QdS, vol. 16, 2014

SAS routines for variance estimation of poverty measures based on sample cumulated over waves of a panel

Francesca Gagliardi
Department of Economics and Statistics, University of Siena
E-mail: gagliardi10@unisi.it

Summary: The main issue of the present work is to pool data with objective to enhance the sample size, in particular with reference to subnational (regional) estimates for which sample sizes are usually too small. This introduces an additional issue of dealing with correlations between samples from consecutive waves of a rotational panel such as EU-SILC survey. The poverty status (and related indicators) of an individual are defined independently for each cross-sectional sample; hence the direct way to estimate variance of a cumulative sample is to pool the cross-sectional samples and apply a suitably adapted standard variance estimation procedure. Such direct estimation requires full information on the sample structure. At a minimum this includes specification at the micro level of: sample weights, stratum, primary sampling unit (PSU), permitting linking of data across waves. The SAS routines presented in this work have been developed for application when full information on the sample structure is available. Our empirical application is based on micro-data for the survey of Spain for year 2009, 2010 and 2011, to which we have had a privileged access through a project with OECD. Unfortunately, EU-SILC micro data available to researcher generally lack full information on sample structure. In general, the variance estimation procedure would need adaptation (and some additional assumptions) to deal with the situation when full information on sample structure is lacking. While this work does not address alternative procedures for the purpose, we have developed and applied those in previous research. For completeness, the technical steps involved have been outlined in the concluding section.

Keywords: Poverty measures, JRR, Panel survey, SAS routines.

1. Introduction

Poverty and social exclusion indicators are an essential monitoring tool, most useful when comparable across countries.

However, the extent to which income inequality and poverty vary within countries across different regions is actually relevant for policy decisions and monitoring. Sub-national measures are scarce, given the complexity of producing indicators at the regional level from the available data and the methodological issues related to cross-countries comparability. Implementing informed policies often requires statistics disaggregated to lower levels than those which meet national needs. National estimates are particularly insufficient for monitoring poverty and social exclusion, as these fields require complex statistics that take into account geographical distribution.

Distributional statistics are necessarily based on intensive and relatively small-scale surveys of households and individuals.

Survey data can be used in different forms or manners to construct regional indicators.

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

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

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

4. Using survey data in conjunction with data from other (especially administrative) sources which are larger in size but less detailed in content than survey data in order to produce improved estimates using small area estimation (SAE) techniques.

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

In this paper poverty and inequality measures have been produced on the basis of the so-called cumulation method (Verma $et\ al.$ 2013).

The reference data for this purpose are based on EU Statistics on Income and Living Conditions (EU-SILC), which is the major source of comparative statistics on income and living conditions in Europe. EU-SILC covers data and data sources of various types: cross-sectional and longitudinal; household-level and person-level; on income and social conditions; and from registers and interview surveys depending on the country. A standard integrated design has been adopted by nearly all EU countries. It involves a rotational panel in which a new sample of households and persons is introduced each year to replace one quarter of the existing sample. Persons enumerated in each new sample are followed-up in the survey for four years. The design yields each year a cross-sectional sample, as well as longitudinal samples of various durations.

The quantification of efficiency gains from averaging across multiple years is not straightforward in surveys, such as EU-SILC, that are based on a rotational panel. We have developed and tested two different methods to produce variance estimates for three-year averaged indicators in EU-SILC. A first, direct approach defines a common structure of strata and PSUs for the three waves of the sample, and applies the standard Jackknife Repeated Replication (JRR) methodology to the union of the three cross-sectional samples. An alternative (indirect) method has been developed to approximate the correlation across the cross-sectional waves using information from the longitudinal data of EU-SILC, which enables linking individuals and households across years when this is not possible in the cross-sectional datasets (Piacentini, 2014).

The issue of this paper is to develop efficient SAS routines for cumulation and standard errors estimation using JRR methodology when full information on sample structure is available. Section 2 presents the variances of cumulated measures. Section 3 describes the methodology chosen to estimate such variances, namely the Jackknife Repeated Replication. Section 4 specifies some practical aspect in dealing with variance estimation. In Section 5 the gain in sampling precision from pooling over waves using EU-SILC survey is quantified. Section 6 describes the SAS routines developed for the presented methodologies. Section 7 presents some empirical results obtained using such routines. Finally Section 8 concludes.

2. Cumulative measures of poverty

Consider that for each wave of a survey like EU-SILC, a persons poverty status (poor or non-poor) is determined from his/her income within the income distribution of that wave, independently for each EU-SILC year, and then the proportion of poor at each wave is computed. These proportions are then averaged over a number of consecutive waves.

The issue is to quantify the gain in sampling precision from such pooling, compared to results based on a single wave.

The quantification of efficiency gains from averaging across multiple years is not straightforward in surveys, such as EU-SILC, that are based on rotational panel, given that data from different waves of a rotational panel are highly correlated.

With a panel design, a new sample of households and individuals is introduced each year to replace only a fraction of the existing sample (1/4 in most of the EU-SILC country surveys, see Figure 1). A large proportion of the individuals are common in the different cross-sections. However, a certain proportion of individuals are different from one wave to the other. The cross-sectional samples are thus not independent, resulting in correlation between measures from different waves.

Apart from correlations at the individual level, we have to deal also with additional correlation that arises because of the common structure (stratification and clustering) of the waves of a panel. Such correlation would exist in, for instance, samples coming from

Figure 1.

the same clusters even if there is no overlap in terms of individual households.

In order to quantify the gain in precision from averaging over waves of a rotational panel, we provide the following simplified procedure that could be of help in better clarifying the point. It illustrates the statistical mechanism of how the gain is achieved.

Indicating by p_j and p'_j the (1, 0) indicators of poverty of individual j over the two adjacent waves, we have the following result for the population variances:

$$var(p_j) = \sum (p_j - p)^2 = p(1-p) = V;$$

similarly,

$$var(p'_j) = p'(1-p') = V',$$

$$cov(p_j, p'_j) = \sum (p_j - p)(p'_j - p') = a - pp' = c_1\,,$$

where a is the persistent poverty rate over the two adjacent years.

Under the two waves model and in the extreme case of a completely full sample overlap and $p' = p$, the variance V_A of the average over two waves of the concerned poverty measure can be estimated as:

$$V_A = (V/2)(1+\rho), \tag{2.1}$$

where ρ represents the correlation between the two waves that in our simplified case can be quantified by

$$\rho = \frac{c_1}{V} = \frac{(a - p^2)}{(p - p^2)}$$

Alternatively, if the overlap between the two waves is only partial like in the EU-SILC survey, and cross-sectional variances are not necessarily equal, it is necessary to allow for variations in cross-sectional sample sizes and partial overlaps:

$$V_A = \frac{1}{2}(\frac{V_1 + V_2}{2})(1 + \rho(\frac{n}{n_H})), \tag{2.2}$$

where V_1 and V_2 are the variances in each of the two waves, n is the sample overlap, n_H is the harmonic mean of different waves sizes, ρ as above (Verma $et\ al.$ 2013).

3. A replication method for variance estimation

In the previous section we have presented simplified formulae for variances when we construct measures averaged over waves of a panel. Now the question is: how can these variances be computed? In this section we present the variance estimation methodology chosen in our work.

The Jackknife Repeated Replication (JRR) is one of the class of practical methods for variance estimation in complex samples based on measures of observed variability among replications of the full sample.

All replicated variance estimation procedures are based on comparisons among replications generated through repeated re-sampling of the same parent sample. Once the set of replications has been appropriately defined for any complex design, the same variance estimation algorithm can be applied to a statistic of any complexity.

The basic requirement is that the full sample is composed of a number of subsamples or replications, each with the same design and reflecting complexity of the full sample, enumerated using the same procedures. A replication differs from the full sample only in size. But its own size should be large enough for it to reflect the structure of the full sample, and for any estimate based on a single replication to be close to the corresponding estimate based on the full sample.

At the same time, the number of replications available should be large enough so that comparison among replications gives a stable estimate of the sampling variability in practice.

JRR provides a versatile and straightforward technique for variance estimation in situations like the ones we are concerned with.

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

Briefly, the standard JRR involves the following.

Let z be a full-sample estimate of any complexity. We use the subscript i to indicate a sample primary sampling unit (PSU) and h indicate its stratum; $a_h 2$ is the number of PSUs in stratum h. Let $z_{(hi)}$ 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 an appropriate factor g_h (see below). Let $z_{(h)}$ be the simple average of the $z_{(hi)}$ over the a_h sample units in h. The variance of z is then estimated as:

$$var(z) = \sum_h [(1 - f_h) g_h \sum_i (z_{(hi)} - z_h)^2], \tag{3.1}$$

$(1 - f_h)$ is the finite population correction and it is usually 1 for samples in typical social surveys.

While one may take factor g_h as

$$g_h = \frac{a_h}{a_h - 1}, \tag{3.2}$$

it is more appropriate to use

$$g_h = \frac{w_h}{w_h - w_{hi}}, \tag{3.3}$$

where $w_h = \sum_i w_{hi}$, with $w_{hi} = \sum_j w_{hij}$ as the sum of sample weights of ultimate units j in primary selection units i. This means that in each replication (hi), the weights for individual units are redefined and rescaled as follows:

- for unit j not in stratum h: $w'_{hij} = w_{hij}$;
- for unit j in stratum h but not in PSU i: $w'_{hij} = g_h w_{hij}$;
- unit j in stratum h and in PSU i: $w'_{hij} = 0$.

The second form for g_h retains the total weight of the included sample cases unchanged across the replications created, so as to have the same total as that for the full sample. With the sample weights scaled such that their sum is equal (or proportional) to some external more reliable population total, population aggregates from the sample can be estimated more efficiently, often with the same precision as proportions or means (Verma and Betti, 2011).

4. Practical aspects: specification of sample structure variables

Practical variance estimation methods need to make some basic assumptions about the sample design because of the type and kind of sample design employed in social surveys. These assumption are generally met or they can be reasonably approximated in most population-based surveys.

These assumption are the following.

a) The survey is based on a probability sample.

b) The sample size is large enough.

c) The sample structure meets (or has been redefined to meet) the following requirements:

   c1) The sample selection is independent between strata.

   c2) Two or more primary selections are drawn from each stratum.

   c3) These primary selections are drawn at random, independently and with replacement.

   c4) The number of primary selections is large enough (i.e. greater than 30) for valid use of the approximations involved in the variance estimation equations. This assumption is needed to ensure that the sampling distribution of the measures constructed from large enough samples tends towards normal probability distribution, even if the distribution of variables like income (that is the reference variable for the measures constructed in this work) is highly skewed.

   c5) The primary selections within the same stratum do not differ greatly in size, meaning in the number of ultimate units selected and in the sum of the sample weights.

To these assumptions a final one should be added that, differently form the previous ones, frequently is not met in practice. The assumption is:

d) essential information on sample structure is provided.

Sampling error computations need to take into account variations in the sampling design. This is done through the definition of the sample structure.

In order to apply the JRR technique (and any other resampling technique and also, for example, Linearization methodology) it is necessary to have full access to the variables that define the structure of the sample, namely the stratification and the primary sampling units.

For the type of sample designs involved in EU-SILC, and in the practical procedures for variance estimation used, generally all the necessary information about the sample structure can be provided in the form of two variables defined for each unit:

- the *computational stratum*, namely the explicit and implicit stratification and
- the *computational primary sampling unit (PSU)*, to which the unit belongs.

Considering the EU-SILC survey, normally the variable computational stratum is related (and sometimes identical) to UDB variable DB050; similarly for computational PSU and DB060. However, very often the UDB variables require some redefinition before they can be used for the purpose of variance estimation.

In order to correctly define the computational strata and PSUs, information concerning the following three aspects must be available:

(1) Codes of the sample structure in the micro-data files.

(2) Detailed description of the sample design, for instance identifying features such as the presence of self-representing units, systematic selection etc.

(3) Information connecting the sample structure codes in the micro-data with descriptions of the particular sample design features, so as to be able to identify the design features applicable to particular units.

For EU-SILC, currently this information is not readily available at the central level for all countries. Presumably (and hopefully) it is available within each country for its own national survey.

As noted, 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 variance estimation. To do this in a statistically valid way requires sampling expertise.

The computational structure can differ from the actual sample structure because of various considerations such as the following.

Firstly, it is often necessary to define computational strata and PSUs to meet the basic requirement of practical methods of variance estimation for complex samples. Below we report some common situations.

(1) It may be necessary to regroup (collapse) strata so as to ensure that each stratum has at least two sample PSUs the *minimum number required* for the computation of variance.

(2) Units which are included into the sample automatically (self-representing units) are in fact strata rather than PSUs, and computational PSUs have to be defined at a lower stage within each such unit.

(3) In samples selected systematically, the implied implicit stratification is often used to define explicit strata, from each of which an independent sample is supposed to have been selected. Such strata have to be formed by pairing or otherwise grouping of PSUs in the order of their selection from the systematic list, ensuring that each resulting computational stratum has at least two primary selections.

(4) Sometimes non-response can result in the disappearance from the sample of whole PSUs. This can disturb the structure of the sample, such as leaving fewer than two PSUs in some strata. Variance computation requires some redefinition of the computational units to *meet the basic requirement of having at least 2 PSUs per stratum.*

(5) The above-mentioned problem arises more frequently and seriously when computing sampling errors for subclasses (subpopulations or small regions). The risk can be reduced by aggregating PSUs and strata to create fewer, larger computational units.

Considerations such as the above apply equally irrespective of whether the JRR or some other form of variance computation algorithm is used (Verma $et\ al.$ 2010).

(6) In a procedure like the JRR, the number of replications is equal or at least similar to the number of PSUs in the sample. In a large sample where elements (households, persons) have been selected directly, the number of replications which can be formed will be of the order of the sample size, normally running into thousands. This necessitates forming much fewer computational units, such as creating pseudo-cluster from random groupings of sample elements, and then random pairing of these clusters to construct computational strata.

(7) The above issue in fact arises in the case of any sample irrespective of its structure when we want to estimate not only variances but also design effects. The denominator of the design effect is variance under simple random sampling (SRS). That variance can be normally estimated by assuming the sample structure to be SRS.

(8) At a minimum, the replication approach requires re-computation of the statistic of interest at each replication. For complex statistics such as poverty rates, this may require a considerable amount of computer time, and it can be desirable to reduce the number of times the process has to be repeated. The same also applies to many other forms of complex analysis, such as estimation involving multivariate analysis and complex parameters, especially if they require iterative procedures.

(9) Variance estimation with replications captures the effect on variance of those features of the data treatment and estimation process used in the actual survey which are repeated for each replication, in the same way that they were applied to the full sample. For instance, in order to fully capture the effect of calibration on variance, it is necessary to recalibrate the sample of each replication using the same procedure as used in the actual sample. The same applies to other aspects of sample weighting, such as adjustment for non-response. Another even more demanding example is imputation for missing data. The need to repeat such heavy procedures at each replication can greatly increase the computational task. Means are required to reduce the number of replications involved.

(10) There are restrictions on the detail with which information identifying individual sampling units, PSUs, strata etc. can be included in the public-release micro data. *Grouping of units and strata can help in preserving confidential nature of the data*. Reducing the detail included in this manner would *make unnecessary the suppression of information on sample structure, such as the suppression done in the microdata disseminated by Eurostat*.

5. *Quantifying the gain in sampling precision from pooling over waves using EU-SILC survey*

The formulae presented in Section 2 and the methodology presented in Section 3 have been applied to the EU-SILC $cross-sectional$ datasets in order to get averaged measures over waves.

When complete information on sample structure is available and, specifically, when identifiers are provided to link strata and PSUs throughout different EU-SILC $cross-sectional$ datasets, it is possible to cumulate waves and quantify the gain in sampling precision achieved with this methodology.

When the above requirement is met, that is full information on sample structure is available, the gain in sampling precision can be easily quantified applying the standard JRR methodology presented in Section 3 on the basis of the following considerations.

The total sample of interest is formed by the union of all the cross-sectional samples being compared or aggregated.

Using as basis the common structure of this total sample, a set of JRR replications is defined in the usual way.

Each replication is formed such that when a unit is to be excluded in its construction, it is excluded simultaneously from every wave where the unit appears.

For each replication, the required measure is constructed for each of the cross-sectional samples involved, and these measures are used to obtain the required averaged measure for the replication.

Variance of the statistic of interest is then estimated from the replication estimates in the usual way.

Let us clarify this procedure, presenting an empirical example.

Consider that we have the cross-sectional dataset of the EU-SILC survey for three consecutive years and want to estimate the average of a given poverty measure over the three years. We proceed as follows.

We first construct a common structure of strata and PSUs from the union of the three cross-sectional datasets (see Figure 2); that is, we keep the list of all the strata and PSUs of each of the three datasets and construct a new list that is the result of the union the three samples. So we will have, as example, PSUs that are common to the three years, PSUs that are common only for two of the three years, and PSUs that are present only in one year. Our final structure of PSUs will be the union of all these.

Then we will create the replications from this common structure.

In the standard JRR methodology, replications are created by eliminating one PSU at a time, a replication being identified by the particular PSU (say k) eliminated in constructing it. In the combined dataset, the concerned PSU, if present, is eliminated from all the three cross-sectional datasets to obtain a combined replication (see Figure 3). Note that, for reasons noted in Section 4, there may be some differences across replications in the final sample structure obtained.

Next, we assign to this common structure new weights equal to the average of the

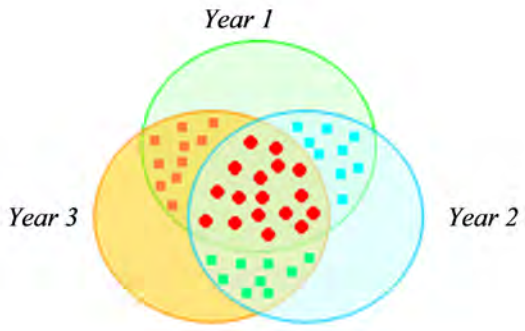

Figure 2.

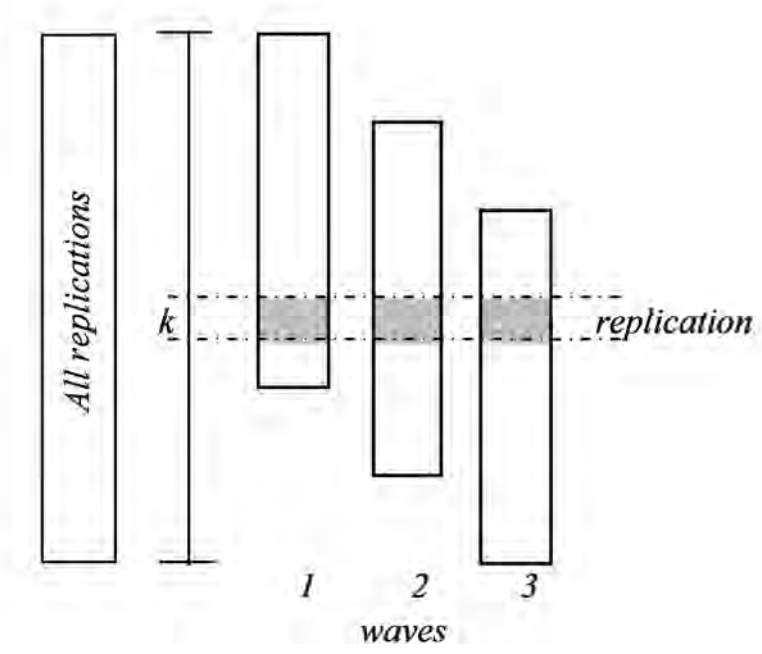

Figure 3.

weights of the three years:

$$w^{(t)Common} = (w)^{Average} = (w^{(1)} + w^{(2)} + w^{(3)})/3 \tag{5.1}$$

For each year (t) and for each replication (k), we can estimate ${y_k}^{(t)}$ where $t = 1, 2, 3$ and from this, the required statistic

$${y_k}^{Average} = \sum_t a_t {y_k}^{(t)}; \tag{5.2}$$

that in our example on three waves is just

$${y_k}^{Average} = (y^{(1)} + y^{(2)} + y^{(3)})/3 \tag{5.3}$$

6. SAS routines for variance estimation of cumulative poverty measures with JRR

In order to develop routines for variance estimation of cumulative poverty measures with JRR methodology, we have used the SAS software.

The SAS software (Statistical Analysis System) is one of the most well known statistical software. It is a software suite developed by SAS Institute for advanced analytics, business intelligence, data management, and predictive analytics. It is the largest market-share holder for advanced analytics (D'Agostino *et al.*, 2015). It is also used by many international and national statistical offices, such as Eurostat and NSIs of Member States.

Our routines have been developed for cumulation of three consecutive waves of EU-SILC. In this specific example, they give results for three poverty measures (see below) at NUTS2[1] regional level for Spain.

6.1. The required datasets

The dataset for which our routines have been developed and adapted is EU-SILC cross-sectional dataset for Spain for years 2009, 2010 and 2011. Thanks to the cooperation in a OECD project (Piacentini, 2014), we have availability of full information on the sample structure, namely the strata, the PSUs and their linkage across the three waves.

Our routines, developed on the basis of the procedure described in Section 5, can be used only if *full information on sample structure is available*.

When complete information on the sample structure is not available and its linkage at cross section level is not feasible, an alternative (indirect) method has been developed by

[1] NUTS is an abbreviation for Nomenclature of Statistical Territorial Units. This is Eurostats hierarchical classification of regions, from Member States (NUTS 0) down to smaller areas

our research group to approximate the correlation across the cross-sectional waves, using information from the longitudinal data of EU-SILC which enables linking individuals and households across years (Betti *et al*. 2015).

6.2. The SAS program

Our routines are collected in a unique program that can be divided into three parts.

(1) A first part where a common structure of strata and PSUs is created for the three years concerned, as described in Section 5, beginning from merging the public EU-SILC dataset (UDB) with a dataset with additional variables on the sample structure, namely the strata and PSUs.

(2) Then the datasets for the three years are prepared with all the needed variables.

(3) This is the *core* of the program. In this part all the developed routines for the estimation of the indicators and of their variances can be found.

Let now describe the core of the program.

This part allows the computation at national and regional NUTS2 level of the estimates, of the standard errors and confidence intervals of three poverty measures: the at-risk-of-poverty rate (HCR) using the national poverty line as the 60 percent of the median equivalised disposable income, the ratio of income shares of the percentile S80/S20, and the Gini index [2]. In the present work the equivalised income is set equal to the equivalised disposable income, as defined by the OECD, with reference to the EU-SILC dataset:

$$equivalised\ dispsable\ income = max((HY020/\sqrt{HX040}), 0)$$

where HY020 is the total disposable household income and HX040 is the household size in the EU-SILC dataset.

The core program is composed of several *macros*: two large ones, one embedded into the other, that contain the whole core program and five specific computational ones. Macros can help in several ways . First, with macros you can make one small change in your program and have SAS echo that change throughout your program. Second, macros can allow you to write a piece of code and use it over and over again in the same program

[2] A brief definition of the indicators is the following.
Head Count Ratio or at-risk-of-poverty rate: proportion of the population with equivalised disposable income below 60 percent of the national median.
Inequality of income distribution Gini coefficient: it is defined as the relationship of cumulative shares of the population arranged according to the level of equivalised disposable income, to the cumulative share of the equivalised total disposable income received by that population.
Inequality of income distribution S80/S20 income quintile share ratio: ratio of the shares of equivalised disposable income of the top and the bottom 20 percent of the population.

or in different programs. Third, you can make your programs data driven, letting SAS decide what to do based on actual data values (Slaughter and Delwiche, 2004).

Lets start describing the five computational macros, that are easier and shorter than the others. They are the following.

There are three macros for the computation of each of the three chosen poverty measures:*%macro stat1* (HCR), *%macro stat10* (S80/S20) and *%macro stat11* (Gini). These macros compute the required poverty measures estimates and merge them to the main dataset. They can be applied and reused in any other of SAS program.

The macros for the computation of the HCR and the S80/S20 use also a forth macro called *% macro perc_bound*. This macro allows calculating any kind of weighted percentile of a distribution. The value that it calculates is the linear interpolation of the percentile. It means that if any real value of the distribution lies in this percentile, a value between the two nearest values - *above* and *below* the percentile- is interpolated. The SAS program doesnt calculate percentiles in such a manner. In fact if any real value of the distribution lies in the required percentile, the SAS function takes the nearest real value above the percentile. In calling this macro, *%macro perc_bound (perc)*, the percentile to be computed (i.e. a number between 0 and 100) should be specified inside the parenthesis.

The fifth computational macro is the one that implements the JRR methodology described in Section 3 and it is called $\%macrojrr\_var(local)$. In this case, when calling the macro, the number of PSUs of the dataset should be specified inside the parenthesis.

This macro is a cycle that is repeated for each replication (PSU). In order to estimate the standard errors, the required measures are estimated inside the replications. In fact, inside the cycle of the macro, so for each replication, there is the computation of the $'\%stat\&'$, the macros for the poverty measures. Inside the replications we also reallocate the weights. Once a PSU is deleted, its weights are assigned to the other PSUs in the same stratum, such that the total sum of the weights does not change (see formula 3.3 and the subsequent description).

The output of this macro is a dataset where the observations (the lines of the dataset) are all the PSUs, with the estimates of the poverty measures computed for each of them.

Let now describe the two main large macros that contain the whole program.

The first one is the macro *%macro waves (waves)*. It is embedded in the second large macro that we are going to present below. This macro is a cycle that repeats a series of computations for each considered wave. In our case we have three waves, so, in calling the macro, the number 3 should be inserted inside the parenthesis.

It begins by keeping as input the datasets prepared with all the necessary variables in part 2 ($working\_pop_1$). Then it transforms the original weights so that their sum is equal to 1, dividing them by the sum of the weights; it also compute the sum of the weights by stratum and by PSU that are needed inside the $\%macrojrr\_var$. After the rescaling of the weights, the $\%macrojrr\_var$ can be called and applied for each wave.

Finally the poverty measures for the entire dataset are estimated recalling their macros.

```
data working_pop0;set working_pop_1;w0=db090;w_sub=w0*I; ;run;
proc sort data=working_pop0; by country ; run;
proc univariate data=working_pop0 noprint;output out=sum_w sum=sum_w
sum=sum_w_sub ;var w0 w_sub ;by country ;run;
data sum_w;set sum_w (keep=sum_w sum_w_sub country );run;
data working_pop; merge working_pop0 sum_w;by country ;
w0=db090/sum_w; ws=w_sub/sum_w_sub;whij=w0; run;
proc sort data=working_pop;by stratum;run;
proc univariate data=working_pop noprint;output out=weight_str sum=w_h ;
var w0 ;by stratum;run;
proc sort data=working_pop;by psu;run;
proc univariate data=working_pop noprint; output out=weight_notuse sum=w_c ;
var w0 ;by psu;run;
proc sort data=working_pop;by stratum;run;
data working2;merge working_pop weight_str;by stratum;run;
proc sort data=working2;by psu;run;
data working3;merge working2 weight_notuse;by psu;w0_old=w0;ws_old=ws; run;
data h;set h0;run; %jrr_var(&psu);
```

Figure 4. Note: DB090 are the EU-SILC cross-sectional weighs. h0 is an empty dataset where the output of JRR is stored. h0 contains the following variables: country (a country code), ah (the number of PSUs per stratum), psu, stratum, stat (the required estimates at replication level).

Two datasets for each wave of the three considered are the output of this macro: one dataset is the output of the macro jrr_var and the second one contains the estimates of the chosen indexes for each wave. Lets now describe the largest macro that contains the whole program. It is $\%macrosub\_ciclo(sub\_ciclo\_start, sub\_ciclo\_end, psu)$.

Also in this case, the macro is a cycle that repeats the whole program for each required measure. In calling the macro, inside the parenthesis three numbers should be specified. They are the following.

sub_ciclo_start and sub_ciclo_end represents the indexes to be computed. In this specific program, they can range from 1 to 60 according to the list that can be found in Appendix A. This list is present at the end of the program. &j corresponds to the 60 measures that can be computed: the three chosen poverty measures at national level and the estimate for each of them for all the NUTS2 regions of the Spain.

If we choose $sub\_ciclo\_start = 1$ and $sub\_ciclo_end = 60$, the output will contain results for all the 60 measures. The number chosen for sub_ciclo_start should be smaller or equal to the one chosen for sub_ciclo_end.

The third number to be specified inside the parenthesis is the total number of PSUs of the dataset.

This macro begins with the above described $\%macrowaves$, so $\%macrowaves$ is repeated for each chosen index.

The present macro keeps the outputs from the macro $\%macrowaves$ and merges them over the years. The results are two datasets, one at replication level with the estimate for the three waves and a second with the final required estimate for each wave. In each of these two datasets the average of the measures is computed

measure=(est1+est2+est3)/3; run;

where est1, est2 and est3 are the estimate computed for each wave.

From the file at replication level (called *h*) variances are computed following the equations presented in Section 3, using the above defined measure and ah, the number of PSUs per stratum.

Variances are also trimmed if too large. If some of them are larger then 6 times the mean of the variances (*limit*, in the routine), they are set equal to exactly 6 times the mean of the variances. This is a very common procedure in order to avoid unstable estimates.

Finally the standard errors are computed.

The final step, outside the macro is the computation of the confidence interval for all the measures.

The final output gives a table with the names ($subpopulation$), the estimates ($measures$), the standard errors ($stat\_se$) and the confidence intervals (ci_upp, ci_low) of all the required measures. Below we report an example for 2 indexes (Table 6.1).

```
proc univariate data=h noprint; output out=jks mean= ah; var ah; by stratum; run;
proc sort data=h;by country stratum;run;
proc univariate data=h noprint;output out=jkm sum= yhsum_stat;var measure;by
country stratum;run;
proc sort data=jkm; by stratum; run;
proc sort data=jks; by stratum; run;
data jk; merge jkm jks; by stratum; run;
data jk; set jk; yh_stat=yhsum_stat/ah;run;
proc sort data=h; by stratum; run;
proc sort data=jk; by stratum; run;
data prova; merge h jk; by stratum; factor=(ah-1)/ah; run;
data jk2_0;set prova;statdif2_0=(measure-yh_stat)**2;run;
proc sort data=jk2_0; by country; run;
proc univariate data=jk2_0 noprint; output out=mean mean=mean; var statdif2_0; by
country; run;
data jk2; merge jk2_0 mean; by country; statdif2=statdif2_0;limit=6*mean;
if statdif2_0 gt limit then statdif2=limit;run;
proc univariate data=jk2 noprint;  output out=var_stat
sum= stat_v; var statdif2 ; weight factor;by country;run;
data se_stat; set var_stat;stat_se=stat_v**0.5;run;
```

Figure 5.

Table 1. Example of the final output.

subpopulation	measure	*stat_se*	*ci_upp*	*ci_low*
S80/S20 ES70	6.842648	0.839775	8.488607	5.196688
Gini ES11	0.302188	0.021438	0.344206	0.260169

Table 2. Average over three years(2009, 2010, 2011), Spain, national results.

SPAIN	(a)	(b)	(c)	(d)	(e)
HCR 60% national poverty line	22.0	0.48	0.31	0.65	1.12
S80/S20	6.5	0.15	0.11	0.73	1.27
Gini	33.8	0.30	0.24	0.80	1.39

(a) Estimate 2011
(b) s.e. 2011
(c) s.e. 3-years average
(d) ratio s.e. 3-years average over s.e. single year
(e) ratio s.e. 3-years average over s.e. 3-years average for independent samples

7. Empirical results

As already mentioned, our routines have been applied to the EU-SILC datasets for years 2009, 2010, 2011 for Spain. We have already mentioned that we have access to full information on Spain sample structure, thanks to a project with OECD.

Below we report some of the results at national and regional level.

The results of Table 2 are at national level and they show a sensible reduction of the standard error (s.e.) using the three years average with all the three measures concerned. The reduction of the standard errors that we get using the three years averages compared to the estimate for a single year (column (d)), ranges from 20% for Gini index, up to 35% for HCR.

Our estimates of the standard errors averaged over three years are, correctly and reasonably, higher than those in case of average on completely independent sample (column (e)). Because of the partial overlap of the EU-SILC sample through the years, the gain in precision of the cumulated estimates is reduced from the one that we could have in case of completely independent samples.

In Table 3 results for Spain are presented at the regional NUTS2 level. The comparison of standard errors between one-year and three-year estimates is more complex here, given the instability of the one-year estimates because of small samples. This problem is particularly evident for regions with a small number of PSUs. The cumulated estimates in fact have been chosen to overcome to the high instability of the single year estimates.

Generally also in this case we can appreciate a reduction of the standard error, both in mean and median, for all the three measures. The reduction can be better appreciated considering the median, which is not affected by extreme values that are present in the results given the instability of the estimates for single years. The largest reductions in this case are in S80/S20, where, in median, we have a decrease of 38%; for HCR the decrease in median is of 20% and for Gini index is of 16%.

	HCR 60%, national p.l.				S80/S20				Gini			
	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)
ES11	18,4	1,17	0,83	0,71	5,6	0,36	0,18	0,50	31,2	2,49	2,14	0,86
ES12	10,3	1,04	0,71	0,68	4,7	0,42	0,22	0,52	29,7	2,09	2,40	1,15
ES13	19,9	0,95	2,08	2,20	6,1	0,29	0,22	0,78	32,8	3,45	2,69	0,78
ES21	11,0	0,84	0,47	0,57	5,1	0,28	0,17	0,59	30,1	1,30	1,15	0,89
ES22	9,5	0,83	0,60	0,72	4,8	0,48	0,19	0,40	28,7	1,93	1,43	0,74
ES23	25,0	1,50	0,98	0,65	7,6	0,54	0,28	0,51	34,5	3,30	2,59	0,79
ES24	17,9	1,52	2,72	1,79	6,2	0,36	0,25	0,69	31,1	2,70	2,09	0,77
ES30	16,4	1,91	0,93	0,49	6,0	0,25	0,21	0,83	32,4	1,81	1,49	0,82
ES41	23,0	1,39	1,33	0,96	6,3	1,64	0,60	0,36	33,2	3,30	2,76	0,84
ES42	31,8	1,74	1,18	0,68	7,4	0,97	0,47	0,48	36,3	4,72	4,09	0,87
ES43	35,2	2,05	2,33	1,14	7,3	0,59	0,36	0,62	36,0	5,01	5,31	1,06
ES51	17,1	1,03	0,55	0,53	5,4	0,29	0,17	0,58	30,8	2,02	1,57	0,78
ES52	20,1	1,01	1,04	1,03	5,4	0,41	0,27	0,66	31,2	2,77	2,45	0,88
ES53	18,6	1,13	1,84	1,64	6,7	0,48	0,43	0,90	32,6	2,75	2,20	0,80
ES61	32,0	1,17	0,94	0,80	8,6	0,52	0,31	0,61	36,9	3,76	3,15	0,84
ES62	24,7	1,56	1,31	0,84	5,4	0,35	0,38	1,08	30,2	4,29	3,93	0,91
ES63	23,5	1,34	2,46	1,84	5,3	0,39	0,85	2,16	35,5	2,93	4,96	1,69
ES64	31,8	2,15	1,51	0,70	10,3	0,93	0,64	0,68	39,1	3,93	2,85	0,73
ES70	32,0	1,19	1,07	0,90	7,6	0,52	0,84	1,61	37,9	4,28	3,50	0,82
Mean				0,99				0,77				0,90
Median				0,89				0,62				0,84

Table 3: Spain, NUTS 2 results

(a) Estimate 2011 (b) s.e. 2011 (c) s.e. 3-years average

(d) ratio s.e. 3-years average over s.e. single year

8. Conclusions

In this paper we have described practical procedures for the estimation of variances of complex statistics such as poverty rates under complex sample designs. Our specific concern is with estimation at subnational (regional) level, where additional numerical difficulties can arise as a consequence of the reduced sample sizes available.

We have decided to treat this problem as cumulating the estimates for three consecutive years. We have developed SAS routines to get the estimates and standard errors of poverty measures cumulated over three waves at regional level applied to EU-SILC survey data.

It is necessary to underline again that such procedures can be applied only if *full information on the sample structure is available*.

We have developed an alternative procedure for dealing with the situation where full information on the sample structure is lacking. In particular, the most serious lack of information in public-use EU-SILC datasets is that micro-level linkage across surveys waves is possible only in the longitudinal version of the data, and not in the larger cross-sectional version. As reported elsewhere (Verma *et al.* 2010), in nutshell the procedure is as follows.

If it is not possible to link the cross-sectional datasets, variances of the measures cannot be computed through the formula 2.2, because the correlation cannot be quan-

tified. We have decided to impute this correlation from the *longitudinal dataset* of the EU-SILC, in which annual data are linkable across the waves. With this assumption we are able to compute the variances for averaged measures.

To compute these variances at regional level, in order to simplify the calculation, we estimate them from the decomposition of the design effect (see Betti *et al*., 2015 for more details) as:

$$V^{(G)} = {V_{SRS}}^{(G)} \cdot {d_W}^{2(G)} \cdot {d_H}^{2(G)} \cdot {d_D}^{2(G)} \cdot {d_X}^{2(G)} \cdot {d_R}^{2(G)}$$

where
(G) stay for region;
V_{SRS} is the required variance under a simple random sample;
${d_W}^2$ is the effect of the weights, known also as Kish factor;
${d_H}^2$ is the effect of clustering of persons within households;
${d_D}^2$ is the effect of clustering of persons and households within dwellings;
${d_X}^2$ is the effect of multi-stage sampling, stratification and other design complexities;
${d_R}^2$ is the effect of correlation in non independent samples.

All the previous quantities can be easily computed on the basis of the following considerations:

- Quantities d_W, d_H and d_D do not depend on structure (especially clustering) of the sample, and can be easily estimated from samples of elements at the regional level.
- d_X and d_R can be set equal to the corresponding values at national level.
- Also ${V_{SRS}}^{(G)}$ can be inferred from ${V_{SRS}}^{(C)}$, the corresponding value at country level (C stays for country) on the basis of the very reasonable assumption that the coefficient of variation of a required Y measure.

$$CV^2(Y) = n \cdot \frac{V_{SRS}(Y)}{Y^2}$$

at the regional level are the same as that at the country level, so that

$${V_{SRS}}^{(G)} = \frac{CV^{2(G)}}{n^{(G)}} \cdot (\frac{Y^{(1)} + Y^{(2)} + Y^{(3)}}{3})^{2(G)}$$

.

Appendix A. List of indexes computed by the routines

if &j eq 1 then Subpopulation='HCR 60% national p.l.';
if &j eq 2 then Subpopulation='S80/S20';

```
if &j eq 3 then Subpopulation='Gini';
if &j eq 4 then Subpopulation='HCR 60%, national p.l. ES11';
if &j eq 5 then Subpopulation='HCR 60%, national p.l. ES12';
if &j eq 6 then Subpopulation='HCR 60%, national p.l. ES13';
if &j eq 7 then Subpopulation='HCR 60%, national p.l. ES21';
if &j eq 8 then Subpopulation='HCR 60%, national p.l. ES22';
if &j eq 9 then Subpopulation='HCR 60%, national p.l. ES23';
if &j eq 10 then Subpopulation='HCR 60%, national p.l. ES24';
if &j eq 11 then Subpopulation='HCR 60%, national p.l. ES30';
if &j eq 12 then Subpopulation='HCR 60%, national p.l. ES41';
if &j eq 13 then Subpopulation='HCR 60%, national p.l. ES42';
if &j eq 14 then Subpopulation='HCR 60%, national p.l. ES43';
if &j eq 15 then Subpopulation='HCR 60%, national p.l. ES51';
if &j eq 16 then Subpopulation='HCR 60%, national p.l. ES52';
if &j eq 17 then Subpopulation='HCR 60%, national p.l. ES53';
if &j eq 18 then Subpopulation='HCR 60%, national p.l. ES61';
if &j eq 19 then Subpopulation='HCR 60%, national p.l. ES62';
if &j eq 20 then Subpopulation='HCR 60%, national p.l. ES63';
if &j eq 21 then Subpopulation='HCR 60%, national p.l. ES64';
if &j eq 22 then Subpopulation='HCR 60%, national p.l. ES70';
if &j eq 23 then Subpopulation='S80/S20 ES11';
if &j eq 24 then Subpopulation='S80/S20 ES12';
if &j eq 25 then Subpopulation='S80/S20 ES13';
if &j eq 26 then Subpopulation='S80/S20 ES21';
if &j eq 27 then Subpopulation='S80/S20 ES22';
if &j eq 28 then Subpopulation='S80/S20 ES23';
if &j eq 29 then Subpopulation='S80/S20 ES24';
if &j eq 30 then Subpopulation='S80/S20 ES30';
if &j eq 31 then Subpopulation='S80/S20 ES41';
if &j eq 32 then Subpopulation='S80/S20 ES42';
if &j eq 33 then Subpopulation='S80/S20 ES43';
if &j eq 34 then Subpopulation='S80/S20 ES51';
if &j eq 35 then Subpopulation='S80/S20 ES52';
if &j eq 36 then Subpopulation='S80/S20 ES53';
if &j eq 37 then Subpopulation='S80/S20 ES61';
if &j eq 38 then Subpopulation='S80/S20 ES62';
if &j eq 39 then Subpopulation='S80/S20 ES63';
if &j eq 40 then Subpopulation='S80/S20 ES64';
if &j eq 41 then Subpopulation='S80/S20 ES70';
if &j eq 42 then Subpopulation='Gini ES11';
if &j eq 43 then Subpopulation='Gini ES12';
if &j eq 44 then Subpopulation='Gini ES13';
if &j eq 45 then Subpopulation='Gini ES21';
if &j eq 46 then Subpopulation='Gini ES22';
if &j eq 47 then Subpopulation='Gini ES23';
if &j eq 48 then Subpopulation='Gini ES24';
```

```
if &j eq 49 then Subpopulation='Gini ES30';
if &j eq 50 then Subpopulation='Gini ES41';
if &j eq 51 then Subpopulation='Gini ES42';
if &j eq 52 then Subpopulation='Gini ES43';
if &j eq 53 then Subpopulation='Gini ES51';
if &j eq 54 then Subpopulation='Gini ES52';
if &j eq 55 then Subpopulation='Gini ES53';
if &j eq 56 then Subpopulation='Gini ES61';
if &j eq 57 then Subpopulation='Gini ES62';
if &j eq 58 then Subpopulation='Gini ES63';
if &j eq 59 then Subpopulation='Gini ES64';
if &j eq 60 then Subpopulation='Gini ES70';
```

References

Betti, G., Verma, V. and Gagliardi, F. (2015). Variance estimation for cumulative and longitudinal poverty indicators from panel data at regional level, in *Analysis of poverty data by small area methods* (ed. Pratesi M.), Wiley and Sons, New York.

DAgostino, A., Neri, L. and Gagliardi, F. (2015). Appendix on software and codes used in the book, in *Analysis of Poverty Data by Small Area Methods* (ed. Pratesi M.), Wiley and Sons, New York.

Piacentini, M. (2014). *Measuring Income Inequality and Poverty at the Regional Level in OECD Countries*, OECD Statistics Working Papers, 2014/03, OECD Publishing. http://dx.doi.org/10.1787/5jxzf5khtg9t-en.

Slaughter, S.J., Delwiche, L.D., SAS (2004). *Macro Programming for Beginners*, SUGI 29 Proceedings, Montral, May 9-12, 2004.

Verma, V., Betti, G. (2011). *Taylor linearization sampling errors and design effects for poverty measures and other complex statistics*, Journal of Applied Statistics, 38(8), pp. 1549-1576.

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., Gagliardi, F. and Ferretti, C. (2013). Cumulation of poverty measures to meet new policy needs, in *Advances in Theoretical and Applied Statistics*. Torelli, N. and Pesarin, F.; Bar-Hen, Avner (Eds.) 2013, XIX, Springer.