poverty and deprivation variables of interest, and that the same value applies across different regions in the country.

Table 2 Design and Kish factors for income-related variables

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	mean
	DE	DK	NL	BE	LU	FR	UK	IE	IT	GR	ES	PT	AT	FI	SE	
k_i	1,07	1,06	1,10	1,08	1,08	1,03	1,04	1,11	1,13	1,09	1,13	1,36	1,19	(1,10)	(1,10)	1,11
d_c	1,12	1,00	1,00	1,11	1,00	1,12	1,13	1,19	1,86	1,23	1,14	1,67	1,13	(1,00)	(1,00)	1,16
k_i*d_c	1,20	1,06	1,10	1,20	1,08	1,15	1,17	1,32	2,10	1,34	1,29	2,27	1,34	(1,10)	(1,10)	1,29

(..) assumed values; PL & RO: the design effect was assumed as 1,20. The Kish factor in Romania is 1,31.

4.4 Subpopulation factor (s_V)

For a subpopulation distributed reasonably uniformly across the population, the sampling error for an estimate over the subpopulation (s) can be related in a simple form to that for an estimate over the total population (c). Examples are HCR for children or old persons, compared to the HCR estimated for the total population. The approximate relationship is:

$$s_V = \left(\frac{se_s}{se_c}\right) = \left(\frac{n_c}{n_s}\right)^{1/2} . \left(1 + \left(\frac{n_s}{n_c}\right) . \left(d_c^2 - 1\right)\right)^{1/2} , \ d_c \geq 1 ,$$

where n_s is the number in the sample from the subpopulation (children, elderly persons, recent school-leavers, etc.), and n_c is number in the sample from the total population. The first factor is the increase in sampling error because of the reduced sample size when we consider only the subpopulation of interest. This is partly balanced by the second factor which gives the reduced design effect. (The design effect is reduced because of reduced cluster size when units belonging to the subpopulation only are considered.)

4.5 Reduction due to aggregation over waves (f_V)

Factor f_V reduces the standard error because of consolidation of measures over waves. Of course, we cannot merely add up the sample seizes over waves since ECHP is a panel survey and there is a high positive correlation in the poverty measures among the years, which reduces the gain from cumulation. The correlation can be estimated as follows. Consider two adjacent waves, with proportion poor as p and p', respectively, with the following individual-level overlaps between the two waves:

	Wave w+1		
Wave w	Poor (p'_i=1)	Non-poor (p'_i=0)	total
Poor (p_i=1)	a	b	p=a+b
Non-poor (p_i=0)	c	d	1-p=c+d
total	p'=a+c	1-p'=b+d	1=a+b+c+d

Indicating by p_j and p'_j the {1,0} indicators of poverty of individual j over the two waves, we have, with the sum over all (n) individuals:

$$\mathrm{var}(p_j) = \Sigma(p_j - p)^2 / n = p.(1-p) = v_1 \text{, say;}$$

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

For data averaged over two adjacent years (and ignoring the difference between p and p'), variance is given by:

$$v_2 = \frac{1}{4}.(v_1 + v_1 + 2.c_1) = \frac{v_1}{2}.\left(1 + \frac{c_1}{v_1}\right).$$

The correlation (c_1/v_1) between two periods is expected to decline as the two become more widely separated. Let (c_i/v_1) be the correlation between two points i waves apart. A simple and reasonable model of the attenuation with increasing i is:

$$(c_i/v_1) = (c_1/v_1)^i.$$

Now in a set of I periods (waves) there are (I-i) pairs exactly i periods apart, i=1 to (I-1). It follows from the above that variance v_I of an average over I periods relates to variance v_1 of the estimate from a single wave as:

$$f_c^2 = \left(\frac{v_I}{v_1}\right) = \frac{1}{I}.\left(1 + 2.\Sigma_{i=1}^{I-1}\left(\frac{I-i}{I}\right).\left(\frac{c_1}{v_1}\right)^i\right), \text{with } \left(\frac{c_1}{v_1}\right) \approx a - p^2,$$

where a is the overall rate of persistent poverty between pairs of adjacent waves (averaged over I-1 pairs), and p is the (cross-sectional) poverty rate averaged over I waves. The ratio of the corresponding standard errors is f_c. Due to averaging over I waves, the effective sample size is increased by $(1/f_c^2)$. We take factor f_c to be country-specific, more or less independent of the particular variable in the set.

Table 3 Reduction in standard error resulting from cumulation over waves

	1 DE	2 DK	3 NL	4 BE	5 LU	6 FR	7 UK	8 IE	9 IT	10 GR	11 ES	12 PT	13 AT	14 FI	15 SE
p (average HCR)	0,12	0,11	0,11	0,14	0,12	0,15	0,19	0,19	0,19	0,21	0,19	0,21	0,13	0,10	0,10
a (persistent poverty rate)	0,07	0,06	0,06	0,09	0,09	0,10	0,12	0,13	0,13	0,14	0,12	0,15	0,08	0,06	0,01
I (no of waves)	8	8	8	8	7	8	8	8	8	8	8	8	7	6	5
Gain over single wave f_c	0,59	0,54	0,57	0,59	0,69	0,60	0,60	0,63	0,60	0,60	0,57	0,65	0,61	0,66	0,45
Effective number of waves	2,90	3,39	3,13	2,86	1,86	2,76	2,77	2,49	2,76	2,80	3,02	2,36	2,67	2,29	5,00

Poland and Romania: no cumulation over waves is involved

Table 3 shows values of the parameters actually obtained for ECHP data over 8 waves. The last two rows show, respectively, the gain in precision (reduced standard error) over a single wave as a result of cumulation, and the factor by which the effective sample size achieved exceeds the average sample size for a single wave.

4.6 Reduction from averaging different poverty thresholds (g_V)

One of the important recommendation of this research is that in constructing regional poverty rates and similar statistics from limited sample sizes, some gain in efficiency can be achieved by computing those using different poverty thresholds (such as 50, 60 and 70% of the median income), and then taking an appropriately weighted average of those. It is desirable to take these weights as externally determined constants. Consider three poverty line thresholds, with poverty rates p_i, $p_1 < p_2 < p_3$, such the with fixed weights W_i, the final rate is computed as $p = \Sigma_i W_i.p_i$. Its variance is given by:

$$\text{var}(p) = \Sigma_i W_i^2.\text{var}(p_i) + 2.\Sigma_{j<i} W_i W_j.\text{cov}(p_i, p_j).$$

By considering the poverty indicator variables $p_{i,k} = \{0,1\}$ for individual j in the population, it can be easily seen that the above equation becomes:

$$\mathrm{var}(p) = \Sigma_i W_i^2 .p_i .(1-p_i) + 2.\Sigma_{j<i} W_i W_j .p_j .(1-p_i).$$

It is this variance that we compare with the variance of a rate (p_2) computed using a single poverty line such as 60% of the median, as is normally done: $\mathrm{ver}(p_2) = p_2 .(1-p_2)$. The ratio $g_V = (\mathrm{var}(p)/\mathrm{var}(p_2))^{1/2}$ gives the required factor by which the standard error is reduced. Table 4 gives the actual factors obtained for ECHP data, using appropriately weighted consolidation over three poverty line thresholds, namely 50%, 60% and 70% of the median as explained above. In fact, computations have been performed, also using different poverty line levels in the sense described earlier, that is by defining the median for population aggregations to different levels such as NUTS2, NUTS1, Country or EU where possible. The factors are remarkably robust to such changes in the level as seen in Table 4. (Only country and NUTS2 poverty line results are shown in the table, as they are the most relevant.)

Table 4 Reduction in standard error from consolidation over different poverty line thresholds

Poverty line level	1 DE	2 DK	3 NL	4 BE	5 LU	6 FR	7 UK	8 IE	9 IT	10 GR	11 ES	12 PT	13 AT	14 FI	15 SE	25 PL
NUTS2	0,93	0,87	0,91	0,90	0,87	0,91	0,94	0,88	0,94	0,96	0,94	0,95	0,89	0,88	0,92	0,95
Country	0,93	0,87	0,91	0,90	0,87	0,91	0,93	0,88	0,95	0,97	0,95	0,96	0,89	0,87	0,92	0,96

This factor equals 1.0 for RO as no consolidation over poverty line thresholds was carried out in this country.

4.7 Standard error of ratio of estimates in a hierarchy (r_V)

As noted earlier, it is more efficient to model the small area estimates in a hierarchical manner. In place of estimating the absolute value of any statistic (say e_2), we estimate instead the *ratio* ($r=e_2/e_1$) of the statistic at one level such as NUTS2, to its estimate at the preceding (higher) level such as NUTS1. The objective is to obtain var(r), given var(e_2) obtained as described in the steps above. We have:

$$\mathrm{var}(r) = \mathrm{var}\left(\frac{e_2}{e_1}\right) = \frac{1}{e_1^2}.\left(\mathrm{var}(e_2) + r^2.\mathrm{var}(e_2) - 2.r.\mathrm{cov}(e_1, e_2)\right).$$

The covariance is easily evaluated by noting that sample "2" is just a subsample of "1", with the same measurements so that correlation between them is 1.0. It can be shown that with n_2 as the size of the subsample of sample n_1:

$$\mathrm{cov}(e_1, e_2) = [\mathrm{var}(e_1).\mathrm{var}(e_2).(n_2/n_1)]^{1/2}.$$

With the reasonable assumption:

$$\frac{\mathrm{var}(e_2)}{e_2^2} = \frac{\mathrm{var}(e_1)}{e_1^2}.\left(\frac{n_1}{n_2}\right), \text{ that is, } \mathrm{var}(e_2) = r^2.\mathrm{var}(e_1).\left(\frac{n_1}{n_2}\right),$$

we get the simple expression for the required factor (r_V):

$$r_V = \left(\frac{\mathrm{var}(r)}{\mathrm{var}(e_2)}\right)^{1/2} = \left(\frac{1}{e_2}\right).\left(\frac{n_1 - n_2}{n_1}\right)^{1/2},$$

the second factor representing the gain resulting from the fact that sample "2" is simply a subsample of "1".

5. Standard errors for other statistics

In this application we have considered 13 main poverty measures, listed in Table 1 above, for which estimates of standard errors are required at various levels (Country, NUTS1, NUTS2, …). Using the factors for the head count ratios [statistics 1-2] derived above, the corresponding factors for the other statistics have been obtained using the following simplified procedures.

5.1 Measures related to income levels

The main differences from the HCR sampling error concern the computation of the standard deviation sd_V, and factor g_V which equals 1.0 since no consolidation over poverty lines is involved. We have assumed all other factors to be the same as those for HCR.

For a variable y such as log-income, standard deviation is computed as:

$$sd_{c,v,w} = \left(\Sigma w_j.(y_j - \bar{y})^2 / \Sigma w_j\right)^{1/2},$$

with $\bar{y} = \left(\Sigma w_j y_j . / \Sigma w_j\right)$, and w_j as the sample weights. The subscripts have been used in the above to indicate that the expression is specific to country (or region), variable and wave. In order to average values over waves, it is preferable to work with the coefficient of variation $cv_{c,v,w} = sd_{c,v,w} / \bar{y}$, which is scale-free and therefore not affected by inflation or the unit of measurement. This permits its straightforward averaging over waves:

$$cv_{c,v} = \frac{1}{T}.\sum_{w:1}^{T} cv_{c,v,w}.$$

After that, averaged value of standard deviation can be calculated as:

$$sd_{c,v} = \frac{cv_{c,v}}{T}.\sum_{w:1}^{T} \bar{y}_{c,v,w}$$

5.2 Fuzzy measures

We assume that the same structure and parameters as above for sampling error of HCR apply for related fuzzy measures of the degree of poverty and deprivation. Of course, standard deviation is computed with reference to proportion p_v for the variable concerned, which may differ significantly from p for HCR. Fuzzy measures have been computed here with reference to a single poverty threshold (60% of the median income), rather than consolidated over three thresholds as was done in the case of the HCR. Consequently, factor g_c=1. On the other hand, however, we expect fuzzy measures to have smaller variance than conventional HCR based on a dichotomous (yes-no) variable. We have not investigated the magnitude of this effect in the present (ECHP) data, but have simply kept the HCR g_c<1 unchanged to make an allowance for it. (See Betti *et al.* (2006) for a complete discussion of such measures).

6. Empirical results

The methodology proposed in Section 3 and implemented to poverty measures in Section 4 and Section 5 has been applied in EU-25 and Candidate countries.
The availability of data for the estimation of such measures is presented in Table 5. Sufficient information is not available in the ECHP surveys in Germany, Luxembourg and Sweden to construct deprivation measures in special dimensions (variables Fuzzy Supplementary 1-5). Only income related measures could be computed from the survey in Romania. It should also be noted that some of the non-monetary measures for Poland lack comparability with corresponding ECHP measures because of differences in the survey questions used.

Table 5 The availability of the target variables in EU25 and Candidate countries

		target variable	1	2	3	4	5	6	7	8	9	10	11	12	13
			HCR_c	HCR_n2	logEqInc	logIncPC	FM_c	FS_c	LAT_c	MAN_c	FSUP-1	FSUP-2	FSUP-3	FSUP-4	FSUP-5
1	DE	Germany	X	X	X	X	X	X	X	X					
2	DK	Denmark	X	X	X	X	X	X	X	X	X	X	X	X	X
3	NL	Netherlands	X	X	X	X	X	X	X	X	X	X	X	X	X
4	BE	Belgium	X	X	X	X	X	X	X	X	X	X	X	X	X
5	LU	Luxembourg	X	X	X	X	X	X	X	X					
6	FR	France	X	X	X	X	X	X	X	X	X	X	X	X	X
7	UK	United Kingdom	X	X	X	X	X	X	X	X	X	X	X	X	X
8	IE	Ireland	X	X	X	X	X	X	X	X	X	X	X	X	X
9	IT	Italy	X	X	X	X	X	X	X	X	X	X	X	X	X
10	GR	Greece	X	X	X	X	X	X	X	X	X	X	X	X	X
11	ES	Spain	X	X	X	X	X	X	X	X	X	X	X	X	X
12	PT	Portugal	X	X	X	X	X	X	X	X	X	X	X	X	X
13	AT	Austria	X	X	X	X	X	X	X	X	X	X	X	X	X
14	FI	Finland	X	X	X	X	X	X	X	X	X	X	X	X	X
15	SE	Sweden	X	X	X	X	X	X	X	X					
23	PL	Poland	X	X	X	X	X	X	X	X	X	X	X	X	X
27	RO	Romania #	X	X	X	X	X								
16	CY	Cyprus	*		*										
17	CZ	Czech Republic	*		*										
18	EE	Estonia	*		*										
19	HU	Hungary	*		*										
20	LT	Latvia	*		*										
21	LV	Lithuania	*		*										
22	MT	Malta	*		*										
24	SI	Slovenia	*		*										
25	SK	Slovakia	*		*										
26	BG	Bulgaria	*		*										
28	TR	Turkey	*		*										

\# Romania: Consumption instead of income variables
* Only published indicators available on HCR (poverty line 60% of national median), and national median income, mostly based on HBS.

6.1 Performance measures

Table 6 shows some 'performance measures' of SAE Model 1. For each model (i.e., target) variable, three measures are shown:

i) the model parameter gamma (γ). It is the ratio between the model variance and the total variance, and is the share of the weight given to the direct survey estimate in the final composite estimate.

ii) ratio (a)*/(a), i.e., the ratio between the EBLUP estimated value of (a)* and the corresponding direct estimation (a). This is to check the extent to which the modelling changes the input direct estimates.

iii) ratio (b)*/(b), i.e., the ratio between mean-squared error (MSE) of the EBLUP estimate of the (NUTS1: Country) ratio, and MSE of direct survey estimate of this ratio. This is to check the extent to which the modelling has improved precision of the estimates.

For each of the above, the following summary statistics are given: the mean value over all NUTS1 areas in the model; the coefficient of variation of those values; and the minimum and maximum values.

Overall the results are as expected: the SAE Model1 for NUTS1 level does not provide much gain, as can be seen from the mean ratio of mean-squared errors. This is because the sample sizes for most NUTS1 areas are actually quite large; NUTS1 can hardly be called 'small areas'. The large sample sizes are achieved by cumulation of data over survey waves.

The largest gains in efficiency are for Manifest Deprivation Rate and HCR with NUTS2 poverty line. It is particular noteworthy for HCR_N2, where the MSE is reduced to two-thirds, which implies more than doubling the effective sample size. Since this variable is based on income distributions within each NUTS2 region (even though the modelling being discussed is from country to NUTS1 level of aggregation), it is possible that the sampling errors of the direct estimates are larger. The main reason for the better performance of the model, however, must be a stronger relationship of HCR_N2 with the predictor variables used, compared to the same relationship for HCR_C. This is an important observation because of the substantive importance, as noted earlier, of HCR_N2 as a regional indicator of poverty.

Table 6 Performance measures for SAE Model 1

	Gamma				Estimate EBLUP/direct estimate				Mean-squared error (MSE) MSE(EBLUP)/MSE(direct estimate			
	mean	CV	min	max	mean	CV	min	max	mean	CV	min	max
1 HCR_C	0,86	0,15	0,41	0,99	0,99	0,10	0,70	1,49	0,95	0,19	0,35	1,90
2 HCR_N2	0,35	0,47	0,03	0,73	1,00	0,05	0,84	1,14	0,67	0,23	0,23	0,93
3 logEqInc	0,95	0,05	0,71	0,99	1,00	0,00	1,00	1,00	0,98	0,02	0,89	1,00
4 logIncPC	0,95	0,05	0,71	0,99	1,00	0,00	1,00	1,00	0,98	0,02	0,89	1,00
5 FM_C	0,83	0,16	0,35	0,98	0,99	0,05	0,72	1,05	0,92	0,07	0,68	0,99
6 FS_C	0,83	0,16	0,39	0,98	1,00	0,05	0,84	1,28	0,93	0,07	0,70	0,99
7 Latent	0,86	0,14	0,38	0,98	1,00	0,03	0,81	1,11	0,94	0,06	0,70	0,99
8 Manifest	0,66	0,36	0,15	0,96	0,98	0,12	0,60	1,39	0,83	0,18	0,43	0,99
9 Fsup_1	0,93	0,05	0,74	0,99	1,00	0,02	0,96	1,03	0,97	0,02	0,89	1,00
10 Fsup_2	0,86	0,10	0,65	0,98	1,00	0,03	0,89	1,11	0,94	0,04	0,84	0,99
11 Fsup_3	0,70	0,32	0,08	0,98	0,99	0,17	0,36	1,32	0,86	0,16	0,29	1,00
12 Fsup_4	0,88	0,09	0,65	0,98	1,00	0,02	0,94	1,06	0,96	0,04	0,84	0,99
13 Fsup_5	0,88	0,07	0,73	0,98	1,00	0,02	0,96	1,05	0,96	0,03	0,89	0,99

Table 7 shows some 'performance measures' of SAE Model 2. For each model (target variable), three measures are shown as in Table 6.

The performance of the model in terms of gain in efficiency is obviously better for Model 2 (NUTS2 level) compared to Model 1 (NUTS1 level). This is because in the former the sample sizes available for direct estimates are smaller. The highest gains, of 20-25%, are for Latent, Manifest and FSUP-1 deprivation measures. Again, as with Model1, the gain for HCR_N2 is almost twice as large as that for HCR_C. This is important in the context of constructing regional indicators. The gain for HCR_C, FSUP-2, FSUP-4 and FSUP-5 is around 10%, while no prediction is possible for FSUP-3 for lack of adequate data. For logarithm of equivalised income and the logarithm of the per capita income, the relative gains are the smallest among the variables.

Table 7 Performance measures for the SAE Model 2

	Gamma				Estimate EBLUP/direct estimate				Standard error (SE) SE(EBLUP)/SE(direct estimate)			
	mean	CV	min	max	mean	CV	min	max	mean	CV	min	max
1 HCR_C	0,80	0,22	0,45	0,98	1,01	0,08	0,86	1,34	0,90	0,11	0,71	1,00
2 HCR_N2	0,66	0,38	0,19	0,95	1,01	0,07	0,83	1,30	0,82	0,22	0,47	1,00
3 logEqInc	0,81	0,23	0,44	0,98	1,00	0,00	1,00	1,01	0,94	0,18	0,68	1,35
4 logIncPC	0,85	0,14	0,65	0,99	1,00	0,00	0,99	1,01	0,92	0,12	0,74	1,21
5 FM_C	0,75	0,27	0,40	0,98	1,02	0,14	0,80	1,63	0,88	0,12	0,66	1,02
6 FS_C	0,68	0,32	0,32	0,97	1,02	0,09	0,85	1,45	0,85	0,14	0,63	0,99
7 Latent	0,61	0,36	0,23	0,96	1,01	0,08	0,84	1,41	0,81	0,16	0,50	0,98
8 Manifest	0,55	0,49	0,12	0,97	1,06	0,25	0,71	2,25	0,76	0,24	0,36	1,00
9 Fsup_1	0,60	0,41	0,22	0,97	1,01	0,08	0,86	1,28	0,80	0,18	0,54	1,00
10 Fsup_2	0,73	0,22	0,47	0,97	1,01	0,07	0,87	1,26	0,88	0,09	0,70	0,99
11 Fsup_3												
12 Fsup_4	0,77	0,15	0,51	0,97	1,01	0,05	0,88	1,24	0,90	0,06	0,76	0,99
13 Fsup_5	0,76	0,22	0,49	0,98	1,00	0,05	0,87	1,11	0,89	0,10	0,72	1,01

Let us now pass to EBLUP models for going from NUTS2 to NUTS3 level. SAE Models 3 are estimated for Italy (only this database makes possible the access to area-coded survey at NUTS3 level). Given the high level of disaggregation we decided to consider only three poverty indicators (consequently three models): the HCR_C, the HCR_N2, logEqInc. The list of the independent variables available is also more limited; it is confined to the relevant covariates, tables for which are provided in NewCronos at NUTS3 level.

The available set covariates is very limited indeed. It would be important to find additional and better covariates in real-life replications of SAE Model 3.

Table 8 shows some 'performance measures' of SAE Model 3. For each model (target variable), three measures are shown as in Table 7.

In this case we really have small areas with very small sample sizes. The average gain in precision is at least 20%, and it is quite consistent across the target variables. It is interesting to note the minimum value of the ratio between the EBLUP standard error and the direct standard error: the minimum values in all the three models are less the 0.10. This means that in some areas the EBLUP estimator provides a gain in efficiency, compared to the direct survey estimates, that is higher than 90%.

Table 8 Performance measurement for the SAE Model 3

	Gamma				Estimate EBLUP/direct estimate				Standard error (SE) SE(EBLUP)/SE(direct estimate)			
	mean	CV	min	max	mean	CV	min	max	mean	CV	min	max
HCR_C	0.70	0.41	0.01	1.00	1.05	0.27	0.44	2.53	0.81	0.30	0.10	1.00
H_N2	0.76	0.36	0.00	1.00	1.03	0.20	0.46	2.21	0.85	0.27	0.08	1.00
logEqInc	0.62	0.44	0.00	1.00	1.00	0.01	0.96	1.05	0.77	0.32	0.05	0.98

References

Berger Y., Skinner C. (2003), Variance estimation for a low income proportion, *Applied Statistics*, **52(4)**, pp. 457-468.

Berthoud R. (2004), *Patterns of poverty across Europe*. The Policy Press.

Betti G, Cheli B., Lemmi A., Verma V. (2006), Multidimensional and Longitudinal Poverty: an Integrated Fuzzy Approach, in Lemmi A., Betti G. (eds.), *Fuzzy Set Approach to Multidimensional Poverty Measurement*. New York : Springer Science+Business Media, LLC, 2006, pp. 111-137.

Betti G., Lemmi A., Natilli M., Verma V. (2006), Indicators of social exclusion and poverty in Europe's regions, Working Paper n. 59, Dipartimento di Metodi Quantitativi, Università di Siena.

Betti G., Verma V. (1999), Measuring the degree of poverty in a dynamic and comparative context: a multi-dimensional approach using fuzzy set theory, Proceedings of the ICCS-VI, pp. 289-301, Lahore, Pakistan, August 27-31, 1999.

Eurostat (2000), Sampling Errors: Proportions, means and other ratios estimated from the household questionnaire, *DOC. PAN 138/2000* (unpublished).

Kish L. (1965), *Survey Sampling.* New York. Wiley.

National Research Council (1998), *Small-Area Estimates of School-Age Children in Poverty*. Washington DC: National Academy Press.

Neri L., Verma V. (2006), Application of EBLUP for the construction of regional indicators of poverty, Working Paper n. 58, Dipartimento di Metodi Quantitativi, Università di Siena.

US Bureau of Census (1978), An error profile: Employment as measured by the Current Population Survey. *Statistical Policy Working Paper*, 3. Washington DC: OFSPS.

Verma V. (1993), *Sampling Errors in Household Surveys: A Technical Study*. New York: United Nations Department for Economic and Social Information and Analysis. (Statistical Division INT-92-P80-15E).

Verma V., Betti G. (2005), Sampling Errors for Measures of Inequality and Poverty. Invited paper in Classification and Data Analysis 2005 - Book of Short Papers, pp. 175-179, CLADAG, Parma, 6-8 June 2005.

Verma V., Lê T. (1996), An Analysis of Sampling Errors for the Demographic and Health Surveys. *International Statistical Review*, **64(3)**, pp. 265-294.

Verma V., Scott C., O'Muircheartaigh C. (1980), Sample designs and sampling errors for the World Fertility Survey. *Journal of the Royal Statistical Society, Series A*, **143(4)**, pp. 431-473.