ISSN 1977-0375

eurostat
Methodologies and
Working papers

# Robustness of some EU-SILC based indicators at regional level

2010 edition

eurostat
EUROPEAN COMMISSION

eurostat
Methodologies and
Working papers

# Robustness of some EU-SILC based indicators at regional level

2010 edition

eurostat
EUROPEAN COMMISSION

More information on the European Union is available on the Internet (http://europa.eu).

Cataloguing data can be found at the end of this publication

Luxembourg: Publications Office of the European Union, 2010

ISBN 978-92-79-16756-0
ISSN 1977-0375
Doi:10.2785/5605
Cat. No. KS-RA-10-020-EN-N

**Theme: Population and social conditions**
**Collection: Methodologies and working papers**

Eurostat is the Statistical Office of the European Union (EU). Its mission is to provide the EU with high-quality statistical information. To that end, it gathers and analyses data from the National Statistical Institutes (NSIs) across Europe and provides comparable and harmonised data for the EU to use in the definition, implementation and analysis of EU policies. Its statistical products and services are also of great value to Europe's business community, professional organisations, academics, librarians, NGOs, the media and citizens. In the social field, the EU Statistics on Income and Living Conditions (EU-SILC) instrument is the main source for statistics on income, poverty, social exclusion and living conditions.

Over the last years, important progress has been made in EU-SILC. This is the result of the coordinated work of Eurostat and the NSIs, *inter alia* in the context of the EU 'Living Conditions' Working Group and various thematic Task-Forces. Despite these significant achievements, EU-SILC data are still insufficiently analysed and used.

It is in this context that Eurostat launched in 2008 a call for applications with the following aims:

(1) develop methodology for advanced analysis of EU-SILC data;
(2) discuss analytical and methodological papers at an international conference;
(3) produce a number of publications presenting methodological and analytical results.

The 'Network for the Analysis of EU-SILC' (Net-SILC), an ambitious 18-partner Network bringing together expertise from both data producers and data users, was set up as in response to this call. The initial Net-SILC findings were presented at the international conference on 'Comparative EU Statistics on Income and Living Conditions' (Warsaw, 25-26 March 2010), which was organised jointly by Eurostat and the Net-SILC network and hosted by the Central Statistical Office of Poland. A major deliverable from Net-SILC is a book to be published by the EU Publications Office at the end of 2010 and edited by Anthony B. Atkinson (Nuffield College and London School of Economics, United Kingdom) and Eric Marlier (CEPS/INSTEAD Research Institute, Luxembourg).

The present methodological paper is also an outcome from Net-SILC. It has been prepared by Vijay Verma, Gianni Betti and Francesca Gagliardi (University of Siena, Italy). Gara Rojas González was responsible at Eurostat for coordinating the publication of the methodological papers produced by Net-SILC members.

It should be stressed that this methodological paper does not in any way represent the views of Eurostat, the European Commission or the European Union. The authors have contributed in a strictly personal capacity and not as representatives of any Government or official body. Thus they have been free to express their own views and to take full responsibility both for the judgments made about past and current policy and for the recommendations for future policy.

This document is part of Eurostat's *Methodologies and working papers* collection which are technical publications for statistical experts working in a particular field. All publications are downloadable free of charge in PDF format from the Eurostat website:
(http://epp.eurostat.ec.europa.eu/portal/page/portal/income_social_inclusion_living_conditions/publications/Methodologies_and_working_papers ). Furthermore, Eurostat databases are freely available at this address, as are tables with the most frequently used and requested short- and long-term indicators.

# Table of contents

# Robustness of some EU-SILC based indicators at regional level

**Vijay Verma, Gianni Betti and Francesca Gagliardi [1],**
**University of Siena, Italy**

**Abstract:** This working paper addresses some statistical aspects related to the construction from EU-SILC data of indicators of poverty and social exclusion for sub-national regions. Conceptual and methodological issues in going from the national level – for which EU-SILC surveys are primarily designed– to the regional level are discussed. Then five complementary approaches aimed at making the best use of available survey data for regional estimation are identified, and statistical considerations involved in their application are discussed. In addition to direct estimates from the survey, these include the construction of alternative measures utilising the available data more intensively, data cumulation over time, use of survey data in conjunction with external sources using small area estimation methods, and constructing supplementary regional indicators directly from administrative and other large-scale sources. The paper concludes with a plea for the provision of more complete information in EU-SILC microdata on sample structure and implementation and for the identification of regions in sufficient detail – information which would greatly enhance the usefulness of the data for regional analysis.

**Key words:** regional indicators, NUTS regions, data cumulation, small area estimation, regional poverty lines, variance estimation, EU-SILC.

[1] This Working Paper was prepared as part of the "Net-SILC" project. Net-SILC, the Network for the analysis of EU-SILC is an international research project funded by Eurostat and coordinated by the CEPS/INSTEAD Research Institute (Luxembourg). The authors thank Giulia Ciampalini for help in the production of numerical results and Olympia Bover for valuable comments and suggestions. Of course, these persons are not responsible for the present contents. The European Commission also bears no responsibility for the analyses and conclusions, which are solely those of the authors. Addresses for correspondence: vijay.verma46@gmail.com, betti2@unisi.it, gagliardi10@unisi.it.

# 1. Introduction

## 1.1 Objectives

EU Statistics on Income and Living Conditions (EU-SILC) survey was designed for the explicitly stated purpose of constructing indicators of poverty and social exclusion primarily at the *national level* in each country. For example, the minimum required cross-sectional and longitudinal 'effective' sample sizes were stipulated in EU-SILC Commission Regulations on the basis of this requirement. Nevertheless, this valuable instrument must contribute – and to considerable extent also has the potential to contribute – towards the provision of such indicators for *sub-national regions*. This Working Paper addresses some statistical aspects relating to this requirement, in particular concerning the sampling precision of the regional indicators which can be produced.

Indicators of poverty and social exclusion are an essential tool for monitoring progress in the reduction of these problems. In the EU-wide context, these indicators are most useful when they are comparable across countries, so that the situation in individual EU Member States can be evaluated in relation to the situation in other countries. These indicators also need to be comparable over time for monitoring trends. For this purpose, the European Commission has adopted a common set of indicators, referred to as the Laeken Indicators. The set of common indicators is supplemented by country-specific indicators, chosen flexibly according to the requirements and data availability in individual countries. Hitherto, most of the indicators have been defined and constructed only at the national level, except for breakdown for special subpopulations such as children, other groups by age and gender, or different household types. This is because the construction of these indicators is based on sample surveys that are rarely large enough for sufficiently useful (reliable) estimation at the regional level.

Robustness of regional indicators is examined in this Working Paper in terms of sampling errors only. Concerning non-sampling errors, the data collection procedures and problems are normally similar across regions in any given country. The main point of interest is to identify any significant differences in the outcome among regions, such as in the rates of non-response of various types. Regional patterns in non-sampling errors are not pursued here. For discussion of non-sampling and sampling errors in EU-SILC at the national and EU levels, see Verma *et al* (2010).

The general objectives of the present study are the following. The study concerns the robustness of a few EU-SILC based Laeken indicators at regional level and aims to assess their statistical reliability at one point in time, primarily in terms of sampling reliability, that is, in terms of the magnitude of the sampling error of estimates based on EU-SILC. In ‘direct’ estimation of such indicators based on a single wave of cross-sectional survey data, the primary concern is the increased sampling error when the results are broken down by region. Starting with direct estimates from the survey, the study explores what can be done for calculating more reliable regional estimates by, for example, cumulating over waves or using small areas estimates techniques, also addressing in such approaches the issue of robustness of the indicators calculated.

## 1.2 The approach: making the best use of available sample survey data

Section 2 considers in some detail various issues involved in adaptation to the regional level indicators originally defined for use primarily at the national level. 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, in a more consolidated manner, and therefore are like to be statistically more robust.

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

(4) Using in combination the survey data and 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 for sub-national regions using appropriate small area estimation (SAE) techniques.

(5) To these we must add the possibility of going altogether beyond the survey, exploiting external sources - in particular ‘meso’ data such as the highly disaggregated tabulations available in NewCronos - directly for the purpose of constructing indicators for small areas.

Sections 3-8 discuss each of the above aspects in turn. Numerical illustrations using EU-SILC data are provided wherever possible.

In constructing direct estimates for regions from survey data, there are essentially no new technical issues involved in producing the estimates themselves. Technical procedures need to be clarified in relation to the estimation of sampling errors, however. This is done in Section 3. Empirical illustrations using EU-SILC data are provided in Section 4.

In Section 5, a specific form of consolidation is described, namely computing poverty rates using several poverty lines defined as different percentages of the median equivalised income, and taking an appropriate average of those rates. In this way, the measures can be made less sensitive to irregularities in empirical data based on small regional samples.

In cumulating over waves of EU-SILC involving overlaps, technical complications arise because of positive correlation between samples which share the same households and persons in a rotational panel design. Variance estimation procedures at the regional level in the presence of cumulation over survey waves are described in Section 6.

Section 7 describes a particular small area estimation approach (called EBLUP) for constructing regional indicators which has been tested successfully on similar data from European Community Household Panel (ECHP) in our previous work (Verma *et al*, 2005). Such an approach can be expected to be equally suitable for applications using EU-SILC. The method involves the use of survey data in conjunction with a source such as NewCronos available in a comparable form across EU countries.

The discussion in the Working Paper is concerned with methods for exploiting EU-SILC data more effectively for the construction of sub-national (regional) indicators. In the concluding Section 8, we note the possibility of going altogether beyond the survey by exploiting external administrative and other large-scale data sources directly for obtaining additional regional indicators complementing the outputs from EU-SILC. The concluding remarks also reiterate an important concern of this study. This is to explore and expose the barriers which researchers, using the restricted information provided in EU-SILC documentation and its Users Data Base (UDB) in the public domain, face in assessing quality of the data. This issue is important for proper use of the data and for the development and improvement of EU-SILC itself, and needs to be brought out prominently.

# 2. Adapting indicators to the regional level

National level indicators are not necessarily appropriate or sufficient for regional analysis. We need to consider indicators suitable and useful for the regional (sub-national) level, and describe and illustrate the statistical methodology for their construction. The established set of country-level indicators does provide the basis for developing indicators suitable for the regional level, but there are serious limitations. Of course some of the country-level indicators can be usefully classified down to the regional level in their existing form. However, some other may need modification (simplification) before such classification, and more importantly, there are also country-level indicators which are not suitable (meaningful, useful, feasible) for regional breakdown. It is also necessary to consider additional, specifically regional indicators which are not covered in the country-level list.

## 2.1 Choice of units to serve as ‘regions’

The first issue in developing regional indicators concerns the choice of the type of units to serve as ‘regions’. For a number of substantive and practical reasons, geographical-administrative regions, specifically NUTS regions (and LAUs) at various level of classification, appear as the most appropriate choice for EU countries. The reasons for this choice include the following. NUTS regions are the most commonly used units for the formulation and implementation of social policy: the units are well-defined and identifiable, and are already widely accepted and used by different users and producers of statistical information. Despite the fact that NUTS units are not defined in exactly the same way in different countries and can differ greatly in size and homogeneity, this territorial system of classification provides a common framework which enhances comparability of the resulting statistical information.[2] Inter-country, EU-wide research also benefits from the use of units based on the same system of classification.

The NUTS classification covers each country exhaustively, providing a hierarchical set of units for which data can be linked across different levels. A lot of information already exists for these type of units from many different sources. Above all, data availability (not only from EU-SILC but also from many other sources of different types) for the purpose of constructing the required indicators is the major reason for the choice of NUTS regions for the purpose.

[2] In EU-27, the average population size is around 5.5 million per NUTS1 region, around 1.8 million per NUTS2 region, and a little under 0.4 million per NUTS3 region. The units vary considerably in size across the countries, and often also within countries. However, generally the range of variation declines as we go down the hierarchy of NUTS regions.

This by no means precludes NUTS as regions being supplemented by other dimensions. For instance, it is possible to consider ‘functional regions’, such as regions defined in terms of the labour market, production, trade or other economic indicators, or in terms of density and other characteristics of the population distribution (e.g., urban-rural distinction). Indeed, the analysis can accommodate different types of units simultaneously. For instance, NUTS regions at a sufficiently low level can be classified according to whether their character is primarily urban or primarily rural. In fact, indicators can be constructed for geographical-administrative units precisely for the purpose of such classification. Furthermore, NUTS-based indicators can be enriched by subpopulation analysis to the extent available data permit their further disaggregation.

## 2.2 Usefulness of measures of averages

When measures at the regional level are constructed by aggregating information on individual elementary units, two types of measures which can be so constructed should be distinguished:

(1) Average measures, i.e. ordinary measures such as totals, means, rates and proportions constructed by aggregating or averaging individual values. (Examples: area unemployment rate; population proportion in the area having a certain characteristic).

(2) Distributional measures, such as measures of variation or dispersion among households and persons in the region. Such measures may depend on the distribution of characteristics in each region, or on the overall distribution in the whole national (or even EU-level) population.

The patterns of variation and relationship for the two types of measures can differ from each other, and hence involve separate statistical considerations. 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, much 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*.

## 2.3 Focus on more basic of the indicators

It is necessary to adapt the national level Laeken indicators for regional application, taking into account differences in the requirements and the data situation. As a general rule, it is necessary to focus on the more *basic* among the indicators. This is because the data requirements are already increased substantially when the results are to be geographically disaggregated. Detailed disaggregation of the indicators by age, gender and other characteristics - simultaneously with disaggregation by geographical region - has to be severely restricted, especially when the information comes from sample surveys of limited size (such as no more than 1,000-2,000 sample households per region). Broad classifications, such as distinguishing children, youth and elderly persons, may be possible, but even those have to be subsidiary to the need for adequate regional breakdown for the total population.

For the same reason, emphasis has to be shifted away from the study of trends over time and longitudinal measures to essentially *cross-sectional measures*. Furthermore, it is more appropriate to aggregate such measures over suitable time periods, such as over a number of years, so as to illuminate the more stable aspects of the patterns of variation across regions. This is not to preclude longitudinal indicators which are produced at the national level, but to suggest that they should be simplified and consolidated for application at the regional level. For instance, with limitations in the data and the regional sample sizes, looking at longitudinal poverty in terms of persistence of poverty over sets of *two-year periods* may well be more suitable for regional comparisons than the four-year period used in the standard longitudinal indicator adopted in the Laeken list.

Consequently, for the purpose of regional indicators the focus has to be primarily on *the standard poverty rates for the total population*, possibly with some major breakdowns. Certain more complex poverty and inequality measures - measures which are more sensitive to details and irregularities of the empirical income distribution - are less suited to disaggregation to small populations and small samples. Examples are Gini coefficient, relative median at-risk-of-poverty gap, and at-risk-of-poverty rate before social transfers.

On the other hand, poverty rates have to be supplemented by other indicators not considered explicitly in the Laeken list. Perhaps the most important of these is simply the mean income levels of the regions, the dispersion among which provides a measure of regional disparities. General entropy measures may also be useful because they can be decomposed into within and between region components.

## 2.4 Using poverty lines defined at different 'levels' and 'thresholds'

By the 'level of poverty line' we mean the population level to which the income distribution is pooled for the purpose of defining the poverty line. Essentially all poverty related indicators in the Laeken list are based on country poverty lines (defined as 60% of the national median income). The income distribution is considered separately at the level of each country, in relation to which a poverty line is defined and the number (and proportion) of poor computed. These numbers may then be pooled over countries to obtain the EU poverty rate, or disaggregated by region to obtain regional poverty rates - but still defined in terms of national poverty lines.

It is also useful to consider poverty lines at other levels. For instance, we may pool the data across countries to construct a single income distribution (and hence a single poverty line) for the whole EU, and use this to compute poverty rates at the EU level, or for individual countries, or for any level of regions within any country.

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 alternative national poverty rates – but in all cases they remain based on the regional poverty lines. So defined, the poverty measures are not affected by disparities in mean levels of income among the regions. The measures are more purely relative.

In fact, different levels for the poverty line can be seen as implying a different mix of 'relative' and 'absolute' measures. By relative measures we mean those concerning purely the distribution of income, and by absolute measures those concerning income levels. For analysis at the country level, the use of national poverty lines provides a relative measure for each country, but the use of a EU poverty line introduces quite a high degree of absoluteness into the measure.

Considering analysis at a certain regional level (such as NUTS2), the use of the regional poverty line provides a relative measure of poverty determined only by the income distribution within the region, independently of the degree of regional disparities in the country. Use of poverty lines defined at a higher level (such as NUTS1 in this example) introduces an element of 'absoluteness' in the sense defined, since the resulting poverty rate in a NUTS2 region now also depends on differences in income levels among NUTS2 regions in the same NUTS1 region. The degree of absoluteness in the measure increases as the poverty line level is raised to country and then to EU level - meaning that increasingly the resulting poverty rates reflects differences among regions in the *level of mean income*, in addition to the extent of disparity within the regions.

In fact we can mix any level of analysis or aggregation with any poverty line level. The former concerns the units for which the measures are computed; the latter refers to the population of which the income distribution is considered in defining the poverty line.

The poverty line level chosen can make a major difference to the resulting poverty rates when it is higher than the level of analysis or aggregation. The extent depends on the degree of disparity between the units of analysis. However, we find that the poverty line level chosen often makes only a small difference to the resulting poverty rates when it is the same as or lower than the level of analysis or aggregation. For instance, while country poverty rates can differ greatly when a EU poverty line is used, the country rates tend to differ much less whether we use a poverty line defined at the national, NUTS1 or NUTS2 level (Verma *et al*, 2006).

### Poverty line threshold

At any level of analysis, it is also possible to consider several poverty lines defined at different 'thresholds'. By 'poverty line threshold' we mean the percentage of the median income defining the poverty line. Different values of the poverty rate are obtained using different poverty line thresholds. These values may be consolidated, for example by taking an appropriately weighted average, to obtain more robust indicators at the regional level. This idea is developed and illustrated in Section 5.

## 2.5 Placing greater emphasis on indicators of non-monetary deprivation

In addition to the level of monetary income, the standard of living of households and persons can be described by a host of indicators, such as housing conditions, possession of durable goods, the general financial situation, perception of hardship, expectations, norms and values. The data required for the construction of non-monetary indicators are generally simpler to collect than detailed data on monetary incomes. *This makes such indicators more convenient and suitable for regional analysis.* An index of non-monetary deprivation which summarises a range of indicators of living conditions should be developed and analysed in its own right. It is also useful to combine monetary and non-monetary measures in order to study the extent to which they overlap. If individuals are subject both to income poverty and non-monetary deprivation simultaneously, their overall deprivation is more intense. Similarly, if they are subject to only one of the two, their deprivation can, in relative terms, be considered less intense. See for instance, Giorgi and Verma (2002); Betti and Verma (2008).

# 3. Cross-sectional regional indicators based on a single wave: some methodological and practical considerations in variance estimation

We consider in this section cross-sectional regional indicators based on a single wave. Some technical procedures need to be clarified in relation to variance estimation at the regional level.

## 3.1 A variance estimation procedure for EU-SILC

We begin by summarising basic features of the variance estimation procedure, common to any application, whether at the country or the regional level.

### Jackknife Repeated Replication (JRR)

The JRR method has been adopted by Eurostat for EU-SILC. The basic model of the JRR may be summarised as follows. Consider a design in which two or more primary selection units (PSUs) have been selected independently from each stratum in the population. Within each PSU, subsampling of any complexity may be involved, including weighting of the ultimate units. In the "standard" version, each JRR replication is formed by eliminating one PSU from a particular stratum at a time, and increasing the weight of the remaining PSUs in that stratum appropriately, so as to obtain an alternative but equally valid estimate to that obtained from the full sample.

Let z be a full-sample estimate of any complexity, and $z_{(hi)}$ be the estimate produced using the same procedure after eliminating primary unit i in stratum h and increasing the weight of the remaining $(a_h - 1)$ units in the stratum by the factor $g_h = w_h/(w_h - w_{hi})$. Let $z_{(h)}$ be the simple average of the $z_{(hi)}$ over the $a_h$ sample units in h. The variance of z is estimated as [3]:

$$\mathrm{var}(z) = \Sigma_h \left[ \left( \frac{a_h - 1}{a_h} \right) . \Sigma_i \left( z_{(hi)} - z_{(h)} \right)^2 \right]. \qquad (1)$$

[3] The 'finite population correction', trivial in a survey such as EU-SILC, is neglected in (1).

The same relatively simple variance estimation formula holds for z of any complexity. Furthermore, apart from variance estimation of ordinary cross-sectional measures, application of the JRR methodology can be readily extended to more complex indicators based on the EU-SILC rotational panel design. These include longitudinal measures, measures of net change, as well as measures of aggregates and averages over two or more waves (Verma and Betti, 2007). The last-mentioned extension is of particular interest for regional estimates based on cumulation of data over survey waves.

## Defining sample structure: 'computational' strata and PSUs

In many practical situations some aspects of sample structure need to be redefined to make variance computation possible, efficient and stable. Of course, any such redefinition is appropriate only if it does not introduce significant bias in the variance estimation. Such redefinition is often necessary because practical variance estimation methods require the sample design to satisfy certain conditions:

(1) The sample selection is independent between strata.

(2) Two or more primary selections are drawn from each stratum.

(3) These primary selections are drawn at random, independently and with replacement.

(4) The number of primary selections is large enough for valid use of the variance estimation procedure.

Though these basic assumptions regarding the structure of the sample for application of the variance estimation methods are met reasonably well in many EU-SILC surveys, often the assumptions are not met exactly.

A very convenient approach in practice is to summarise the most essential information about the sampling design in the form of two variables, coded for each unit in the microdata file: the 'computational stratum' and the 'computational PSU' to which the unit belongs. This can be done in most cases for the type of sample designs involved in EU-SILC.

The computation stratum has to incorporate all information about the stratification of the PSUs, including both explicit stratification and, where applicable, implicit stratification resulting from systematic sampling of the PSUs. It has also to ensure that each computational stratum contains at least two computational PSUs (which are then assumed to have been selected at random with replacement).

Starting from the actual PSUs, the variable computational PSU should seek to create units reasonably large and uniform in size, and small enough in number so as to avoid excessive computational burden.

Redefinition for the above purposes usually involves some ‘collapsing’ of the sample structure. Some technical procedures for this purpose include: reducing the number of replications formed by deleting units in groups rather than singly as assumed in the basic model; dropping some of the replications from the computation; random grouping PSUs within strata so as to reduce the number of units to be dealt with; grouping PSUs across strata; and grouping PSUs within as well as across strata. It has been demonstrated that appropriately done collapsing usually does not introduce additional bias or variability in the variance estimates (Rust, 1985). Nevertheless, to do the above in a statistically valid way requires sampling expertise.

It is not possible here to go into further technical details of how the required computational strata and PSUs may be defined most appropriately in the case of each EU-SILC national sample design. An extensive discussion may be found in the accompanying publication Verma *et al* (2010).

## 3.2 Special issues in variance estimation for regions and other subpopulations

### Extension of the JRR procedure

Regional indicators are a special case of measures calculated for subpopulations. Each region normally involves a part of the total sample at the level of strata and primary sampling units (PSUs); while in general, subpopulations (such as age groups) refer to any divisions of the sample up to the level of ultimate units (households or individual persons). For variance estimation for subpopulations using the JRR method, the same formulae as those used for the total population apply, except that sample elements which are not members of the subpopulation of interest are simply disregarded. However, some practical aspects need further consideration. When dealing with regions, the main problems are the small numbers of strata and PSUs which may be available for individual regions, and the fact that the regional boundaries may cut across design strata and PSUs. For subpopulations such as age-groups, a similar complication which can arise is that, considering only the subpopulation members, some strata and PSUs may become empty. This problem is unlikely to arise when we merely move from the national level to the level of a geographical region, but are still considering the total population at each level.

Any of these problems would normally require some re-definition of the sample structure for the purpose of variance estimation. Implementing this requires sampling expertise.

If a region coincides with a ‘design domain’ of the sample or is composed only of one or more whole strata, and the statistic of interest is such that it depends only on units within the region (i.e. is independent of the sample in the rest of the country), then variance computations can be performed for each region separately, in exactly the same way as at the national level. Examples of such statistics are equivalised income and other mean values, and of special interest here, poverty rates defined with reference to the regional poverty line.

However, in the context of poverty and inequality, the subpopulation measures of interest, including measures at the regional level, are often of a special type: while all (or some) of the parameters involved in the definition of the measure are estimated from the *full sample*, the measure itself is estimated only for the subpopulation concerned. The most important example is the poverty rate for a subpopulation, but with an individual’s poverty status defined in relation to the poverty line determined from income distribution of the whole population.

Consider the common case when the poverty rate is calculated at, say, NUTS2 level, while the common poverty line used for this purpose is calculated at the national level. The JRR variance estimation procedure can be easily adapted for this purpose as follows. Replications are constructed for the full sample as usual. For each replication, the statistic (such as the poverty rate) is re-estimated only for units in the subpopulation of interest; however, the parameters involved in the definition of the statistic (such as the poverty line) are estimated using all units in the replication of the total sample.

### Stability of variance estimates for indicators at the regional level

In estimating variances at the regional level, one of the main problems is the small numbers of strata and PSUs which may be available in the samples for individual regions. Of course this problem arises only in multi-stage samples. In EU-SILC surveys using direct samples of elements (persons, households, addresses), the samples normally contain numerous units, even for fairly small regions. However, in multi-stage samples, the stability of the variance estimates depends primarily on the number of sample PSUs available, which may be quite small for individual regions. In such situations, it is not wise to rely on the results from individual computations separately. It is preferable to average or smooth the results from many computations in an appropriate way. In order to ameliorate this problem, and also to avoid excessive amount of complex computations when dealing with many regions, we have developed the methodology elaborated later in this paper.

## Limitations owing to availability of information on sample structure

Appropriate coding of the sample structure, most preferably in the survey microdata, is an essential requirement in order to ensure that sampling errors can be computed properly, taking into account the actual sample design. Furthermore, information in the microdata need to be completed with documentation and description of the sampling procedures and the resulting sample. 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.

The major problem in computing sampling errors for EU-SILC is the lack of sufficient information for the purpose: *the UDB does not include information on sample structure, in particular concerning stratification.* Consequently, from UDB variances can be computed only for countries which have employed simple (unstratified) samples of households or persons, or where it is reasonable to approximate the design as simple random sampling of households or persons (e.g., Denmark, Iceland, Austria, Sweden). In a number of countries, stratified random sample of households or persons are used. For these the effect of stratification *may be* relatively small, at least in comparison with that of stratification in multi-stage designs. Examples are Cyprus, Estonia, Lithuania, and also with some approximation, Slovakia, Finland, and Germany (though the last-mentioned survey lacks proper probability sampling).

We are fortunate in having received additional information on sample structure (in particular on explicit stratification, variable DB050) from Eurostat for this and related research. But this information has some major limitations. The most important limitations include the following:

(1) It is available for only a subset of countries.

(2) The specially provided sample structure information can be linked only to the longitudinal dataset in UDB (through common household identifiers, DB030), but not to the cross-sectional dataset because of randomisation of the identifiers in it. No linkage at the micro-level is possible neither between the cross-sectional and longitudinal components, nor across cross-sectional samples for different years.

(3) For regional indicators, the primary interest is in *most recent cross-sectional estimates*. However, information for the identification of even the sample PSUs is missing or incomplete in the cross-sectional data files (it is generally coded only for the newly introduced panels each year, not for the entire cross-sectional data set). Hence the number and sizes of clusters are not known for the cross-sectional samples, and have to be inferred from the corresponding longitudinal datasets. Unfortunately, however, for a given survey round, the longitudinal data become available a year later than the corresponding cross-sectional data.

(4) Another very critical limitation in the present context is that in many countries regional identifiers are not available at all, or are available only to NUTS1 level.

For the set of countries for which special information on sample structure has been provided to us by Eurostat for the purpose of this research, the situation with regard to region identifiers remains poor. Generally, the coding is only to NUTS1 level, at the most. Empirical illustration is limited by the information available in the UDB.

**Availability of information for identification of regions**

| Country | Lowest level of regional identifiers available |
|---|---|
| DE, UK | = MISSING (NUTS2 and lower levels exist, but no information, even on NUTS1, has been recorded) |
| **CZ**, FI, FR* | NUTS2 |
| AT, BE, EL, **PL** | NUTS1 |
| CY, DK, EE, IS, LT, LV, SE, SK | None (Countries not divided into NUTS1 regions, but lower levels of NUTS exist in most cases, but no information is recorded) |

*FR can be excluded, since no information even on stratification has been recorded in UDB.
Numerical illustrations will be provided for CZ and PL in the following sections.

## 3.3 Design effects

Design effect (Kish, 1995) 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:

$$d^2 = v/v_0 , \quad d = se/se_0 . \qquad (2)$$

Computing design effects requires the additional step of estimating the error under simple random sampling ($se_0$), apart from its estimate under the actual design (se).

Proceeding from standard errors to design effects is essential for understanding the patterns of variation and determinants of the magnitude of the error, for smoothing and extrapolating the results for diverse statistics and population subclasses, and for evaluating the performance of the sampling design. Analysing design effects into components also helps to better understand from where inefficiencies of the sample arise, to identify patterns of variation, and through that, to improve 'portability' of the results to other statistics, designs, situations. In applications to EU-SILC, there is in addition a most important and special reason for having procedures for appropriate decomposition of the total design effect into its components. Because of the limited information on sample structure included in the microdata 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 overcoming at least partly the problem due to the lack of information on sample structure in EU-SILC microdata.

All the above reasons apply even more strongly for statistics at the regional level than they do at the country level. Smaller sample sizes and less information at the regional level make the computation of sampling errors more difficult, sometimes impossible. The results of individual computations also tend to be less stable, and therefore there is a greater need for averaging over them. Disaggregation to the regional level also increases greatly the amount of computations involved, unless the results from a limited set of computations can be extrapolated to other statistics and situations. All these operations require 'portable' measures such as individual components of the design effect.

## 3.4 Components of design effect

We may decompose the design effect into components as follows:

$$v = v_0.d^2 = v_0.(d_W.d_H.d_D.d_X)^2. \qquad (3)$$

Here $\mathrm{v_0}$ is the variance (for the statistic concerned) in a simple random sample (SRS) of *individual persons*; $\mathrm{d_W}$ is the effect of sample weights; if relevant, $\mathrm{d_H}$ is the effect of clustering of individual persons into households and $\mathrm{d_D}$ the effect of clustering of households into dwellings; and finally, $\mathrm{d_X}$ is the effect of other complexities of the design, mainly clustering and stratification.

All factors other than $\mathrm{d_X}$ do not involve clusters or strata, but essentially depend only on the number elements (households, persons etc.), and the sample weight associated with each such element in the sample. Hence normally they are well estimated, even for quite small regions. Procedures for estimating components of the design effect are summarized below.

### *Effect of weights* $(d_W)$

The effect of weights $d_W$ does not depend on the sample structure, other than the presence of unequal sample weights for the elementary units of analysis. Weighting generally inflates variance (weighting is primarily introduced to reduce bias). With the complex weighting procedures of EU-SILC, variation in weights can become large, inflating the design effect. This effect needs to be evaluated and controlled. In principle (but rarely in practice) the factor can be <1, for example with particularly efficacious calibration.

### *Clustering of persons within households* $(d_H)$

Factor $d_H$ applies if $v_0$ refers to variance in a simple random sample of individuals, while v refers to a variable measured at the household level.

For example, this factor equals square-root of household size for variables relating to household income when $v_0$ is defined to refer to a SRS of *individual persons*.[4] This applies equally to register and survey countries in EU-SILC, since in both cases income is defined and measured at the household level.

The factor equals 1 for personal interview variables in register countries, since there is only one such interview per household.

For variables constructed to the household level on the basis of separate but correlated observations on individual household members, $d_H$ will be lower than the square-root of household size, depending on the strength of the correlation.

In principle, $d_H < 1$ for variables which are negatively correlated among members of the same household, but this situation is rare.

### *Clustering of persons and households within dwellings* $(d_D)$

The effect of clustering of households within dwellings or addresses is absent ($d_D$=1) when we have a direct sample households or persons, or when such units are selected directly within sample areas - as is the case in most of the EU-SILC surveys. This effect is present when the ultimate units are dwellings, some of which may contain multiple households, but it is small in so far as there is generally a one-to-one correspondence between addresses and households.

[4] Actually, such a design involving a SRS of individual *persons* is never used in EU-SILC, because income of any individuals is defined and measured only in terms of income of all its household members.

Factor $d_D$ cannot be estimated separately unless we have unit identifiers linking households to the dwellings from which they come. Such information has not been recorded in EU-SILC for the few countries using samples of dwellings or addresses.

Note that when the sample has multiple stages, with one or more area stages preceding the selection of dwellings, $d_D$ can be incorporated into the estimation of $d_X$ - the effect of clustering, stratification and other complexities - and hence into the estimation of the overall design effect d; the separation of $d_D$ requires unit identifiers linking households to the dwellings from which they come. (Examples: the, Greece, Latvia, Poland). By contrast, in a direct sample of dwellings, $d_D$ cannot be estimated at all in the absence of linking information, and therefore is neglected in the estimated overall design effect. (Example: Austria.)

*Multi-stage sampling, stratification and other design complexities* $(d_X)$

Factor $d_X$ represents the effect on sampling error of various complexities of the design such as multiple stages and stratification. Normally this effect exceeds 1 because the loss in efficiency of the sample due to clustering tends to be larger than the gain from stratification. We can expect it to be less than 1 in stratified random samples of element.

Components on the design effect other than $d_X$ can be estimated without reference to the information on sample structure, except for weighting and identifiers linking different types of units (e.g. persons with their households). By contrast, computation of $d_X$ requires information on the sample structure linking elementary units to their strata and higher stage units.

## 3.5 Estimating the components of design effect

*Effect of weights* $(d_W)$

A very simple expression for estimating $d_W$ is the following from Kish (1965):

$$d_W^2 = \left[ \frac{n.\sum\left(w_i^2\right)}{\left(\sum w_i\right)^2} \right] = 1 + cv^2\left(w_i\right). \qquad (4)$$

This provides a very good approximation when the sample weights are ‘external’, not correlated with survey variables. Generally it over-estimates the effect.

In situations for which the ‘linearization method’ of variance estimation can be formulated, the effect can be estimated more precisely as:

$$d_W^2 = \left[\frac{n}{\sum w_i}\right] \cdot \frac{\sum\left(w_i^2 \cdot z_i^2\right)}{\sum\left(w_i \cdot z_i^2\right)}.$$

Here $z_i$ is the ‘linearized variate’ corresponding to a complex statistic, which is used in the linearization method to estimate the variance of the complex statistic.

In the linearization variance estimation method, the above-mentioned linearized variate is always in the form $z_i = z_{1i} + t_i$, where: $z_{1i}$ is defined as $z_{1i} = (y_i - r.x_i)$, with the complex statistic concerned written as if it were a simple ration of the form $r = \Sigma w_i.y_i / \Sigma w_i.x_i$; and $t_i$ are additional, generally complex, terms.[5] Now, empirically we have found that the following simpler expression yields values indistinguishable to the above for all the complex statistics encountered in analysis of income inequality and poverty:

$$d_W^2 = \left[\frac{n}{\sum w_i}\right] \cdot \frac{\sum\left(w_i^2 \cdot z_{1i}^2\right)}{\sum\left(w_i \cdot z_{1i}^2\right)}. \qquad (5)$$

The complex variate $z_i$ is available only when the linearisation procedure for variance estimation can be developed, but the simpler $z_{1i}$ is available in most cases.[6] Hence (5) can be used with other procedures such as JRR to estimate the effect of weighting.

### Estimating other components using random grouping of elements

The estimation of design effect due to design complexity with a replication method such as JRR requires an indirect approach. Consider variance computed under the following two assumptions about structure of the design:

(i) Variance (v) under the actual design.

(ii) Using the same procedure, variance (say $v_R$) computed by assuming the design to be (weighted) simple random sampling of elements. This can be estimated from a ‘randomised sample’ created from the actual sample by completely disregarding its structure other than the weight attached to individual elements.

[5] For example, a poverty rate seen as a simple proportion (which is just a special case of a ratio statistic).

[6] However, the above expressions cannot be applied the median and other quantiles of the distribution for which the linearized variable $z_i$ does not contain the term $z_{1i}$. For these we either have to borrow the result from ‘similar’ variables where the above procedure is applicable (see for example Verma and Betti, forthcoming), or use the simpler expression such as the one based on Kish (1965).

For computation (ii), the JRR replications are constructed as in the normal application of the JRR, but in place of the actual strata and primary selections, *random grouping of the sample elements* are used for this purpose. This provides a variance estimate corresponding to a sample of elements i.e. without the effect of stratification, clustering or other complexities, but which still differs from the SRS estimate due to the effect of sample weights on variance. Actually, the result $v_R$ depends on random grouping of *which type of elements* are used. When we have random groupings of *persons* (that, without regard to whether they come from the same or from different households), the variance estimate obtained is:

$$v_R = v_0.d_W^2 . \qquad (6)$$

In a sample of households, we may use random groupings of *households* instead (that is, keeping all members of a household together in the same group), the variance estimate obtained is:

$$v_R = v_0.(d_W.d_H)^2 . \qquad (7)$$

In a sample with dwellings as the ultimate units, *provided that the identifiers linking households to their dwellings are available in the micro-data*, we may use random groupings of *dwelling* (that is, keeping all persons and households of a dwelling together in the same group), and obtain the variance estimate:

$$v_R = v_0.(d_W.d_H.d_D)^2 . \qquad (8)$$

With v computed from the standard application of JRR to the actual design, and $d_W$ estimated from (5), the application of (6) gives $v_0$ - and hence overall design effect $d^2 = v/v_0$ without the need to separate out other components.

If applicable and necessary, the separate components can be obtained from (7) and (8): (7) gives $d_H$ and (8) gives $d_D$. The ratio $(v/v_R)$ gives $(d_X)^2$ if we use (6), gives $(d_X.d_D)^2$ using (7), and gives $(d_X.d_D.d_H)^2$ using (8).

# 4. Empirical illustrations on computing variance and design effects

The first two subsections below present illustrations of variance and design effect computations at the national level. As explained below, because of limitations in the available information on sample structure, it is necessary to first perform the computations for the UDB longitudinal data set (Section 4.1), and then to extrapolate some results from it to complete computations for the full cross-sectional sample (Section 4.2).[7] These illustrations are provided for two countries, Poland and the Czech Republic, for which regional identifiers were also available in the microdata.

Sections 4.3-4.5 extend the results to the regional level, NUTS1 regions in Poland and NUTS2 regions in the Czech Republic. The objective is to illustrate how variance computations at the national level can support and supplement variance computations at the level of regions.

## 4.1 Computing for the longitudinal sample (country level)

On the basis of the additional information provided by Eurostat for the purpose of this research, sampling errors have been computed for illustration for Poland and the Czech Republic, extended to the regional level to be described in the next section.

The sample basis considered are the 2006 sample in the *longitudinal data set* for the year 2006. This data set covers the preceding 2 or 3 years depending on the country. The set of 'rotation groups' included in this data set are those appearing in both the 2006 and 2005 samples, including any which also appeared earlier in 2004. The computations illustrated in Table 1 cover the following three cross-sectional indicators for the year 2006 at the *national level*:

- mean household equivalised income,
- at-risk-of-poverty rate, national poverty line,
- at-risk-of-poverty rate, regional poverty line.

[7] This indirect method restricts the illustrations to the last year for which the longitudinal data were available, 2006 at the time of the present research, even though cross-sectional data (but without the necessary sampling information) were already available for 2007.

## Poverty lines at the regional level

The last-mentioned indicator refers to NUTS1 regions in Poland and to NUTS2 regions in the Czech Republic, these being the lowest level regions for which identifiers are available in the data.

As noted in Section 2.4, ‘level of poverty line’ refers to the population level to which the income distribution is pooled for the purpose of defining the poverty line.

All poverty related indicators in the Laeken list are based on country poverty lines. This applies even when the indicators are aggregated over countries, or are disaggregated to regions within a country. The income distribution is considered separately at the level of each country, in relation to which a poverty line is defined and the number (and proportion) of poor computed. We can disaggregate these numbers of poor by region and obtain regional poverty rates defined according to the national poverty line in each country.

Table 1 also shows results for poverty rates at the national level, but computed from the numbers defined as poor in terms of the *regional poverty line* within each region. That is, a poverty line is 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 alternative national poverty rates, or even further to EU level to produce an EU poverty rate – but in all cases based on the regional poverty lines.

## Defining sample structure for variance estimation

The computational procedures for Table 1 are based on the standard JRR methodology. The main technical task involved was to appropriately (in a statistically valid manner) define ‘computational strata’ and ‘computational PSUs’ for each of the national samples on the basis of available information on the sample structure. The following procedures were applied in the case of the samples of Poland and the Czech Republic for defining the needed computational units.

The sample in Poland is composed of a large number of very small clusters (PSUs), selected from also a large number of strata. The original structure for the longitudinal dataset 2005 contained 4,103 PSUs, many of them with just one or two households. A proportion of the original strata were also very small in size. A more suitable structure for computation of variances thus involved two steps: (i) collapsing of the smallest strata and linking of PSUs across them to create larger computational units; and (ii) random grouping of original PSUs within each of the remaining strata so as to create two ‘computational’ PSUs in each stratum

## Table 1: Estimation of variance and design effects at the national level

**POLAND**

**Longitudinal data set 2006)**

| | Estimate | sample size | | %se* | %se* | Components of design effect | | | | des.eff | %se* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | persons | households | rand | actual | $d_X$ | $d_W$ | $d_H$ | $d_D$ | d | SRS |
| | (1) | (2) | (3) | (4) | (5) | (6)=<br>(5)/(4) | (7) | (8) | (9) | (10)=<br>(6)*(7)*(8)*(9) | (11)=<br>(5)/(10) |
| Mean equivalised disposable income | 3 684 | 33 535 | 10 846 | 0.75 | 0.71 | 0.94 | 1.21 | 1.76 | 1.00 | 2.00 | 0.36 |
| HCR - National poverty line | 18.5 | 33 535 | 10846 | 0.44 | 0.45 | 1.02 | 1.08 | 1.76 | 1.00 | 1.93 | 0.23 |
| HCR - Regional (NUTS1) poverty line | 18.3 | 33 535 | 10846 | 0.52 | 0.55 | 1.05 | 1.08 | 1.76 | 1.00 | 1.98 | 0.28 |

**Full cross-sectional data set (2006)**

| | Estimate | sample size | | %se* | | %se* | | | | | %se* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | persons | households | rand | $d_X$ | actual | $d_W$ | $d_H$ | $d_D$ | d | SRS |
| | (1) | (2) | (3) | (4) | (5) | (6)=<br>(4)*(5) | (7) | (8) | (9) | (10)=<br>(5)*(7)*(8)*(9) | (11)=<br>(6)/(10) |
| Mean equivalised disposable income | 3 704 | 45 122 | 14 914 | 0.61 | 0.94 | 0.57 | 1.22 | 1.74 | 1.00 | 1.99 | 0.29 |
| HCR - National poverty line | 19.1 | 45 122 | 14914 | 0.50 | 1.02 | 0.51 | 1.09 | 1.74 | 1.00 | 1.94 | 0.26 |
| HCR - Regional (NUTS1) poverty line | 19.0 | 45 122 | 14914 | 0.58 | 1.05 | 0.61 | 1.09 | 1.74 | 1.00 | 1.99 | 0.30 |

**CZECH REPUBLIC**

**Longitudinal data set (2006)**

| | Estimate | sample size | | %se* | %se* | Components of design effect | | | | des.eff | %se* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | persons | households | rand | actual | $d_X$ | $d_W$ | $d_H$ | $d_D$ | d | SRS |
| | (1) | (2) | (3) | (4) | (5) | (6)=<br>(5)/(4) | (7) | (8) | (9) | (10)=<br>(6)*(7)*(8)*(9) | (11)=<br>(5)/(10) |
| Mean equivalised disposable income | 5 434 | 9 287 | 3 852 | 1.12 | 1.14 | 1.02 | 1.16 | 1.55 | 1.00 | 1.83 | 0.62 |
| HCR - National poverty line | 10.2 | 9 287 | 3852 | 0.66 | 0.66 | 1.00 | 1.23 | 1.55 | 1.00 | 1.91 | 0.35 |
| HCR - Regional (NUTS2) poverty line | 9.9 | 9 287 | 3852 | 1.08 | 0.83 | 0.77 | 1.25 | 1.55 | 1.00 | 1.49 | 0.56 |

**Full cross-sectional data set (2006)**

| | Estimate | sample size | | %se* | | %se* | | | | | %se* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | persons | households | rand | $d_X$ | actual | $d_W$ | $d_H$ | $d_D$ | d | SRS |
| | (1) | (2) | (3) | (4) | (5) | (6)=<br>(4)*(5) | (7) | (8) | (9) | (10)=<br>(5)*(7)*(8)*(9) | (11)=<br>(6)/(10) |
| Mean equivalised disposable income | 5 403 | 17 830 | 7 483 | 0.90 | 1.02 | 0.92 | 1.20 | 1.54 | 1.00 | 1.88 | 0.49 |
| HCR - National poverty line | 9.8 | 17 830 | 7483 | 0.57 | 1.00 | 0.57 | 1.23 | 1.54 | 1.00 | 1.90 | 0.30 |
| HCR - Regional (NUTS2) poverty line | 9.7 | 17 830 | 7483 | 0.64 | 0.77 | 0.50 | 1.27 | 1.54 | 1.00 | 1.51 | 0.33 |

See notes below.

**Notes to Table 1:**

| | |
|---|---|
| %se* | For mean statistics e.g. equivalised disposable income, error is expressed as percentage of the mean value. |
| | For proportions and rates (e.g. poverty rates), error is given as absolute percentage points (pp). |
| | Terms (%se* actual), (%se* rand) and (%se SRS) relate, respectively, to the variances v, $v_R$ and $v_0$ in the text. |
| $d$ | Overall design effect |
| | ***Components of design effect:*** |
| $d_X$ | design effect due to clustering and stratification of ultimate sampling units (dwellings or households) |
| $d_W$ | effect of unequal sample weights |
| $d_H$ | effect of clustering of persons within households |
| $d_D$ | effect of clustering of households within dwellings (if applicable) |

The computations refer to 2006 data in the 2-year (2005-2006) panel.

In PL for example, standard error (col. 5) for mean equivalised disposable income is 0.71% of the mean value (euro 3,686). For at-risk-of-poverty rate of 18.5%, standard error is 0.45 in (absolute) percentage points (implying a 95% confidence interval of 17.6-19.4%, for instance).

Col. (4) gives standard error computed by ignoring any clustering and stratification of the ultimate sampling units (dwellings or households). The ratio of the actual to this 'randomised sample' standard error (col. 6) isolates the effect of clustering and stratification of dwellings/households in the sample.

Col. (11) is an estimate of standard error which would be obtained in a simple random sample of *persons*, of the same size as shown in col. (2).

For this purpose, we first sorted the dataset by region and inside each region sorted the strata by the number of household in each. Then we merged the smallest strata to create new strata such that the minimum number of household in each of them was 30 and the new strata did not cross regional boundaries. Then, within each of these strata, as well as the other (original) strata, the existing PSUs were randomly grouped, finally giving 426 computational PSUs in 213 computational strata.

In the Czech Republic longitudinal dataset 2005, mostly each original stratum already had an appropriate number of households suitable for the variance computations, so that the existing strata could be directly used as ‘computational strata’. The only step involved was their renumbering and sorting by region. By contrast, the original PSUs were generally very small in terms of the number of households. Within each stratum, the original PSUs were collapsed (randomly grouped) such that each PSU contained 16 households at the minimum so long as each stratum contained at least two computational PSUs.

## 4.2 Variance estimation for the full cross-sectional sample (country level)

Table 1 also shows variance and design effect estimations for the same three variables for Poland and the Czech Republic for the full cross-sectional sample for 2006.

The major additional problem in computing variances for the full cross-sectional sample is that the additional information on sample structure which was provided to us by Eurostat especially for the purpose of this research can be linked only to the longitudinal microdata in UDB (this linkage is through common household identifiers, DB030), but *it cannot be linked to the cross-sectional data set because the household identifiers in the latter have been randomised*.

In the absence of information on the sample structure, the effect of clustering and stratification of households, $\mathrm{d}_{\mathrm{X(C)}}$, cannot be directly estimated for the cross-sectional sample base. We need to impute or infer somehow this quantity from computed value $\mathrm{d}_{\mathrm{X(L)}}$ based on the longitudinal data set. We describe in the next subsection a simplified model which links the $\mathrm{d}_{\mathrm{X}}$ values for two samples with similar design applied on the same population. On the basis of that model $\mathrm{d}_{\mathrm{X(C)}}$ can be inferred from its longitudinal counterpart $\mathrm{d}_{\mathrm{X(L)}}$ through respective cluster sizes of the two samples. However, even this simple model cannot be used in the case of EU-SILC since the number of sample clusters and hence the average size per cluster is not available in the data for the cross-sectional sample. Table 1 has been constructed on the assumption (very reasonable one in this case) that $d_{X(C)} = d_{X(L)}$.

The practically important point to note is that the complexity of the sample design at stages above the selection (and weighting) of households, i.e. the complexity resulting from stratification and clustering of households, is represented by factor $d_X$ only; *all other components of design effect are independent of this complexity, and hence can be estimated despite the lack of information on sample structure* in EU-SILC data files, so long as there is information for the identification of individual persons and households (and if relevant, dwelling units), and their sample weights.

Hence for the cross-sectional sample in the table, cols. (4) and (7)-(9) are computed directly as they do not involve the sample structure. Parameter $d_X$ in col. (5) is taken from the corresponding figure from the longitudinal sample, and on this basis actual standard error in col. (6), and hence also cols. (10) and (11) can be estimated.

## 4.3 Design effect due to clustering and stratification ($d_X$) at the regional level

As noted earlier, in estimating variances at the regional level, one of the main problems is the small numbers of strata and PSUs which may be available in the regional samples. In multi-stage samples, the stability of the variance estimates depends primarily on the number of sample PSUs available, which can be quite small for individual regions, making results from individual computations unstable. Simplifications are also desirable in order to avoid excessive amount of complex computations in dealing with many regions.

All factors other than $d_X$ do not involve clusters or strata, but essentially depend only on the number elements (household, persons etc.) in the sample. Hence normally these factors are well estimated, even for quite small regions.

Factor $d_{X(G)}$ for a region may be estimated in relation to $d_{X(C)}$ estimated at the country level on the following lines.

(1) For large regions, each with a large enough number (say over 25 or 30) of PSUs, we may estimate v for the actual sample, and hence $d_{X(G)}$ directly at the regional level.

(2) Sometimes a region involves a SRS of elements, even if the national sample is multi-stage in other parts; here obviously, $d_{X(G)} = 1$.

(3) 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 $d_{x(G)} = d_{x(C)}$.

(4) It is common that the main difference between the regional and the total samples is the average cluster size (b). In this case we may use the relationship:

$$d^2_{x(G)} = 1 + \left(d^2_{x(C)} - 1\right) . \frac{b_{(G)} - 1}{b_{(C)} - 1} \qquad (9)$$

This relationship is based on the (often very reasonable assumption) that, for a given variable, the intra-cluster correlations in the region and the total country are the same, $\text{roh}_{(G)} = \text{roh}_{(C)}$. By definition, the intra-cluster correlation relates to design effect as: $d^2_x = 1 + (b - 1).\text{roh}$, giving (9). A convenient simplification to (9) is:

$$d^2_{x(G)} = 1 + \left(d^2_{x(C)} - 1\right) . \frac{b_{(G)}}{b_{(C)}} \qquad (10)$$

The above 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 or when units within clusters are negatively correlated (both these situations are rare, but not impossible), or simply as a result of random variability in the empirical results. In any case, if $d_{x(C)} < 1$, (10) should be replaced by:

$$d_{x(G)} = d_{x(C)} . \qquad (11)$$

In Table 2, we have used (10), or (11) where applicable, in estimating $d_{x(G)}$ for regions from its estimate $d_{x(C)}$ at the country level.

(5) Sometimes there may be more profound differences in the regional and the overall national designs than simply differences in the average cluster size: for example the nature of clusters and the type of subsampling within clusters may be different. This would affect the intra-cluster correlations. But often it is still reasonable to assume that the ratio of the intra-cluster correlations for the region and the country is similar for different variables. This gives a model of the form:

$$d^2_{X(G)} = 1 + c_{(G)}.\left(d^2_{X(C)} - 1\right).\frac{b_{(G)}}{b_{(C)}}, \qquad (12)$$

where $c_{(G)}$ is the average over variables (or over each of a group of variables) of the ratios regional to national intra-cluster correlations. We can obtain these ratios by performing detailed computation for individual variables, taking their average, and then using the average in (12) to estimate smoothed values of the design effect component $d_X$. Often, a smoothed value provides a better estimate than the raw computations for individual variables. Whether it is sufficient to apply this procedure to all the variables together in a single group, or it is necessary to construct more than one groups of variables for the purpose is an empirical issue.

## 4.4 Standard error under a simple random sample (%se* SRS)

If desired, quantity (%se* SRS) in col. (10) of Table 2 can be directly computed at the regional level as was done for the national level in Table 1 via cols. (4) and (7)-(9), using equation (8), giving in terms of the notations used in the table:

$$(\%\text{se}*\text{rand})^2 = v_R; \quad (\%\text{se}*\text{SRS})^2 = v_0;$$

and

$$v_R = v_0.\left(d_W.d_H.d_D\right)^2.$$

None of the above quantities requires reference to the structure of the sample. However, the above requires JRR computations of $v_R$ for each variable over each region, which can be a heavy task if there are many regions and variables involved. Fortunately, very good approximations can be usually obtained simply. The following model has been used in Table 2. For means such as mean equivalised income over very similar populations, assumption of a constant coefficient of variation is a reasonable one. With this assumption, the region-to-country ratio of relative standard errors (expressed as percentage of the mean value as in Table 2) under simple random sampling is inversely proportional to the square-root of their respective sample sizes:

$$(\%\text{se}*\text{SRS})^2_{(G)} = (\%\text{se}*\text{SRS})^2_{(C)}.\left(n_{(C)}/n_{(G)}\right)$$

For proportions (p, with q = 1-p), with standard error expressed in absolute percentage points (pp) as in Table 2, the corresponding relationship is:

$$(\%se*SRS)^2_{(G)} = (\%se*SRS)^2_{(C)} \cdot \left(\frac{p_{(G)}.q_{(G)}}{p_{(C)}.q_{(C)}}\right). \left(n_{(C)}/n_{(G)}\right)$$

At-risk-of-poverty rates may be treated as proportions for the purpose of applying the above.

## 4.5 Actual standard errors for regional estimates

With standard error corresponding to a simple random sample (%se* SRS) and the effect of clustering and stratification $(d_X)$ imputed for regions as explained above, and the other components of design effect computed direct without the need to refer to the sample structure other than weights, equation (3) has been used to compute standard errors for regional estimates:

$$(\%se*) = (\%se*SRS). (d_W.d_H.d_D.d_X)^{2} . \qquad (13)$$

The results are shown in col. (16) in Table 2:

**Table 2: Estimation of variance and design effects at the regional level**
**2(PL): POLAND NUTS1 regions**

| 2(PL). POLAND NUTS1 regions | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Longitudinal data set (2006) | | | | | Full cross-sectional data set (2006) | | | | | | | | | |
| | | Sample size | | Cluster | Relative | | Estimate | Sample size | | %se* | Components of design effect (d) | | | | | %se* |
| | | persons | clusters | size | size | $d_X$ | | persons | households | SRS | $d_X$ | $d_W$ | $d_H$ | $d_D$ | d | actual |
| | (1) | (2) | (3) | (4)=<br>(2)/(3) | (5)=<br>(4)/(4C) | (6) | (7) | (8) | (9) | (10) | (11)=<br>(6) | (12) | (13) | (14) | (15)=(11)*<br>(12)*(13)*(14) | (16)=<br>(10)*(15) |
| **Mean equivalised disposable income** | | | | | | | | | | | | | | | | |
| Poland | | 33 535 | 3 863 | **8.7** | 1.00 | **0.94** | 3 704 | 45 122 | 14 914 | **0.29** | 0.94 | 1.22 | 1.74 | 1.00 | 1.99 | 0.57 |
| Regions | PL1 | 6 509 | 805 | 8.1 | 0.93 | 0.94 | 4 236 | 8 728 | 3 001 | 0.65 | 0.94 | 1.28 | 1.71 | 1.00 | 2.06 | 1.34 |
| | PL2 | 6 973 | 894 | 7.8 | 0.90 | 0.94 | 3 889 | 9 273 | 3 114 | 0.63 | 0.94 | 1.10 | 1.73 | 1.00 | 1.78 | 1.13 |
| | PL3 | 6 812 | 621 | 11.0 | 1.26 | 0.94 | 3 162 | 9 079 | 2 862 | 0.64 | 0.94 | 1.20 | 1.78 | 1.00 | 2.00 | 1.28 |
| | PL4 | 4 967 | 565 | 8.8 | 1.01 | 0.94 | 3 530 | 6 912 | 2 225 | 0.73 | 0.94 | 1.14 | 1.76 | 1.00 | 1.90 | 1.39 |
| | PL5 | 3 393 | 412 | 8.2 | 0.95 | 0.94 | 3 906 | 4 538 | 1 563 | 0.90 | 0.94 | 1.22 | 1.70 | 1.00 | 1.96 | 1.77 |
| | PL6 | 4 881 | 573 | 8.5 | 0.98 | 0.94 | 3 419 | 6 592 | 2 149 | 0.75 | 0.94 | 1.15 | 1.75 | 1.00 | 1.90 | 1.43 |
| **At-risk-of-poverty rate, national poverty line** | | | | | | | | | | | | | | | | |
| Poland | | 33 535 | 3 863 | 8.7 | 1.00 | **1.02** | 19.1 | 45 122 | 14 914 | **0.26** | 1.02 | 1.09 | 1.74 | 1.00 | 1.94 | 0.51 |
| Regions | PL1 | 6 509 | 805 | 8.1 | 0.93 | 1.02 | 17.1 | 8 728 | 3 001 | 0.57 | 1.02 | 1.07 | 1.71 | 1.00 | 1.85 | 1.06 |
| | PL2 | 6 973 | 894 | 7.8 | 0.90 | 1.02 | 14.7 | 9 273 | 3 114 | 0.52 | 1.02 | 1.06 | 1.73 | 1.00 | 1.86 | 0.97 |
| | PL3 | 6 812 | 621 | 11.0 | 1.26 | 1.02 | 25.2 | 9 079 | 2 862 | 0.64 | 1.02 | 1.14 | 1.78 | 1.00 | 2.09 | 1.34 |
| | PL4 | 4 967 | 565 | 8.8 | 1.01 | 1.02 | 18.7 | 6 912 | 2 225 | 0.66 | 1.02 | 1.10 | 1.76 | 1.00 | 1.98 | 1.32 |
| | PL5 | 3 393 | 412 | 8.2 | 0.95 | 1.02 | 18.6 | 4 538 | 1 563 | 0.82 | 1.02 | 1.10 | 1.70 | 1.00 | 1.91 | 1.56 |
| | PL6 | 4 881 | 573 | 8.5 | 0.98 | 1.02 | 21.4 | 6 592 | 2 149 | 0.71 | 1.02 | 1.10 | 1.75 | 1.00 | 1.95 | 1.40 |
| **At-risk-of-poverty rate, regional poverty lines** | | | | | | | | | | | | | | | | |
| Poland | | 33 535 | 3 863 | 8.7 | 1.00 | **1.05** | 19.0 | 45 122 | 14 914 | **0.30** | 1.05 | 1.09 | 1.74 | 1.00 | 1.99 | 0.61 |
| Regions | PL1 | 6 509 | 805 | 8.1 | 0.93 | 1.04 | 19.8 | 8 728 | 3 001 | 0.70 | 1.04 | 1.07 | 1.71 | 1.00 | 1.90 | 1.34 |
| | PL2 | 6 973 | 894 | 7.8 | 0.90 | 1.04 | 18.5 | 9 273 | 3 114 | 0.67 | 1.04 | 1.06 | 1.73 | 1.00 | 1.91 | 1.27 |
| | PL3 | 6 812 | 621 | 11.0 | 1.26 | 1.06 | 18.6 | 9 079 | 2 862 | 0.68 | 1.06 | 1.13 | 1.78 | 1.00 | 2.14 | 1.45 |
| | PL4 | 4 967 | 565 | 8.8 | 1.01 | 1.05 | 17.5 | 6 912 | 2 225 | 0.76 | 1.05 | 1.10 | 1.76 | 1.00 | 2.04 | 1.54 |
| | PL5 | 3 393 | 412 | 8.2 | 0.95 | 1.04 | 20.9 | 4 538 | 1 563 | 1.00 | 1.04 | 1.10 | 1.70 | 1.00 | 1.97 | 1.96 |
| | PL6 | 4 881 | 573 | 8.5 | 0.98 | 1.05 | 19.1 | 6 592 | 2 149 | 0.80 | 1.05 | 1.09 | 1.75 | 1.00 | 2.00 | 1.60 |

For longitudinal sample, $d_X$ in col. (6) for regions is estimated from its value at the country level, modified taking into account the relative regional to country cluster sizes in col. (5), using equation (10) as explained in the text. This value of $d_X$ is then carried over to the cross-sectional sample for the region concerned. Similarly, %se(SRS) in col. (10) for regions is estimated from its value at the country level, on the basis of statistical considerations as explained in the text. For the remaining, see notes to Table 1.

## 2(CZ): THE CZECH REPUBLIC NUTS2 regions

| 2(CZ). CZECH REPUBLIC NUTS2 regions | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Longitudinal data set (2006) | | | | | Full cross-sectional data set (2006) | | | | | | | | | |
| | | Sample size | | Cluster | Relative | | Estimate | Sample size | | %se* | Components of design effect (d) | | | | | %se* |
| | | persons | clusters | size | size | $d_X$ | | persons | households | SRS | $d_X$ | $d_W$ | $d_H$ | $d_D$ | d | actual |
| | (1) | (2) | (3) | (4)=<br>(2)/(3) | (5)=<br>(4)/(4C) | (6) | (7) | (8) | (9) | (10) | (11)=<br>(6) | (12) | (13) | (14) | (15)=(11)*<br>(12)*(13)*(14) | (16)=<br>(10)*(15) |
| **Mean equivalised disposable income** | | | | | | | | | | | | | | | | |
| Czech Republic | | 9 287 | 681 | **13.6** | 1.00 | **1.02** | 5 403 | 17 830 | 7 483 | **0.49** | 1.02 | 1.20 | 1.54 | 1.00 | 1.88 | 0.92 |
| Regions | CZ01 | 776 | 87 | 8.9 | 0.65 | 1.01 | 6 988 | 1 456 | 676 | 1.70 | 1.01 | 1.18 | 1.47 | 1.00 | 1.75 | 2.99 |
| | CZ02 | 838 | 71 | 11.8 | 0.87 | 1.01 | 5 696 | 1 736 | 751 | 1.56 | 1.01 | 1.05 | 1.52 | 1.00 | 1.61 | 2.51 |
| | CZ03 | 1 071 | 81 | 13.2 | 0.97 | 1.01 | 5 507 | 2 142 | 913 | 1.41 | 1.01 | 1.08 | 1.53 | 1.00 | 1.67 | 2.35 |
| | CZ04 | 1 015 | 80 | 12.7 | 0.93 | 1.01 | 5 168 | 1 996 | 861 | 1.46 | 1.01 | 1.18 | 1.52 | 1.00 | 1.83 | 2.66 |
| | CZ05 | 1 413 | 101 | 14.0 | 1.03 | 1.02 | 5 144 | 2 632 | 1 086 | 1.27 | 1.02 | 1.11 | 1.56 | 1.00 | 1.75 | 2.22 |
| | CZ06 | 1 526 | 105 | 14.5 | 1.07 | 1.02 | 5 102 | 2 888 | 1 151 | 1.21 | 1.02 | 1.08 | 1.58 | 1.00 | 1.75 | 2.11 |
| | CZ07 | 1 287 | 80 | 16.1 | 1.18 | 1.02 | 4 980 | 2 408 | 954 | 1.33 | 1.02 | 1.13 | 1.59 | 1.00 | 1.83 | 2.42 |
| | CZ08 | 1 361 | 82 | 16.6 | 1.22 | 1.02 | 4 864 | 2 572 | 1 091 | 1.28 | 1.02 | 1.18 | 1.54 | 1.00 | 1.85 | 2.38 |
| **At-risk-of-poverty rate, national poverty line** | | | | | | | | | | | | | | | | |
| Czech Republic | | 9 287 | 681 | 13.6 | 1.00 | **1.00** | 9.8 | 17 830 | 7 483 | **0.30** | 1.00 | 1.23 | 1.54 | 1.00 | 1.90 | 0.57 |
| Regions | CZ01 | 776 | 87 | 8.9 | 0.65 | 1.00 | 4.7 | 1 456 | 676 | 0.75 | 1.00 | 1.21 | 1.47 | 1.00 | 1.78 | 1.32 |
| | CZ02 | 838 | 71 | 11.8 | 0.87 | 1.00 | 8.5 | 1 736 | 751 | 0.90 | 1.00 | 1.33 | 1.52 | 1.00 | 2.02 | 1.81 |
| | CZ03 | 1 071 | 81 | 13.2 | 0.97 | 1.00 | 6.1 | 2 142 | 913 | 0.69 | 1.00 | 1.22 | 1.53 | 1.00 | 1.87 | 1.29 |
| | CZ04 | 1 015 | 80 | 12.7 | 0.93 | 1.00 | 16.0 | 1 996 | 861 | 1.10 | 1.00 | 1.15 | 1.52 | 1.00 | 1.75 | 1.93 |
| | CZ05 | 1 413 | 101 | 14.0 | 1.03 | 1.00 | 8.9 | 2 632 | 1 086 | 0.74 | 1.00 | 1.05 | 1.56 | 1.00 | 1.63 | 1.22 |
| | CZ06 | 1 526 | 105 | 14.5 | 1.07 | 1.00 | 8.3 | 2 888 | 1 151 | 0.69 | 1.00 | 1.19 | 1.58 | 1.00 | 1.90 | 1.30 |
| | CZ07 | 1 287 | 80 | 16.1 | 1.18 | 1.00 | 11.2 | 2 408 | 954 | 0.86 | 1.00 | 1.32 | 1.59 | 1.00 | 2.10 | 1.81 |
| | CZ08 | 1 361 | 82 | 16.6 | 1.22 | 1.00 | 15.5 | 2 572 | 1 091 | 0.96 | 1.00 | 1.36 | 1.54 | 1.00 | 2.10 | 2.01 |
| **At-risk-of-poverty rate, regional poverty lines** | | | | | | | | | | | | | | | | |
| Czech Republic | | 9 287 | 681 | 13.6 | 1.00 | **0.77** | 9.7 | 17 830 | 7 483 | **0.33** | 0.77 | 1.27 | 1.54 | 1.00 | 1.51 | 0.50 |
| Regions | CZ01 | 776 | 87 | 8.9 | 0.65 | 0.77 | 12.1 | 1 456 | 676 | 1.27 | 0.77 | 1.19 | 1.47 | 1.00 | 1.35 | 1.70 |
| | CZ02 | 838 | 71 | 11.8 | 0.87 | 0.77 | 10.4 | 1 736 | 751 | 1.08 | 0.77 | 1.34 | 1.52 | 1.00 | 1.57 | 1.70 |
| | CZ03 | 1 071 | 81 | 13.2 | 0.97 | 0.77 | 8.3 | 2 142 | 913 | 0.88 | 0.77 | 1.17 | 1.53 | 1.00 | 1.38 | 1.21 |
| | CZ04 | 1 015 | 80 | 12.7 | 0.93 | 0.77 | 13.1 | 1 996 | 861 | 1.12 | 0.77 | 1.18 | 1.52 | 1.00 | 1.39 | 1.55 |
| | CZ05 | 1 413 | 101 | 14.0 | 1.03 | 0.77 | 7.8 | 2 632 | 1 086 | 0.77 | 0.77 | 1.07 | 1.56 | 1.00 | 1.28 | 0.99 |
| | CZ06 | 1 526 | 105 | 14.5 | 1.07 | 0.77 | 7.0 | 2 888 | 1 151 | 0.70 | 0.77 | 1.21 | 1.58 | 1.00 | 1.47 | 1.04 |
| | CZ07 | 1 287 | 80 | 16.1 | 1.18 | 0.77 | 9.2 | 2 408 | 954 | 0.87 | 0.77 | 1.36 | 1.59 | 1.00 | 1.66 | 1.45 |
| | CZ08 | 1 361 | 82 | 16.6 | 1.22 | 0.77 | 11.5 | 2 572 | 1 091 | 0.93 | 0.77 | 1.46 | 1.54 | 1.00 | 1.72 | 1.60 |

# 5. Consolidation of measures over different poverty line thresholds

## 5.1 Substantive and statistical considerations

In the standard analysis, as for instance in Laeken indicators, the poverty line is defined as a certain percentage (threshold x%) of the median income of the national population. As noted, a 'poverty line threshold' refers the percentage of the median income defining the poverty line, and different values of the poverty rate are obtained depending on the threshold (i.e. on 'x') of the chosen poverty line. The Laeken set of indicators at the national level includes a measure of dispersion around the at-risk-of-poverty threshold, computing the percentage of persons in the population with an equivalised disposable income below, respectively, 40%, 50%, 60% and 70% of the national median equivalised disposable income, 60% being the main threshold. The substantive objective of introducing indicators of dispersion around the poverty line is to take more fully into account differences among countries in the shape at the lower end of the income distribution. Higher thresholds identify broad disadvantaged groups. Lower thresholds isolate the more severely poor and tend to be more sensitive in distinguishing among countries or other population groups being compared. As the threshold is raised, this sensitivity tends to fall: clearly in the extreme case when 'x' is taken as 100% (poverty line equal to the median), the poverty rate in all situations is 50%, by definition.

In addition to the above systematic differences, the results from using different poverty line thresholds are also likely to be affected by irregularities in the empirical income distribution. Irregularities are larger when the distributions are estimated from smaller samples, as normally is the case for disaggregated estimates by region. It is this consideration which is likely to dominate in the context of constructing regional measures.

In view of the reduced sample sizes in moving to the regional level, it is desirable to avoid producing too many individual figures each subject to large sampling variability. Instead, it would seem a better idea to compute poverty rates with reference to several different thresholds, but then to consolidate them, such as by taking an appropriately weighted average, for comparisons across regions. In specific terms, a single measure based on suitable consolidation over poverty lines defined as, say, 50%, 60% and 70% of median, would be preferable to separate indicators for each of these levels.

There are also substantive considerations in such consolidation. The rate consolidated over different thresholds provides a *summary or overall measure* of different degrees of severity of poverty contained within the given income distribution.

The procedure described below gives relatively higher weights to poverty rates defined with reference to poverty lines at lower thresholds of the median. This aspect is in fact a substantively desirable one, since those rates correspond to more acute conditions of poverty.

## 5.2 Potential gain in precision

Some gain in sampling precision can be obtained by computing poverty rates using different thresholds, and then taking their weighted average using some appropriate pre-specified (i.e., constant or external) weights. A quantitative indication of the magnitude of this gain may be obtained on the following lines.

Consider three poverty line thresholds, giving poverty rates:

$$p_i, \quad p_1 < p_2 < p_3 .$$

With fixed weights $W_i$, $\Sigma W_i = 1$, a consolidated rate is computed as $p = \Sigma W_i . p_i$ .

For simplicity, in the following exposition we will take the sample as a simple random sample (SRS) and approximate the complex statistic ‘poverty rate’ as an ordinary proportion. This assumption is not likely to be consequential, since the design effects due to departures from SRS are likely to be very similar for the various statistics being considers. Neglecting them should not substantially affect the conclusions. On this basis we can compare the precision of the weighted average of poverty rates at different thresholds against that of the conventional measure defined with referent to a single, usually 60% of the median, threshold. The effect of design complexities common to both can then be brought in, if desired, to estimate the actual variances of the measures.

Under the above model, variance of the consolidate poverty rate p is given by:

$$\mathrm{var}(p) = \Sigma_i W_i^2 . \mathrm{var}(p_i) + 2.\Sigma_{j<i} W_i W_j . \mathrm{cov}(p_i, p_j) . \qquad (14)$$

By considering the poverty indicator variables $p_{i,k} = \{0,1\}$ for individuals 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: $\mathrm{var}(p_2) = p_2.(1 - p_2)$. The ratio:

$$g_V = (\mathrm{var}(p)/\mathrm{var}(p_2))^{1/2}, \tag{15}$$

gives the required factor by which the standard error is reduced.

The ‘constant’ weights may come from poverty rates estimated at the country level, and then the same weights applied to each region. An appropriate choice for the weights is the following:

$$W_1 = \frac{1}{3}.\left(\frac{p_2}{p_1}\right), \quad W_2 = \frac{1}{3}, \quad W_3 = \frac{1}{3}.\left(\frac{p_2}{p_3}\right) \tag{16}$$

where subscripts 1, 2 and 3 refer to the rates computed at the national level with poverty line thresholds, respectively, as 50%, 60% and 70% of the national median equivalised income.

## Table 3: Weighted average of poverty rates at different thresholds: gain in precision

| | | | **At-risk-of-poverty rate, national poverty line** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Sample size** | **Poverty line threshold** | | | | | | | | |
| | | **(persons)** | **50% of** | **60% of median** | | | **70% of** | **Weighted estimate (A)** | | | |
| | | | **median** | **Estimate (P)** | **"V(P)"** | **%se*** | **median** | **estimate** | **"V(A)"** | **gain** | **%se*** |
| | | **(1)** | **(2)** | **(3)** | **(4)=<br>(3)*(100-(3))** | **(5)** | **(6)** | **(7)** | **(8)** | **(9)=<br>(8)/(4)** | **(10)=<br>(5)*sqrt(9)** |
| | | | | | **Poland** | | | | | | |
| Country | | 45 122 | 12.3 | **19.1** | 1 545 | 0.51 | 26.9 | **19.1** | 1 355 | 0.88 | 0.48 |
| Regions | PL1 | 8 728 | 11.2 | 17.1 | 1 420 | 1.06 | 11.2 | 14.1 | 1 284 | 0.90 | 1.00 |
| (NUTS1) | PL2 | 9 273 | 9.7 | 14.7 | 1 253 | 0.97 | 9.7 | 12.2 | 1 136 | 0.91 | 0.92 |
| | PL3 | 9 079 | 16.0 | 25.2 | 1 886 | 1.34 | 35.0 | 25.0 | 1 624 | 0.86 | 1.25 |
| | PL4 | 6 912 | 11.1 | 18.7 | 1 518 | 1.32 | 26.7 | 18.3 | 1 273 | 0.84 | 1.20 |
| | PL5 | 4 538 | 12.1 | 18.6 | 1 515 | 1.56 | 24.7 | 18.4 | 1 347 | 0.89 | 1.47 |
| | PL6 | 6 592 | 14.3 | 21.4 | 1 682 | 1.40 | 28.5 | 21.3 | 1 509 | 0.90 | 1.32 |
| weight | | | 0.519 | 0.333 | | | 0.237 | | | | |
| | | | | | **Czech Republic** | | | | | | |
| Country | | 17 830 | 4.9 | **9.8** | 884 | 0.57 | 17.9 | **9.8** | 745 | 0.84 | 0.52 |
| Regions | CZ01 | 1 456 | 1.9 | 4.7 | 449 | 1.32 | 9.2 | 4.5 | 337 | 0.75 | 1.15 |
| (NUTS2) | CZ02 | 1 736 | 4.9 | 8.5 | 774 | 1.81 | 15.9 | 9.0 | 719 | 0.93 | 1.74 |
| | CZ03 | 2 142 | 3.4 | 6.1 | 569 | 1.29 | 11.8 | 6.4 | 522 | 0.92 | 1.24 |
| | CZ04 | 1 996 | 6.2 | 16.0 | 1 344 | 1.93 | 24.1 | 13.8 | 952 | 0.71 | 1.62 |
| | CZ05 | 2 632 | 4.7 | 8.9 | 810 | 1.22 | 17.9 | 9.3 | 706 | 0.87 | 1.13 |
| | CZ06 | 2 888 | 3.9 | 8.3 | 761 | 1.30 | 17.5 | 8.5 | 613 | 0.81 | 1.17 |
| | CZ07 | 2 408 | 5.5 | 11.2 | 991 | 1.81 | 21.2 | 11.3 | 826 | 0.83 | 1.65 |
| | CZ08 | 2 572 | 9.2 | 15.5 | 1 311 | 2.01 | 25.4 | 15.9 | 1 226 | 0.93 | 1.94 |
| weight | | | 0.662 | 0.333 | | | 0.182 | | | | |

Source: EU-SILC Users' database

Reading note: Col. (7) is computed as the weighted sum of cols. (2), (3) and (6), with weights given in the last row of the panel for each country. Cols.(4) and (8) are estimates of population variances of cols. (3) and (7) respectively, assuming them to be simple proportions under simple random sampling. Square root of their ratio in col. (9) is applied to actual standard error (se) of the conventional rate in (5) to obtain col. (10) for the weighted measure

Note that, in order to facilitate comparison, these weights have been chosen such that the weighted poverty rate p at the national level is numerically the same as the conventional rate $p_2$ with poverty line at 60% of the national median.

## 5.3 Illustration

Table 3 illustrates the application of the procedure to poverty rates at the regional level, all computed with reference to the national lines but at three different thresholds of the median – 50%, 60% and 70%. As before, we consider NUTS1 regions in Poland and NUTS2 regions in the Czech Republic. Cross-sectional EU-SILC data from these countries have been used for the illustration.

For simplicity, all variances in Table 3 have been computed using the above-mentioned expressions, which take the sample as SRS and approximate the complex statistic ‘poverty rate’ as an ordinary proportion. In the table, “A” refers to the consolidated poverty rate, taken as a weighted average of the rates based on different poverty line thresholds. The weights are taken according to equation (16), applied at the national level in each country, and these weights are then assumed as constants in evaluating the variances.

For the weighted measure, variance is lowered by 12% in the case of Poland and 16% in the Czech Republic. This is the estimated gain at the national level, and these are also the average of regional gains in each of the two countries.[8]

[8] Note that in these estimates the weights in equation (16) taken as constants. This can be approximately the case if the weights come from a large ‘external’ source, such as from pooled data over EU or other grouping of countries. If they are taken simply from the figures at the national level, then by definition there can be no gain in precision at that level since $p=p_2$ always, by definition. Nevertheless, the gain at the provincial level may still be realistically estimated with such as assumption, especially when dealing with lower level regions.

# 6. Averaging over survey waves to improve precision of regional indicators

## 6.1 Cumulating estimates or data over waves

Consider that for each wave in a panel, a person's poverty status (poor or non-poor) is determined based on the income distribution separately of that wave. Then the results from several waves are pooled together (each individual appearing in the pooled data set as many times as he/she appears in the sample during those waves). The proportion poor among the 'cases' in the pooled data gives the poverty rate averaged over waves. This gives to each wave a weight in the cumulation proportional to its sample size.

Alternatively, we may choose to give the same importance to the results from each wave. Actually this is preferable from a substantive point of view, though from the point of sampling error the first option is likely to be a little more efficient. In any case, in practice the difference between the two modes of constructing the average may be minor in so far as wave sample sizes are similar.

The main issue is to determine the gain in sampling precision from such pooling. With a panel design, the statistical problem is the following. A large proportion of the individuals are common in the different panels. However, a certain proportion of individuals are different from one wave to another. The cross-sectional samples are not independent, resulting in correlation between measures from different waves. Apart from correlations at the individual level, we have to deal also with the additional correlation that arises because of the common structure (stratification and clustering) of the waves of a panel. Such correlation would exist, for instance, in samples coming from the same clusters even if there is no overlap in terms of individual households. For this purpose, the JRR approach can be extended on the following lines for estimating variance of estimates cumulated over time.

Consider the second of the above procedures for constructing the average poverty rate (i.e., computing the poverty rate for each wave separately, and then taking an unweighted average of these rates). Using the common sample structure of the cross-sections in a panel, a 'common set of JRR replications' (see below) is defined. For each replication, the required measure is constructed for each cross-section involved. These replication-specific cross-sectional measures are aggregated to obtain the required average measures for the replication. Variance is then estimated from the resulting replicated measures in the usual way.

For constructing a ‘common set of replications’ the total sample of interest is formed by the union of all the cross-sectional samples being aggregated. Using as basis the common structure of this total sample, a common set of JRR replications is defined for it in the usual way. Constructing a ‘common set of replications’ requires that when an element is to be excluded in the construction of a particular replication, it must be excluded simultaneously from every cross-sectional sample included in which the element appears. As noted, for each replication the required measure is then constructed for each of the cross-sectional samples involved.

## 6.2 Indication of the gain in precision from cumulation

The following provides a simplified procedure for quantifying the gain in precision from averaging over waves of the EU-SILC panel. As noted above, more accurate variance estimates for the cumulated estimates can be made using, for instance, the JRR methodology. But the following illuminates the statistical mechanism of how the gain is achieved.

In assessing the reduction in standard error because of consolidation of measures over T waves, of course we cannot merely add up the sample sizes over the waves. EU-SILC is a (rotational) 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 in the poverty status between the two waves form the same panel:

| | 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'=ab+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 the following for the population variances:

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

Similarly,

$$\mathrm{var}(p'_j) = p'.(1-p') = v' \,.$$

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

Parameter ‘a’ is the persistent poverty rate over the two years.

For simplicity of exposition, let us consider the simple case in which the difference between p and p’ can be ignored, and the two waves are a part of a single panel, so that there is complete overlap and any differences between the waves in sample sizes do not complicate the picture. With these simplifications, for data averaged over two adjacent years, variance $v_A$ is given by:

$$v_A = \frac{v}{2}.\left(1+\frac{c_1}{v}\right). \tag{17}$$

The correlation

$$b = \left(\frac{c_1}{v}\right) = \left(\frac{a-p^2}{p-p^2}\right) \tag{18}$$

between two adjacent waves is expected to decline as the two become more widely separated. Let $(c_k/v)$ be the correlation between two points *k* waves apart. A simple and reasonable model of the attenuation of correlation with increasing *k* is:

$$\left(\frac{c_k}{v}\right) = \left(\frac{c_1}{v}\right)^k.$$

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

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

with $(c_1/v)$ given by (18), where quantity $a$ is the overall rate of persistent poverty between pairs of adjacent waves (averaged over K-1 pairs), and p is the (cross-sectional) poverty rate averaged over K waves.

Returning to (17) for application to pairs of waves in EU-SILC, it is necessary to allow for variations in cross-sectional sample sizes and for the fact that overlap between cross-sections is partial in the EU-SILC rotational panel design. The result can be seen to be:

$$v_A = \frac{(v_1 + v_2)}{4} . \left( 1 + b. \left( \frac{n}{n_H} \right) \right) \tag{19}$$

where $v_1$ and $v_2$ are the sampling variances for single waves 1 and 2,

b is the correlation coefficient for a statistic (such as mean equivalised income or a poverty rate) over the two cross-sections:

$$b = \left( \frac{a - p^2}{p - p^2} \right), \quad p = \frac{p_{1(n)} + p_{2(n)}}{2},$$

and n is the overlap between the cross-sectional samples, and $n_H$ is the harmonic mean of their sample sizes $n_1$ and $n_2$:

$$n_H = \left( \frac{2.n_1.n_2}{n_1 + n_2} \right). \tag{20}$$

## 6.3 Illustration

The methodology described above has been applied to the 2005-2006 EU-SILC samples for Poland and the Czech Republic at the national level. Results on sampling errors for the full cross-sectional and the longitudinal samples for 2006 come from Tables 1 and 2. Corresponding results for 2005 have been computed exactly in the same way, except that for the longitudinal sample results for both 2006 and 2005, the appropriate weights to be used are the base weight for *2006* in the longitudinal data set. This is because the 2005 data in the 2006 longitudinal data set are from the sample already subject to the additional attrition between 2005 and 2006, which is reflected in the 2006 but not in the 2005 sample weights.

In terms of the quantities defined in the previous subsection, rows (1)-(5) in Table 4 are as follows.

Standard error of average HCR over two years (assuming independent samples):

$$(1) \quad = \quad \frac{1}{2}.(v_1 + v_2)^{1/2} \quad .$$

Increase in standard error due to positive correlation between waves:

$$(2) \quad = \quad \left(1 + b.\left(\frac{n}{n_H}\right)\right)^{1/2} \quad .$$

Standard error of average HCR over two years (given correlated samples):

$$(3) \quad = \quad (1).(2) \quad = \quad (v_A)^{1/2} \, .$$

Average standard error over a single year:

$$(4) \quad = \quad \frac{(v_1)^{1/2} + (v_2)^{1/2}}{2} \, .$$

Average gain in precision (variance reduction, or increase in effective sample size, compared to a single year sample):

$$(5) \quad = \quad 1 - \left(\frac{(3)}{(4)}\right)^2 \, .$$

Averaging over two waves, variance of at-risk-of-poverty rate is found to be 30% less than the variance of the measure estimated from a single wave. This figure comes out to be essentially the same in Poland and the Czech Republic.

**Table 4: Gain in precision from averaging over correlated samples in EU-SILC rotational design**

| | Estimate | Sample size | | Components of design effect | | | %se* | %se* |
|---|---|---|---|---|---|---|---|---|
| | | persons | households | dH | dW | dX | rand | actual |
| | | | | | | (a) | (b) | (c)=<br>(a)*(b) |
| **POLAND** | | | | | | | | |
| **Full cross-sectional sample (2006)** | | | | | | | | |
| HCR* 2006 | 19.1 | 45 122 | 14 914 | 1.74 | 1.09 | 1.02 | 0.50 | 0.51 |
| **Full cross-sectional sample (2005)** | | | | | | | | |
| HCR* 2005 | 20.6 | 49 044 | 16 263 | 1.74 | 1.07 | 1.02 | 0.44 | 0.45 |
| **Longitudinal (panel) sample 2005-06** | | | | | | | | |
| HCR* 2006 | 18.5 | 32 820 | | | | | | |
| HCR* 2005 | 20.2 | 32 820 | | | | | | |
| Persistent poverty rate<br>(2005-06) | 12.5 | 32 820 | | | | | | |
| (1) Standard error of average HCR over two years (assuming independent samples) | | | | | | | | 0.34 |
| (2) Increase in standard error due to positive correlation between waves | | | | | | | | 1.18 |
| (3)=(1)+(2) Standard error of average HCR over two years (given correlated samples) | | | | | | | | 0.40 |
| (4) Average standard error over a single year sample | | | | | | | | 0.48 |
| (5)=1-[(3)/(4 Average gain (variance reduction, or increase in effective sample size, over a single year sample) | | | | | | | | 30% |
| **CZECH REPUBLIC** | | | | | | | | |
| **Full cross-sectional sample (2006)** | | | | | | | | |
| HCR* 2006 | 9.8 | 17 830 | 7 483 | 1.54 | 1.23 | 1.00 | 0.57 | 0.57 |
| **Full cross-sectional sample (2005)** | | | | | | | | |
| HCR* 2005 | 10.4 | 10 333 | 4 351 | 1.54 | 1.25 | 1.00 | 0.65 | 0.65 |
| **Longitudinal (panel) sample 2005-06** | | | | | | | | |
| HCR* 2006 | 10.3 | 9 101 | | | | | | |
| HCR* 2005 | 10.2 | 9 101 | | | | | | |
| Persistent poverty rate<br>(2005-06) | 6.4 | 9 101 | | | | | | |
| (1) Standard error of average HCR over two years (assuming independent samples) | | | | | | | | 0.43 |
| (2) Increase in standard error due to positive correlation between waves | | | | | | | | 1.18 |
| (3)=(1)+(2) Standard error of average HCR over two years (given correlated samples) | | | | | | | | 0.51 |
| (4) Average standard error over a single year | | | | | | | | 0.61 |
| (5)=1-[(3)/(4 Average gain (variance reduction, or increase in effective sample size, over a single year sample) | | | | | | | | 30% |

HCR* = at-risk-of-poverty rate, using national poverty line (60% of national median)
%se* = standard error in absolute percentage points

NB: For cross-sectional and longitudinal results for 2006, see Tables 1 and 2. For formulae used for rows (1)-(5), see text.

# 7. Small area estimates for regional indicators

## 7.1 Outline of a small area estimation procedure for regional indicators

For the estimation of measures at regional level via small area estimation techniques, we believe that a good procedure is to use the Empirical Best Linear Unbiased Predictor (EBLUP) estimator, with appropriate procedures to evaluate the robustness of such measures.[9]

In this methodology an intensive survey with a modest sample size, such as EU-SILC, provides direct poverty-related information at the micro (unit) level. This information can be aggregated to areas such as NUTS regions where the survey contains some sample units from the area and the area identifies are available in the microdata. On the other side, correlates of poverty-related characteristics of the areas can come from aggregated statistics (such as NewCronos). The two sources can be combined to produce composite estimates, provided that (i) the survey data contain information for the identification of the area to which each unit belongs (which, unfortunately, is not always the case in EU-SILC data files), and (ii) the aggregate data on the correlates are available for all the areas in the population of interest (which fortunately is the case for many correlates available in data sources such as NewCronos).

The approach can be to apply area level random-effect models relating small area direct estimates (from EU-SILC) to domain-specific covariates, considering the random area effects as independent. The basic area-level model includes random area specific effects, and in it the area specific covariates, $x_i = (x_{i,1}, x_{i,2}, \dots x_{i,p})$, $i = 1 \dots m$, are related to the target parameters $\theta_i$ (totals, means, proportion, etc.) as:

$$\theta_i = x_i \beta + z_i v_i$$

[9] In the literature small area models are classified as: (i) area-level random effect models, which are used when auxiliary information is available only at area level (such as the prevailing unemployment rate in the area); (ii) nested-error unit level regression models, used if unit specific covariates (such as the individual's or the household's employment situation) are available at unit level.
On the basis of empirical work, it appears that area-level synthetic estimates tend to produce better results than their unit-level counterparts. This is because regression coefficients calculated at unit-level do not always correctly reflect the relationship between the area-level averages involved in the synthetic estimator. In any case, the type of data available for poverty analysis at the regional level generally precludes the use of unit (household or person) level models.

where $z_i$ are known positive constants, $\beta$ is the regression parameters vector (px1), and $v_i$ are independent and identically distributed random variables with 0 mean and variance $\sigma_v^2$. The model assumes that the direct estimators $\hat{\theta}_i$ are available and design unbiased, in the form:

$$\hat{\theta}_i = \theta_i + e_i$$

where $e_i$ are independent sampling errors with zero mean and known variance $\psi_i$. The BLUP estimator is a weighted average of the design-based estimator and the regression synthetic estimator:

$$\tilde{\theta}_i\left(\sigma_v^2\right) = \gamma_i\hat{\theta}_i + \left(1-\gamma_i\right)x_i\hat{\beta}\ , \qquad (21)$$

where

$$\gamma_i = \frac{\sigma_v^2}{\sigma_v^2 + \psi_i} \qquad (22)$$

is a weight (or ‘shrinkage factor’) which assumes values in the range [0-1]. This parameter measures the uncertainty $\sigma_v^2$ in modelling $\theta_i$ in relation to the total uncertainty including the variance of the direct estimator $\varphi_i$ (Gosh and Rao, 1994). The mean square error of the BLUP estimator depends on the variance parameter $\sigma_v^2$, which in practice is replaced by its estimator; hence the estimator obtained is called Empirical BLUP (EBLUP).

Mathematical details for the EBLUP estimators are available in Handerson (1950); see also Ferretti (2010), where it is proven that EBLUP is a special case of the wider Empirical Best (EB) estimator. It is a consistent estimator when the number of statistical units is sufficiently high.

## 7.2 External data source for SAE using EU-SILC

EU-SILC and EU level sources such as NewCronos provide common and presumably reasonably comparable information across countries. A very important consequence of this favourable situation is that in making small area estimations, information may be pooled over countries. This implies the assumption of similar relationships between the model variables in different countries. This is a strong assumption. Its justification results from the 'ratio approach' outlined below. A major positive feature of this approach is that the modelling strategy becomes hierarchical. We begin with poverty rates and other target variables at the national level, using essentially direct survey estimates without involving any modelling. Then we move to NUTS1 level and obtain estimates using the SAE methodology. From there we move to the next level down, if possible and useful, and so on.

## 7.3 The ratio approach in constructing SAEs

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 region i and its country. In this way the effect of the difficult-to-quantify institutional and historical factors, common to the country and its regions, is abstracted. This makes the pooling of data across different countries for the estimation of a common model more reasonable. Similarly, in going from NUTS1 region i to its NUTS2 region j, we express the model variables in the form:

$$R_{ij} = Y_{ij} / Y_i$$

and similarly from NUTS2 regions j to its NUTS3 region k in the form:

$$R_{ijk} = Y_{ijk} / Y_{ij} \, . \qquad (23)$$

This type of modelling is further improved by taking different parts of a large or exceptionally heterogeneous country as separate units, 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. We may simply model the ratio of the subpopulation measure to the total population measure. This methodology is developed in Verma *et al* (2005).

## 7.4 Estimation of standard errors in the ratio approach

As noted, 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 $\text{var}(r)$ for the ratio, given $\text{var}(e_1)$ and $\text{var}(e_2)$ for the absolute quantities. We have:

$$\text{var}(r) = \text{var}\left(\frac{e_2}{e_1}\right) = \frac{1}{e_1^2}.\left(\text{var}(e_2) + r^2.\text{var}(e_2) - 2.r.\text{cov}(e_1, e_2)\right) \quad . \qquad (24)$$

The covariance is 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 $n_1$:

$$\text{cov}(e_1, e_2) = \left[\text{var}(e_1).\text{var}(e_2).(n_2/n_1)\right]^{½} \quad . \qquad (25)$$

## 7.5 Illustration: SAE for regional indicators in Poland and the Czech Republic

Below we present some empirical results of EBLUP estimates based on EU-SILC data for NUTS1 regions in Poland and NUTS2 regions in the Czech Republic. We have chosen these two countries and the above-noted levels of regional disaggregation in them only because of the limitations in the data available on NUTS classification in EU-SILC. We have decided to pool together the eight NUTS2 regions in the Czech Republic and the six NUTS1 regions in Poland in order to reach a sufficient number of observations for efficiently estimating the regression models of the EBLUP estimator.

*Covariates used*

The regressors used in the model have been downloaded from NewCronos and are reported in the table below.

As explained above, we used the ‘ratio approach’ to improve the precision of the models.

Under this approach, the model input consists of:

(1) NUTS1-to-Country (Poland) or the NUTS2-to-Country (the Czech Republic) ratio for the statistic concerned, as directly estimated from the survey;

(2) standard error of this ratio estimate.

The output from the model consists of:

(1) model estimate of NUTS1-to-Country ratio (Poland) or NUTS2-to-Country ratio (the Czech Republic) for the statistic concerned;

(2) mean-squared error of this estimate.

**Covariates available at Nuts1 (PL) and Nuts2 (CZ) levels**

| | | |
|---|---|---|
| 1 | Disposable income | PPS per capita 2006, net |
| 2 | GDP | PPS per capita 2006 |
| 3 | Activity rate | Activity rate for 2006; from domain Regional Labour Market |
| 4 | Unemployment rate | Unemployment rate 2006; from Regional Unemployment: LFS adjusted series |
| 5 | Long-term unemployment | Long-term unemployment rate 2006 (unemployed for 12 months or longer); from Regional Unemployment: LFS adjusted series |
| 6 | Population density | Population density 2006; constructed from demographic statistics |
| 7 | IMR | Infant mortality rate 2006; constructed from demographic statistics |
| 8 | HH Size | Household size 2006; estimated from EU-SILC 2006 |

## Performance measures

Table 5 shows some performance measures of the two SAE Models: the first (panel B in the table) refers to the model relating the *absolute* NUTS values (i.e., not the region-to-country ratios), while the second (panel D in the table) refers to the ratio model described above. For each model three measures of interest are shown:

- 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;
- the ratio between the EBLUP estimated value and the corresponding direct estimate. This is to check the extent to which the modelling changes the input direct estimates;
- the ratio between mean square error (MSE) of the EBLUP estimate of the NUTS: Country ratio, and the MSE of direct survey estimate of this ratio (which in this case is simply its variance, since the estimates are unbiased). This is to check the extent to which the modelling has improved precision of the estimates.

Note that in the first model (panel B), the weight given to direct estimate (gamma) is lower for the Czech Republic compared to Poland. This results from the regional (NUTS2) sample sizes in the Czech Republic being smaller than the regional (NUTS1) sample sizes in Poland. The final estimates are on the whole similar to the original ones. Notable gain in terms of MSE is present in the Czech Republic (0.74), but such gain is not found in the case of Poland.

When analyzing the model in terms of ratios (panel D), the weight given to the regression model is increased in the case of the Czech Republic and decreased in the case of Poland. This is reflected also in the final gain in terms of MSE, which is evident in the Czech Republic (almost 50%). In this case the model does not perform well in Poland, where the estimates have a certain bias, which is not sufficiently compensated for by reduction in terms of variance since the direct estimates already have a very small standard error.

## Table 5: Small area (EBLUP) estimates of at-risk-of-poverty rates for sub-national regions: illustrations

| Czech Republic NUTS2 regions | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Panel | | (A) | | | | (B) | | | (C) | | | | | (D) | | |
| | n | est | se | gamma | est | se | ratio_est | ratio_MSE | r | (r)se | gamma | (r)est | (r)se | ratio_est | ratio_MSE | estimate |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7)=<br>(5)/(2) | (8)=<br>(6)/(3) | (9)=<br>(2)/(2-c'try) | (10) | (11) | (12) | (13) | (14)=<br>(12)/(9) | (15)=<br>[(13)/(10)]^2 | (16)=<br>(12)*(2-c'try) |
| CZ01 - Praha | 1456 | 4.7 | 1.32 | 0.30 | 4.6 | 1.47 | 0.97 | 1.24 | 0.48 | 0.130 | 0.21 | 0.41 | 0.108 | 0.85 | 0.69 | 4.0 |
| CZ02 - Strední Cechy | 1736 | 8.5 | 1.81 | 0.18 | 7.3 | 1.22 | 0.87 | 0.45 | 0.86 | 0.176 | 0.13 | 0.69 | 0.095 | 0.80 | 0.29 | 6.8 |
| CZ03 - Jihozápad | 2142 | 6.1 | 1.29 | 0.30 | 7.5 | 1.21 | 1.23 | 0.88 | 0.62 | 0.124 | 0.23 | 0.72 | 0.094 | 1.17 | 0.57 | 7.1 |
| CZ04 - Severozápad | 1996 | 16.0 | 1.93 | 0.16 | 16.2 | 1.57 | 1.01 | 0.67 | 1.63 | 0.188 | 0.12 | 1.66 | 0.144 | 1.02 | 0.59 | 16.3 |
| CZ05 - Severovýchod | 2632 | 8.9 | 1.22 | 0.33 | 8.8 | 1.18 | 0.99 | 0.95 | 0.91 | 0.115 | 0.26 | 0.87 | 0.091 | 0.96 | 0.63 | 8.6 |
| CZ06 - Jihovýchod | 2888 | 8.3 | 1.30 | 0.30 | 8.3 | 1.20 | 1.00 | 0.84 | 0.85 | 0.122 | 0.24 | 0.88 | 0.092 | 1.03 | 0.57 | 8.6 |
| CZ07 - Strední Morava | 2408 | 11.2 | 1.81 | 0.18 | 9.8 | 1.20 | 0.88 | 0.44 | 1.14 | 0.171 | 0.14 | 1.00 | 0.092 | 0.88 | 0.29 | 9.8 |
| CZ08 - Moravskoslezsko | 2572 | 15.5 | 2.01 | 0.15 | 12.8 | 1.30 | 0.83 | 0.42 | 1.58 | 0.190 | 0.11 | 1.39 | 0.120 | 0.88 | 0.40 | 13.6 |
| CZ - Czech Republic | 17830 | 9.8 | 0.57 | 0.24 | | | 0.97 | 0.74 | 1.00 | | 0.18 | | | 0.95 | 0.50 | |

| Poland NUTS1 regions | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | n | est | se | gamma | est | se | ratio_est | ratio_MSE | r | (r)se | gamma | (r)est | (r)se | ratio_est | ratio_MSE | estimate |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7)=<br>(5)/(2) | (8)=<br>(6)/(3) | (9)=<br>(2)/(2-c'try) | (10) | (11) | (12) | (13) | (14)=<br>(12)/(9) | (15)=<br>[(13)/(10)]^2 | (16)=<br>(12)*(2-c'try) |
| PL1 - Centralny | 8728 | 17.1 | 1.06 | 0.40 | 17.4 | 1.16 | 1.01 | 1.21 | 0.90 | 0.050 | 0.65 | 0.90 | 0.060 | 1.00 | 1.44 | 17.2 |
| PL2 - Poludniowy | 9273 | 14.7 | 0.97 | 0.44 | 15.5 | 1.10 | 1.06 | 1.29 | 0.77 | 0.045 | 0.69 | 0.79 | 0.053 | 1.03 | 1.35 | 15.2 |
| PL3 - Wschodni | 9079 | 25.2 | 1.34 | 0.29 | 24.5 | 1.48 | 0.97 | 1.20 | 1.32 | 0.063 | 0.54 | 1.31 | 0.076 | 0.99 | 1.45 | 25.0 |
| PL4 - Pólnocno-Zachodni | 6912 | 18.7 | 1.32 | 0.30 | 19.3 | 1.23 | 1.04 | 0.87 | 0.98 | 0.063 | 0.53 | 0.98 | 0.069 | 1.01 | 1.18 | 18.8 |
| PL5 - Poludniowo-Zachodni | 4538 | 18.6 | 1.56 | 0.23 | 19.3 | 1.30 | 1.04 | 0.70 | 0.98 | 0.077 | 0.43 | 1.05 | 0.074 | 1.08 | 0.92 | 20.0 |
| PL6 - Pólnocny | 6592 | 21.4 | 1.40 | 0.27 | 19.9 | 1.26 | 0.93 | 0.81 | 1.12 | 0.068 | 0.50 | 1.07 | 0.072 | 0.95 | 1.14 | 20.4 |
| PL - Poland | 45122 | 19.1 | 0.51 | 0.32 | | | 1.01 | 1.02 | 1.00 | | 0.56 | | | 1.01 | 1.25 | |

Panels

(A) Direct estimates and JRR direct standard errors

(B) EBLUP estimates with simple model (in terms of absolute values, rather than ratios)

(C) Direct estimates of ratios and standard errors

(D) EBLUP estimates with ratio model

In conclusion, EBLUP model has proved very useful in the Czech Republic, where geographical information is available at NUTS2 level, and where the regional sample sizes are not large. In Poland the gain of using EBLUP estimators is not realised since the direct estimates at NUTS1 level are already quite good; moreover, in the case of the ratio model, the final estimates are also biased.

Indeed, EBLUP models can be more useful when they are applied to lower level regions, as demonstrated using ECHP data for Italy at NUTS3 level in Verma *et al* (2005).

# 8. Concluding remarks

## 8.1 The potential of sources such as NewCronos to supplement regional indicators

Among the diverse external sources which may be used to supplement EU-SILC for the construction of regional indicators of poverty and social exclusion, almost certainly NewCronos (now termed 'Eurostat Free Dissemination Database') is a front runner. It provides a valuable data resource for the construction of regional indicators. In itself NewCronos is not a source of original data, but a compilation of information from a diversity of sources presented in the form of very detailed tabulations. NewCronos REGIO domain covers the principal aspects of the economic and social life of the European Union: demography, economic accounts, labour force, health, education, etc., by region. The concepts and definitions used are as close as possible to those used by Eurostat for the production and compilation of statistics at national level.

NewCronos is useful not simply as a source of covariates for small area estimation using EU-SILC, as may seem to have been assumed in the previous section. In fact, there are at least three different forms in which we can make use of variables derived from a source like NewCronos for the construction of regional indicators.

> (1) Some statistics in NewCronos can serve, in their own right, as direct indicators pertaining to poverty and living conditions. In fact, the scope for such use is likely to be greater in the context of regional indicators, compared to that in the national context. This is because measures of levels - which are more abundantly available in NewCronos than the generally more complex distributional measures - can themselves serve as indicators of disparity when compared across regions.
>
> (2) A large number of measures correlated with direct indicators of poverty and deprivation can be constructed. In conjunction with direct indicators obtained from more intensive surveys, these measures can be used as 'covariates' or 'regressors' to produce more precise indicators using small area estimation procedures, as described in the previous section.

(3). In addition, NewCronos provides a very large number of measures, giving what has been termed as "intermediate output" indicators. Such indicators express on the one hand the policy effort in favour of those at risk of poverty and social exclusion, and on the other hand the impact of social policies as well as of the economic context. NewCronos is a unique source of such indicators.

NewCronos has hitherto been under-utilised for these purposes, and there is a great potential for more thorough exploitation of the information which already exists in this source. While direct indicators of regional poverty and living conditions are generally not available with sufficient regional breakdown in NewCronos, several exceptionally positive aspects of the resource need to be appreciated; some of these become even more important as we move down from the national to the regional level. Firstly, a wide range of subject-matter areas are covered in the very detailed tabulations provided. These can be utilised to construct many direct indicators pertaining to poverty and living conditions, as well as to obtain many more variables correlated with direct indicators. Secondly, detailed break-downs are available for many variables correlated with indicators of poverty and deprivation, mostly to NUTS2, and in a few cases to NUTS3 level.

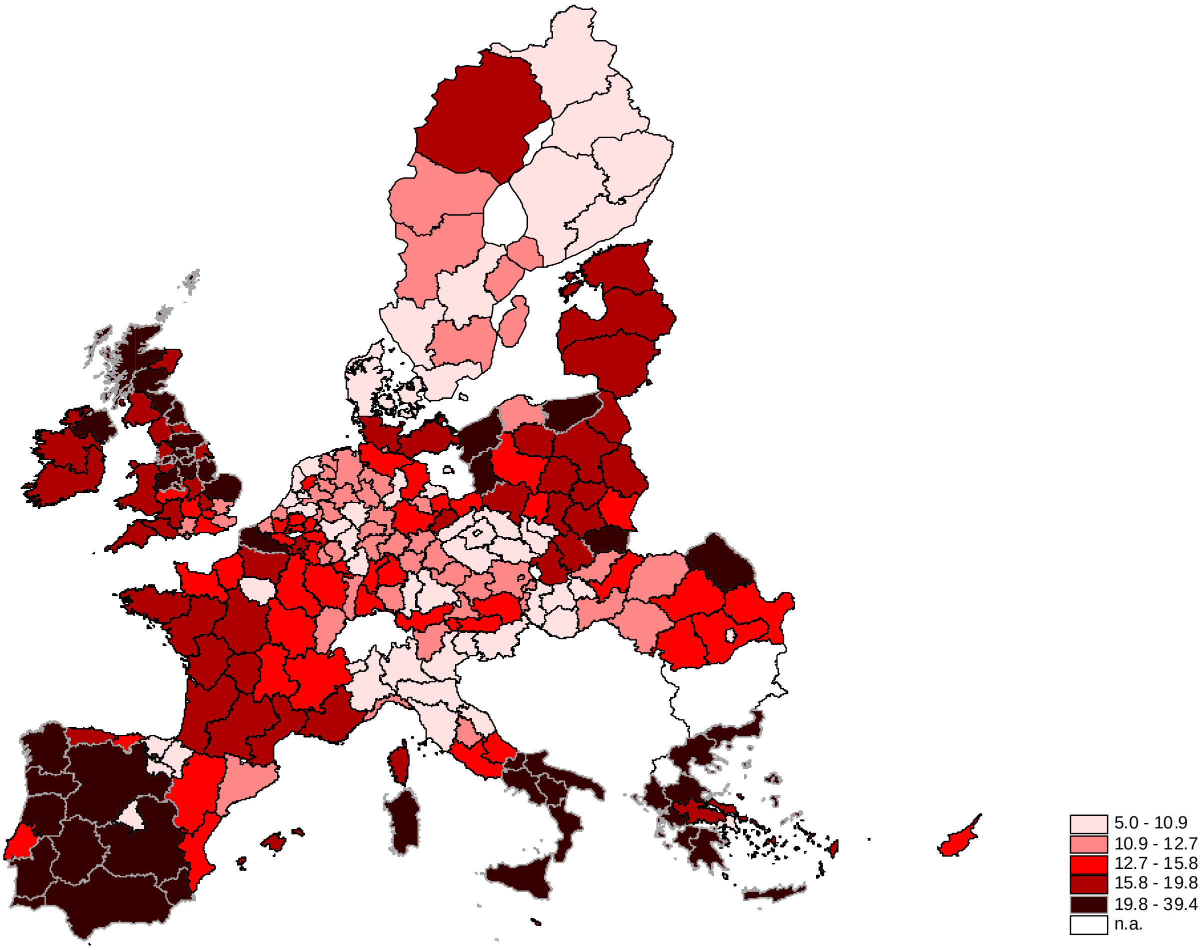
## 8.2 The potential with more complete information on sample structure and regional identifiers

It is clear from the research presented in this Working Paper that the production of indicators of poverty and social exclusion from EU-SILC at the level of sub-national regions is severely limited because of the lack of information in the microdata on sample structure and for the identification of regions.

It would be very useful if steps are taken to improve data availability in this respects. In fact, the situation was better in relation to data of the European Community Household Panel which preceded EU-SILC. Just as an illustration of what could be done with data from ECHP supplemented by NewCronos and some other external sources, below we reproduce some results from an earlier application by the authors and colleagues of an extended application of the small area estimation methodology in order to produce at-risk-of-poverty rates for NUT2 regions in the EU.

The figure below shows the at-risk-of-poverty rates for NUTS2 regions, but based on the national poverty line for regions in any given country.

**Figure 1: NUTS2 at-risk-of-poverty rates (country poverty lines)**



Source: Verma *et al* (2005). ECHP data, supplemented from NewCronos and other sources.

# References

Betti, G. and Verma, V. (2008), ‘Fuzzy measures of the incidence of relative poverty and deprivation: a multi-dimensional perspective’, *Statistical Methods and Applications*, **12**(2), pp. 225-250.

Ferretti, C. (2010), ‘Multidimensional poverty measures: issues in small area estimation’, PhD Thesis, University of Florence.

Giorgi, L. and Verma, V. (2002), ‘European Social Statistics: Income, Poverty and Social Exclusion: 2nd Report’ Luxembourg: Office for Official Publications of the European Communities.

Gosh, M. and Rao, J.N.K (1994), ‘Small Area Estimation: An Appraisal (with discussion)’, *Statistical Science*, 9 (1), pp. 55-93.

Handerson, C.R. (1950), ‘Estimation of Genetic Parameters’, *Annals of Mathematical Statistics*, 21, pp. 309-310.

Kish, L. (1965), Survey Sampling. Wiley.

Kish, L. (1995), ‘Methods for design effects’, Journal of Official Statistics, 11: 55-77.

Rust, K. (1985), Efficient replicated variance estimation. University of Michigan and Australian Bureau of Statistics.

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

Verma, V. and Betti, G. (forthcoming), ‘Taylor linearization sampling errors and design effects for poverty measures and other complex statistics’, *Journal of Applied Statistics.*

Verma, V. , Betti, G. and Gagliardi, F. (2010), ‘An assessment of survey errors in EU-SILC’, *Eurostat methodologies and working papers*, Eurostat, Luxembourg.

Verma, V., Betti, G., Lemmi, A., Mulas, A., Natilli, M., Neri, L. and Salvati, N. (2005), ‘Regional indicators to reflect social exclusion and poverty. Final report.’ Project VT/2003/45, European Commission, Employment and Social Affairs D.G.

Verma, V., Betti, G., Natilli, M. and Lemmi, A. (2006), ‘Indicators of social exclusion and poverty in Europe’s regions’, Working Paper no. 59, Department of Quantitative Methods, University of Siena.

European Commission

**Robustness of some EU-SILC based indicators at regional level**

Luxembourg: Publications Office of the European Union

2010 — 58 pp. — 21 x 29.7 cm

ISBN 978-92-79-16756-0
ISSN 1977-0375
Doi:10.2785/5605
Cat. No KS-RA-10-020-EN-N

**Theme: Population and social conditions**

**Collection: Methodologies and working papers**

## How to obtain EU publications

**Free publications:**

- via EU Bookshop (http://bookshop.europa.eu);
- at the European Union's representations or delegations. You can obtain their contact details on the Internet (http://ec.europa.eu) or by sending a fax to +352 2929-42758.

**Priced publications:**

- via EU Bookshop (http://bookshop.europa.eu).

**Priced subscriptions (e.g. annual series of the *Official Journal of the European Union* and reports of cases before the Court of Justice of the European Union):**

- via one of the sales agents of the Publications Office of the European Union (http://publications.europa.eu/others/agents/index_en.htm).

KS-RA-10-020-EN-N

ISBN 978-92-79-16756-0

9 789279 167560

Publications Office