Design effect estimation methodology by means of intraclass correlation.

Silvia Biffignandi[1], *Univ. of Bergamo*

Stefano Falorsi[2], *National Statistical Institute of Italy (ISTAT)*

Abstract: The paper presents a methodology for obtaining a more efficient estimator of design effect with reference to complex multistage sampling designs used for large scale surveys on households. Finally, the results of a simulation using real data are presented.

1. Introduction

The methodology proposed in this paper aims at providing, for each estimate, an efficient estimator of the *design effect* (*deff*) with reference to each target domain. This statistic, first proposed by Kish (1965), is expressed by the ratio between the variance of the estimator of the parameter of interest under the complex sample design employed with respect to that of an hypothetical simple random sample of equal size in terms of elementary units. Thus it measures the inflation or deflation of the variance resulting from the sample design adopted, compared to that of the design of the simple random sample used as a basic reference design. Since its first formulation, *deff* has been used extensively in the field of sampling, both at the sample design planning stage and in the critical analysis of the design adopted. *Ex-ante*, during the planning phase of sampling design, *deff* should be estimated on the basis of information derived from previous surveys of the same type (the same variables of interest has to be considered). On the contrary, *ex-post*, during the estimation phase it is possible to use data derived from the survey itself.

The proposed methodology may be applied either in the case of *planned domains*, obtained as aggregation of complete design strata, or when *unplanned domains*, cutting across design strata, are of interest. These domains are considered as *small domains* when sampling errors of the direct domain estimates are considered too high to allow their publication. As noted by Kalton (1994), the estimator, of the sampling variance may prove to be particularly complex in the case of multi-stage sample designs, such as those adopted in surveys on households and individuals carried out by the main centers for official statistical information at national and international levels. In fact, of itself, it may happen frequently that the number of primary units selected in each stratum, and/or overall in each domain of interest, it is low (no more than a few units) and

[1] Silvia Biffignandi, Dipartimento di Matematica, Statistica, Informatica e Applicazioni, Università di Bergamo, Via Caniana n. 2, 24127 Bergamo – Italia.

[2] Stefano Falorsi, Research Manager Methodological Direct., ISTAT, National Statistical Institute of Italy.

Paper prepared within the research projects grant ex 60% , 2007 and 2008, University of Bergamo, responsible Silvia Biffignandi.

ultimate clusters variance estimation – which is based on deviations between estimates of the totals of the variable of interest with reference to primary sampling units falling within the domain (or the stratum) – may be extremely imprecise with few degrees of freedom. In such cases, Kalton suggests calculating a *synthetic estimation* of the variance relative to the domain of interest. This is obtained by multiplying the estimator of the domain (or stratum) sampling variance with reference to the simple random sample design for a synthetic estimation of *deff* calculated on a suitable macro-domain (including the domain, or the stratum of interest). The methodology proposed in this paper is influenced by the above-mentioned ideas. Getting a more precise estimation of sampling variances and *deff*s will produce positive effects both on sample allocation, as well as on the choice of the estimator.

As regards the allocation phase in large-scale surveys, it should be noted that, in most cases, surveys have multiple objectives, which means it is unrealistic to hope for sample sizes that can guarantee predetermined levels of precision for all estimates of interest. Furthermore, an additional problem – which arises in nearly all the large scale surveys – is represented by the need to produce parameter estimations for a high number of planned domains under study. In seeking out – independently for each planned domain under study – an optimum solution to this problem, as a result we are trying to reconcile different tasks, each of which demands for a different type of response and whose solutions may be at odds with each other. Following a direction pursued by many other national statistical institutions to solve such problems, National Statistical Institute of Italy (ISTAT) has investigated multivariate allocation methodologies that take a global view of the problem of the optimum determination of sample size given a multiplicity of objectives and ties. More precisely, the methodology in question allows for the determination of the minimum sample size able to guarantee – with the desired level of precision – the production of parameter estimations of interest with reference to a variety of planned domains. Clearly, such a solution to the problem may be excellent in a global sense, but at the level of each individual domain under study it provides solutions that are generally less efficient than those which may be obtained via an autonomous determination of sample size for each planned domain under study. The methodology studied and applied by ISTAT, presented in Falorsi and Russo (2001), generalizes – in the context of multi-stage sample design and in the context of multiple planned domains under study – the method proposed by Bethel (1989), aimed at determining optimum size from a multivariate viewpoint, and related to the case of a design with one stage of stratification and with a single domain under study. More specifically, the generalization for multi-stage designs, proposed in Falorsi and Russo (2001), is based on the inflation of the estimator of the variance for each stratum, under a simple random sampling, by means of an estimator of design effect referred to the stratum itself.

2. The estimation of the intraclass correlation coefficient

2.1. General formulation

The proposed methodology for the estimation of design effect is based on a reformulation of the *intra-class* correlation coefficient. To this aim we introduce the following general notation in which *k* and *l* are, respectively, indices of *cluster* and *elementary units* of target population, *G* is the total number of clusters, *E* is the overall number of elementary units and $\overline{E}$ is the mean number of elementary units per cluster; where, hypothetically, the size, E_k, of each cluster is constant and therefore equal to $\overline{E}$. Furthermore, with reference to the target variable *y*, Y_{kl} is the observed value for elementary unit *l* belonging to cluster *k*, Y_k is the total relative to cluster *k*, Y is the overall total. Given the notation introduced, the intra-class correlation coefficient of variable *y* may be expressed as

$$\rho_y = \frac{\frac{G-1}{G}\frac{\mathrm{S}_{y1}^2}{\overline{E}} - \sigma_y^2}{(G-1)\,\sigma_y^2} \qquad (1)$$

in which

$$\mathrm{S}_{y1}^2 = \frac{1}{G-1}\sum_{k=1}^{G}(\mathrm{Y}_k - \overline{\overline{\mathrm{Y}}})^2 = \frac{1}{G-1}\,(\,\mathrm{SQ}_{y1} - G\,\overline{E}^2\,\overline{\overline{\mathrm{Y}}}^2\,) \qquad (2)$$

and

$$\sigma_y^2 = \frac{1}{G\overline{E}}\sum_{\mathrm{k}=1}^{\mathrm{G}}\sum_{l=1}^{\overline{E}}(\mathrm{Y}_{kl} - \overline{\overline{\mathrm{Y}}})^2 = (G\overline{E})^{-1}\mathrm{SQ}_y \; - \; \overline{\overline{\mathrm{Y}}}^2, \qquad (3)$$

being $\overline{\mathrm{Y}} = \mathrm{Y}/G$, $\overline{\overline{\mathrm{Y}}} = \mathrm{Y}/(G\overline{E})$; SQ_{y1} and SQ_y are the sum of squares (SQ) of the target variable respectively for the cluster totals and for the elementary units. Last quantities are formally expressed as

$$\mathrm{SQ}_{y1} = \sum_{k=1}^{G}\mathrm{Y}_k^2 \qquad , \qquad \mathrm{SQ}_y = \sum_{k=1}^{G}\sum_{l=1}^{E_k}\mathrm{Y}_{kl}^2 \qquad (4)$$

By substituting the expression of S_{y1}^2 with that of ρ_y an alternative formula for the intraclass correlation coefficient is obtained

$$\rho_y = \frac{1}{G-1}\{(G\overline{E})^{-1}\sigma_y^{-2}\;[\mathrm{SQ}_{y1} \; - \; G\overline{E}^2\,\overline{\overline{\mathrm{Y}}}^2] - 1\} \qquad (5)$$

In the case in which *y* is a binary variable equal to "1" if the unit under observation possesses characteristics of interest (for example, is employed), or otherwise equal to "0", the expression of intra-class correlation coefficient, denoted as $_{\text{bin}}\rho_y$, may be further simplified, since the mean value $\overline{\overline{\overline{\text{Y}}}}$ and the sum of squares SQ_y coincide with the relative frequency, P_y , of units having the characteristic of interest and

$$\sigma_y^2 = (G\overline{E})^{-1}\ \text{SQ}_y = \text{P}_y(1-\text{P}_y)\,. \tag{6}$$

Then formula (5) becomes

$$_{\text{bin}}\rho_y = \frac{1}{G-1}\{(G\overline{E})^{-1}[\ \text{P}_y(1-\text{P}_y)]^{-1}\ (\text{SQ}_{y1}\ -\ G\overline{E}^{\,2}\,\text{P}_y^2)\ -1\} \tag{7}$$

As may be noted, the above-mentioned expression depends both on the population size, G and $\overline{E}$, as well as on P_y and SQ_{y1}, the entity of these last two quantities varying with changes in the variable of interest under consideration.

2.2. Approximation of intraclass correlation coefficient

An approximation of (5) may be obtained by substituting the cluster totals, Y_k $(k=1,...,G)$, of the target variable in SQ_{y1} with the corresponding values estimated by means of a linear model. Under the unit level linear mixed model Y_{kl} $(k=1,...,G\,;l=1,...,E_k)$ is a random variable expressed as

$$Y_{kl} = \mathbf{x}'_{kl}\,\boldsymbol{\beta}_k + \nu_k + \varepsilon_{kl} \tag{8}$$

where $\mathbf{x}_{kl}$ and $\boldsymbol{\beta}_k$ are the *q*-dimensional vectors of the auxiliary variables and of the regression coefficients, ε_{kl} and ν_k are independent random variables with 0 mean and constant variances equal to σ_ε^2 and σ_ν^2 respectively.

The simplest model of type (8) is obtained when $\mathbf{x}'_{kl} \equiv 1$ and $\nu_k = 0$ $(k=1,...,G\,;l=1,...,E_k)$. Under this model, the Best Linear Unbiased Predictor (BLUP) of Y_{kl} , on the basis of all the units of the finite population, is

$$Y'_{kl} = E_k\,\overline{\overline{\overline{\text{Y}}}} \tag{9}$$

and the BLUP of SQ_{y1} is

$$SQ'_{y1} = \overline{\overline{Y}} \sum_{k=1}^{G} E_k^2 = \overline{\overline{Y}}\, SQ_E \quad . \tag{10}$$

Substituting (10) in place of SQ_{y1} in (5), the following approximation of intra-class correlation coefficient is obtained

$$\rho'_y = \frac{1}{G-1}\{(G\overline{E})^{-1}\sigma_y^{-2}\,\overline{\overline{Y}}^2\,[\,SQ_E \;-\; G\overline{E}^2\,] - 1\} \quad , \tag{11}$$

that may be expressed in a more compact way as

$$\rho'_y = \frac{1}{G-1}\{\overline{E}^{-1}\,cv_y^{-2}\;\sigma_E^2 \;- 1\}, \tag{12}$$

being

$$cv_y^{-2} = \sigma_y^{-2}\;\overline{\overline{Y}}^2 \quad , \quad \sigma_E^2 = \frac{1}{G-1}\sum_{k=1}^{G}(E_k - \overline{E})^2 . \tag{13}$$

For binary variables formula (12) may be further simplified as

$$_{bin}\rho'_y = \frac{1}{G-1}\{\overline{E}^{-1}\frac{P_y}{(1-P_y)}\;\sigma_E^2 \;- 1\}. \tag{14}$$

Formula (14) is particularly noteworthy because it expresses $_{bin}\rho'_y$ as a exclusive function of P_y, apart from the knowledge of the population sizes (G and $\overline{E}$) and the variability, σ_E^2, of cluster sizes in terms of elementary units. This quantity may be computed via the distribution of cluster sizes to be found from administrative registers or from census data.

2.3. The case of stratified cluster sampling design

The formulas of the previous pages are derived under the condition of equality of cluster sizes, this is not the real situation for large scale surveys conducted by multistage stratified sampling design in which the Primary Stage Units (PSUs), i.e. the clusters selected at the first stage of selection, are selected, inside each stratum, with probability proportional to size and the average cluster size may be highly variable between strata. Furthermore many of these surveys adopt a stratification of PSUs by size and from the list of PSUs ordered according their

size the "biggest" clusters (in terms of size) are included in the first strata and so on until the last strata that include the "smallest" clusters. For the above reasons it is more convenient to evaluate intra-class correlation coefficient inside each stratum. To this aim we utilize the subscript *h* to denote that the quantities above introduced are related to stratum *h*, being *H* the total number of strata in which the population of interest is divided. With reference to *h*-th stratum $(h = 1,...,H)$, let's denote with: *k* and *l* the cluster and the elementary unit indexes; Y_{hkl} the value of the target variable related to the elementary unit *hlk* ; E_{hk} the population size, in terms of elementary units, referred to cluster *hk*; g_h and G_h the population sizes, in terms of clusters; E_h and $\overline{E}_h$ respectively the total number and the mean number of elementary units per cluster. Then for the *h*-th ($h = 1,...,H$) stratum, formula (5) may be written conformably as

$$\rho_{y,h} = (G_h - 1)^{-1} \{ (G_h \overline{E}_h)^{-1} \sigma_{y,h}^{-2} \ [\mathrm{SQ}_{y1,h} \ - \ G_h \overline{E}_h^2 \ \overline{\overline{Y}}_h^2 \ ] - 1 \} \qquad (15)$$

while for binary case, formula (7) become

$$_{\mathrm{bin}}\rho_{y,h} = (G_h - 1)^{-1} \{ (G_h \overline{E}_h)^{-1} [\ P_{y,h}(1 - P_{y,h})\ ]^{-1} \ (\mathrm{SQ}_{y1,h} \ - \ G_h \overline{E}_h^2 \ P_{y,h}^2) \ - 1 \} \quad (16)$$

Correspondent approximations are possible.
The correspondent approximated expressions are

$$\rho'_{y,h} = (G_h - 1)^{-1} \ \{ \overline{E}_h^{-1} \mathrm{cv}_{hy}^{-2} \ \sigma_{E,h}^2 \ - 1 \ \}, \qquad (17)$$

and

$$_{\mathrm{bin}}\rho'_{y,h} = (G_h - 1)^{-1} \ \{ \overline{E}_h^{-1} \frac{P_{y,h}}{(1 - P_{y,h})} \sigma_{E,h}^2 \ - 1 \ \} \qquad (18)$$

2.4. Estimation of intra-class correlation coefficient

Referring to the case of stratified sampling designs, with reference to each stratum *h*, it possible to derive a *direct* estimator, $\hat{\rho}_{y,h}$ of the correspondent intra-class correlation coefficient (15) estimating, on the basis of the sample data, the unknown quantities dependent by *y*, i.e. $\sigma_{y,h}^{-2}$, $\overline{\overline{Y}}_h^2$ and $\mathrm{SQ}_{y1,h}$. The direct *Horvitz-Thompson* (*HT*) estimators of these unknown quantities, denoted as $\hat{\sigma}_{y,h}^{-2}$, $\hat{\overline{\overline{Y}}}_h^2$ and $\hat{\mathrm{SQ}}_{y1,h}$ are calculated by weighting the observations of the elementary sampling units, selected in stratum *h*, by means of the inverse of their inclusion

probabilities. The direct estimator, ${}_{\text{bin}}\hat{\rho}_{y,h}$, of (16) is obtained by means of the direct estimators $\hat{\mathrm{P}}_{y,h}$, of $\mathrm{P}_{y,h}$, and $\mathrm{S}\hat{\mathrm{Q}}_{y1,h}$.

The same is for the direct estimators of (17) and (18): $\hat{\rho}'_{y,h}$ is a function of the direct estimator, $\mathrm{c}\hat{\mathrm{v}}^{-2}_{y,h}$ $(=\hat{\sigma}^{-2}_{y,h}\ \hat{\bar{\bar{\mathrm{Y}}}}^{2}_{\mathrm{h}})$, of $\mathrm{cv}^{-2}_{y,h}$ while ${}_{\text{bin}}\hat{\rho}'_{y,h}$ is dependent only by $\hat{\mathrm{P}}_{y,h}$.

It is reasonable to expect too that population parameters $\mathrm{cv}^{-2}_{y,h}$ and $\mathrm{P}_{y,h}$ $(h=1,...,H)$ will vary little from stratum to stratum, compared to a high level of variation for the corresponding direct estimates which are based on strata sample sizes. Then for each stratum h $(h=1,...,H)$ a much more precise estimation, denoted as $\mathrm{c}\tilde{\mathrm{v}}^{-2}_{y,h}$ and $\tilde{\mathrm{P}}_{y,h}$, could be obtained by means of a linear mixed model with strata random effects. The fixed effect of the model may borrow strength from the overall population or from sub-populations including the strata. It is important to note that the choice of the more appropriate model will have to balance between variance reduction and increase of the bias. The resulting estimators of intra-class correlation coefficient, denoted as $\tilde{\rho}_{y,h}$, ${}_{\text{bin}}\tilde{\rho}_{y,h}$, $\tilde{\rho}'_{y,h}$, ${}_{\text{bin}}\tilde{\rho}'_{y,h}$, are known as *small area estimators* (Rao, 2003) and in the following will be called *indirect estimators*.

2.5. Regression model for the intra-class correlation coefficient

The condition of equality of cluster sizes rarely holds also inside each stratum, then with the scope of improving the reliability of the estimation of intra-class coefficient, derived in par 2.3 a further passage is proposed aimed to the adaptation of a linear regression model between the intra-class correlation coefficients calculated for the different strata and corresponding factors derived from formulas (17) and (18). To this aim by means of a logarithmic transformation expression (17) and (18) become

$$\begin{aligned}\ln\{(G_h-1)\,\rho'_{y,h}-1\} = \quad & 2\ln(\sigma_{E,h}) \ - \ \ln(G_h-1) \\ & - \ \ln(\bar{E}_h) \ - \ 2\ln(\mathrm{c}\hat{\mathrm{v}}_{y,h})\end{aligned} \tag{19}$$

$$\begin{aligned}\ln\{(G_h-1)\ {}_{\text{bin}}\rho'_{y,h}-1\} = \ & 2\ln(\sigma_{E,h}) \ - \ \ln(G_h-1) \\ & -\ln(\bar{E}_h) \ + \ \ln(\mathrm{P}_{y,h}) \ - \ \ln(1-\mathrm{P}_{y,h})\end{aligned} \tag{20}$$

In the above expressions, the terms $\ln(\mathrm{c}\hat{\mathrm{v}}_{y,h})$ and $\ln(\mathrm{P}_{y,h})$ are unknown and need to be estimated from the sample data while $\ln(\sigma_{E,h})$ may be evaluated from the

sample data too, using a direct estimator $\hat{\sigma}_{E,h}$ of $\sigma_{E,h}$, or may be calculated exploiting the available information on PSU's sizes coming from administrative registers. This estimator will be denoted as $\sigma'_{E,h}$ and will be utilized as in the subsequent expressions.

Then for the scatter of H points $(\hat{\rho}_{y,h}, \sigma'_{E,h}, G_h, \overline{E}_h, \mathrm{c}\hat{\mathrm{v}}_{y,h})$ $(h=1,...,H)$ the following working model may be adapted

$$\begin{aligned}\ln\{(G_h-1)\,\hat{\rho}_{y,h}-1\} = {} & \alpha_1 \ln(\sigma'_{E,h}) + \alpha_2 \ln(G_h-1) \\ & + \alpha_3 \ln(\overline{E}_h) + \alpha_4 \ln(\mathrm{c}\hat{\mathrm{v}}_{y,h}) + \varepsilon_h\end{aligned} \qquad (21)$$

while for binary variables to the scatter of H points $({}_{\mathrm{bin}}\hat{\rho}_{y,h}, \sigma'_{E,h}, G_h, \overline{E}_h, \hat{\mathrm{P}}_{y,h})$ $(h=1,...,H)$ the following model may be fitted

$$\begin{aligned}\ln\{(G_h-1)\ {}_{\mathrm{bin}}\hat{\rho}_{y,h}-1\} = {} & \alpha_1 \ln(\sigma'_{E,h}) + \alpha_2 \ln(G_h-1) \\ & + \alpha_3 \ln(\overline{E}_h) + \alpha_4 \ln(\mathrm{P}_{y,h}) + \alpha_5 \ln(1-\mathrm{P}_{y,h}) + \varepsilon_h\end{aligned} \qquad .(22)$$

Taking into account that $(1-\mathrm{P}_{y,h}) \cong 1$ when $\mathrm{P}_{y,h}$ is small and then $\ln(1)=0$, in many real situations model (22) may be reduced to

$$\begin{aligned}\ln\{(G_h-1)\ {}_{\mathrm{bin}}\hat{\rho}_{y,h}-1\} = {} & \alpha_1 \ln(\sigma'_{E,h}) + \alpha_2 \ln(G_h-1) \\ & + \alpha_3 \ln(\overline{E}_h) + \alpha_4 \ln(\mathrm{P}_{y,h}) + \varepsilon_h\end{aligned} \qquad .(23)$$

Let's denote with $\hat{\alpha}_k$ (k=1,…,5) the least squares estimators of α_k . Then the *direct regression estimator* (*dre*) of intra-class correlation coefficient of the *h*-th stratum $(h=1,...,H)$ is

$$\begin{aligned}\hat{\rho}''_{y,h} = {} & (G_h-1)^{-1}\{\exp[\hat{\alpha}_1 \ln(\sigma'_{E,h})] \cdot \exp[\hat{\alpha}_2 \ln(G_h-1)] \\ & \exp[\hat{\alpha}_3 \ln(\overline{E}_h)] \cdot \exp[\ln(\mathrm{c}\hat{\mathrm{v}}_{y,h})] + 1\}\end{aligned} \quad , \qquad (24)$$

while for binary target variables the *dre* is

$$\begin{aligned}{}_{\mathrm{bin}}\hat{\rho}''_{y,h} = {} & (G_h-1)^{-1}\{\exp[\hat{\alpha}_1 \ln(\sigma'_{E,h})] \cdot \exp[\hat{\alpha}_2 \ln(G_h-1)] \\ & \exp[\hat{\alpha}_3 \ln(\overline{E}_h)] \cdot \exp[\hat{\alpha}_4 \ln(\hat{\mathrm{P}}_{y,h})] + 1\}\end{aligned} \quad . \qquad (25)$$

If in (24) is utilized an *indirect estimator* (*synthetic* or *empirical best linear unbiased predictor*), $\tilde{cv}_{y,h}$, of $cv_{y,h}$ the *indirect regression estimator* (*ire*), $\tilde{\rho}''_{y,h}$, of $\rho''_{y,h}$ is obtained. In the same way for a binary target variable, formula (25) shows that the *ire*, ${}_{bin}\tilde{\rho}''_{y,h}$, of ${}_{bin}\rho''_{y,h}$ is dependent by the indirect estimator, $\tilde{P}_{y,h}$, of $P_{y,h}$.

3. Variance estimation using design effect

In large scale surveys complex multi stage sampling design with stratification of Primary Stage Units (PSUs) and selection of units at different stages with *probability proportional to size without replacement* (*ppswor*) are generally adopted. In this context it is not unusually to consider only the first stage of selection , i.e. ignoring 2nd and later stage of selection, especially when later stages of selection have low probabilities of selecting units and each selected cluster can be considered as an *ultimate cluster*, i.e. the aggregate of all elementary units selected from the same PSU. A common choice, in this context, is to consider the hypothesis of selecting the PSUs with *probability proportional to size without replacement* (*ppswr*). For this simplified framework – in which are available simple estimators formulas for sampling variances not requesting the calculus of second order inclusion probabilities between PSU's - each stratified multistage sampling design can be approximated by a *ppswr* selection of PSUs inside each stratum and all the units selected at later stage of selection are considered as belonging to the same ultimate cluster. Under this simplified context it is useful to add some notation to that given in previous pages. Then let's denote with: Y_{hkl} and π_{hkl} the target variable and the inclusion probability related to the elementary *hlk*-th unit; e_{hk} the sample size, in terms of elementary units, referred to PSU *hk*; U_{dh} the sub-population of elementary units, of size N_{dh} , belonging to domain *d* (*d*=1,…,*D*), being $U_h = \sum_d U_{dh}$ and $N_h = \sum_d N_{dh}$. The symbol $U_d = \sum_h U_{dh}$ denotes the subset of elementary units of the population belonging to domain *d* (*d*=1,…,*D*) and $U = \sum_d U_d$ is the overall population. Furthermore let's denote with $Y_{dhkl} = Y_{hkl} I_{dhkl}$ the value of the target variable for *hkl*-th unit related to its belonging to domain *d*, being $I_{dhkl} = 1$ if unit $(hkj) \in U_d$ and $I_{dhkl} = 0$ otherwise. It is worthwhile to note that when U_d (*d*=1,…,*D*) are *planned domains*, they are obtained as aggregation of complete strata, and population domain do not cut across strata, being $U_{dh} \equiv U_h$, $Y_{dhkl} \equiv Y_{hkl}$ $(d = 1,...,D; h = 1,...,H; k = 1,...,G_h; l = 1,...,E_{hk})$ otherwise in the case of *unplanned domains* they cut across strata.

Given the above,

$$\hat{Y}_d = \sum_{h=1}^{H} \hat{Y}_{dh} = \sum_{h=1}^{H} \sum_{k=1}^{g_h} \hat{Y}_{dhk} = \sum_{h=1}^{H} \sum_{k=1}^{g_h} \sum_{l=1}^{e_{hk}} \hat{Y}_{dhkl} \qquad (26)$$

indicates the *HT* estimator of the total

$$Y_d = \sum_{h=1}^{H} Y_{dh} = \sum_{h=1}^{H} \sum_{k=1}^{G_h} Y_{dhk} = \sum_{h=1}^{H} \sum_{k=1}^{G_h} \sum_{l=1}^{E_{hk}} Y_{dhkl} \qquad (27)$$

where $\hat{Y}_{dhkl} = Y_{dhkl} K_{hkl}$ and $K_{hkl} = \pi_{hkl}^{-1}$ is the sampling weight of unit (*hkl*) on the basis of the adopted multistage sampling design. Then, the sampling variance of the total $\hat{Y}_d$ for the multistage stratified *complex random sampling* (*crs*) design under examination may be expressed as

$$V_{crs}(\hat{Y}_d) = \sum_{h=1}^{H} V_{crs}(\hat{Y}_{dh}). \qquad (28)$$

In the case of two stage sampling plans, e.g. utilized by many large scale surveys on households conducted by face to face interview, under *ppswr* approximation, the variance, $V_{crs}(\hat{Y}_{dh})$, under *crs* for stratum h $(h = 1,...,H)$ is

$$V_{crs}(\hat{Y}_{dh}) = \frac{1}{g_h} \sum_{k=1}^{G_h} Z_{hk} \left(\frac{Y_{dhk}}{Z_{hk}} - Y_{dh}\right)^2 + \frac{1}{g_h} \sum_{k=1}^{G_h} \frac{E_{hk}^2 (1 - f_{hk}) S_{dhk}^2}{e_{hk} Z_{hk}} \qquad (29)$$

where with reference to the *hk*-th PSU, $Z_{hk} > 0$ $(k = 1,...G_h)$ is the probability or relative size assigned to the PSU, being $\sum_{k=1}^{G_h} Z_{hk} = 1$, S_{dhk}^2 is the variance of the target variable Y_{dhkl} values among elementary units and $f_{hk} = (e_{hk}/E_{hk})$ is the sampling rate of the elementary units. An unbiased estimator, $\hat{V}_{crs}(\hat{Y}_d)$, of $V_{crs}(\hat{Y}_d)$ is given by

$$\hat{V}_{crs}(\hat{Y}_d) = \sum_{h=1}^{H} \hat{V}_{crs}(\hat{Y}_{dh}) \qquad (30)$$

being

$$\hat{V}_{crs}(\hat{Y}_{dh}) = \frac{1}{g_h(g_h - 1)} \sum_{k=1}^{G_h} \left(\frac{Y_{dhk}}{Z_{dhk}} - \hat{Y}_h\right)^2 = \sum_{k=1}^{G_h} \frac{g_h}{(g_h - 1)} \left(\hat{Y}_{dhk} - \hat{\bar{Y}}_h\right)^2 . \qquad (31)$$

Formula (29) also holds for multistage stratified sampling under *ppswr* of PSUs, provided that $\hat{Y}_{hk}$ is an unbiased estimator of $\hat{Y}_{hk}$ and that sub-sampling is independent whenever a primary unit is drawn.
In this context let's consider now the sample variances, $V_{crs}(\hat{Y}_{dh})$ and $V_{srs}(\hat{Y}_{dh})$, of $\hat{Y}_{dh}$ $(h=1,...,H)$ referred respectively to the actual complex random sample and to the hypothetical *simple random sample* (*srs*) of equal size, "*e*" , in terms of elementary units, to that related to the real complex sample.
Under *srs* the sampling variances of $\hat{Y}_d$ and $\hat{Y}_{dh}$ $(h=1,...,H)$ are

$$V_{srs}(\hat{Y}_d)=\frac{E^2 S_d^2}{e} \quad \text{and} \quad V_{srs}(\hat{Y}_{dh})=\frac{E_h^2 S_{dh}^2}{e_h} \tag{32}$$

respectively where S_d^2 and S_{dh}^2 represent the overall and *h*-th stratum variances of the target variable values, Y_{dhkl}, among elementary units. In formula (32) the finite population corrections (*fpc*) are ignored supposing $f=(e/E)$ and $f_h=(e_h/E_h)$ near to zero. The direct estimators, $\hat{V}_{srs}(\hat{Y}_d)$ and $\hat{V}_{srs}(\hat{Y}_{dh})$ $(h=1,...,H)$, of variances (32) are obtained by means of direct estimators, $\hat{S}_d^2$ and $\hat{S}_{dh}^2$ $(h=1,...,H)$, of correspondent population variances included in (32).
The design effect of the estimator $\hat{Y}_d$, $\delta(\hat{Y}_d)$, is

$$\delta(\hat{Y}_d)=\frac{V_{crs}(\hat{Y}_d)}{V_{srs}(\hat{Y}_d)} \quad . \tag{33}$$

It is useful to rewrite $V_{crs}(\hat{Y}_d)$ as

$$V_{crs}(\hat{Y}_d)=\sum_{h=1}^{H} V_{srs}(\hat{Y}_{dh})\,\delta(\hat{Y}_{dh}) \tag{34}$$

where $\delta(\hat{Y}_{dh})$ is the design effect of $\hat{Y}_{dh}$ $(h=1,...,H)$

$$\delta(\hat{Y}_{dh})=\frac{V_{crs}(\hat{Y}_{dh})}{V_{srs}(\hat{Y}_{dh})} \quad . \tag{35}$$

For stratum *h*-th $(h=1,...,H)$, under multistage sampling design with *ppswr* selection of PSUs, and supposing PSUs sizes to be reasonably constant in terms of

elementary units, the design effect (Cicchitelli et al., 1992) of $\hat{\mathrm{Y}}_{dh}$ may be approximated via the following function of intra-class correlation coefficient

$$\delta(\hat{\mathrm{Y}}_{dh}) \doteq 1 + \rho_{y_d,h}(\bar{e}_{dh} - 1) \tag{36}$$

where $\rho_{y_d,h}$ denotes that formula (15) is applied to the values Y_{dhkl} $(h = 1,...,H; k = 1,...G_h; l = 1,...E_{hk})$ and $\bar{e}_{dh} = e_{dh}/g_h$ being

$$e_{dh} = \sum_{k=1}^{g_h} \sum_{l=1}^{e_{hk}} \mathrm{I}_{dhkl} \ . \tag{37}$$

Given the above, the design effect formula (33) become

$$\delta(\hat{\mathrm{Y}}_d) \doteq [V_{srs}(\hat{\mathrm{Y}}_d)]^{-1} \sum_{h=1}^{H} V_{srs}(\hat{\mathrm{Y}}_{dh}) \ \rho_{y_d,h}(\bar{e}_{dh} - 1) \tag{38}$$

that under the hypothesis $S_d^2 = S_{dh}^2$ $(h = 1,..., H)$, using (32) may be rewritten as

$$\delta(\hat{\mathrm{Y}}_d) = \frac{e}{E^2} \sum_{h=1}^{H} \frac{E_h^2}{e_h} [\rho_{y_d,h}(\bar{e}_{dh} - 1)] \ . \tag{39}$$

Finally in the case in which larger PSUs are certainly selected, let denote with H_{sr} the number of *self representing* (sr) strata, H_{nrs} the number of non self representing (*nrs*) strata. For *sr* strata are valid the following conditions: $k \equiv h$, $G_h = g_h = 1$, a is the SSU index, M_h and m_h denote the number of population and sample SSUs of *h*-th stratum-PSU respectively, where for stratum (i.e. PSU selected with certainty) *h*-th $(h = 1,..., H_{sr})$ in *sr* domain is

$$E_{sr,h} = \sum_{a=1}^{M_h} E_{sr,ha} \quad , \quad e_{sr,h} = \sum_{a=1}^{m_h} e_{sr,ha} \quad , \quad e_{sr,dh} = \sum_{a=1}^{m_h} e_{sr,hl} \, \mathrm{I}_{dha} \tag{40}$$

If, inside each stratum-PSU, no SSUs are selected, then $M_h = m_h = 1$
while where for stratum *h*-th $(h = 1,..., H_{nsr})$ in *nsr* domain is

$$E_{nsr,h} = \sum_{k=1}^{G_h} E_{nsr,hk} \quad \text{and} \quad e_{nsr,hk} = \sum_{k=1}^{g_h} e_{sr,hk} \quad e_{sr,dh} = \sum_{l=1}^{g_h} e_{sr,hkl} \, \mathrm{I}_{dhkl} \tag{41}$$

If the same number of elementary units in each PSU is selected and under the hypothesis that $S^2_{sr,dh} = S^2_d$ for $h = 1,\ldots,H_{sr}$ and $S^2_{nsr,dh} = S^2_d$ for $h = 1,\ldots,H_{nsr}$ the design effect is given by

$$\delta(\hat{Y}_d) = \frac{e}{E^2} \{ \sum_{h=1}^{H_{sr}} \frac{E^2_{sr,h}}{e_{sr,h}} \delta_{sr}(\hat{Y}_{dh}) + \sum_{h=1}^{H_{nsr}} \frac{E^2_{nsr,h}}{e_{nsr,h}} \delta_{nsr}(\hat{Y}_{dh}) \} \qquad (42)$$

in which $\delta_{sr}(\hat{Y}_{dh})$ $(h = 1,\ldots,H_{sr})$ and $\delta_{nsr}(\hat{Y}_{dh})$ $(h = 1,\ldots,H_{nsr})$ denotes design effect for *sr* and *nsr* strata respectively. Using (36), the above formula may be approximated as

$$\begin{aligned} \delta(\hat{Y}_d) \doteq \frac{e}{E^2} \{ & \sum_{h=1}^{H_{sr}} \frac{E^2_{sr,h}}{e_{sr,h}} [1 + \rho_{y_d, sr, h} (\bar{e}_{sr,dh} - 1)] \\ & + \sum_{h=1}^{H_{nsr}} \frac{E^2_{nsr,h}}{e_{nsr,h}} [1 + \rho_{y_d, nsr, h} (\bar{e}_{nsr,dh} - 1)] \} \end{aligned} \qquad (43)$$

where $\bar{e}_{sr,dh} = e_{sr,dh} / m_h$ and $\bar{e}_{nsr,dh} = e_{nsr,dh} / g_h$ are the mean number of elementary units at PSU level for *h*-th stratum in *sr* and in *nsr* domain. On the basis of what has been described in previous paragraphs, in order to get an estimator, $\tilde{\delta}_{\omega}(\hat{Y}_d)$ of $\delta(\hat{Y}_d)$, as precise as possible, different estimators of the unknown intra-class correlation coefficients $\rho_{y_d, sr, h}$ and $\rho_{y_d, nsr, h}$ may be used:

(1) direct estimators, $\hat{\rho}_{y_d, sr, h}$ and $\hat{\rho}_{y_d, nsr, h}$, or indirect estimators, $\tilde{\rho}_{y_d, sr, h}$ and $\tilde{\rho}_{y_d, nsr, h}$, of expression (15) (or expression (16) if the target variable is binary) as described in par. 2.4. The resulting estimator of $\delta(\hat{Y}_d)$ is denoted as $\tilde{\delta}_{\text{dir}}(\hat{Y}_d)$, i.e. $\tilde{\delta}_{\omega}(\hat{Y}_d) = \tilde{\delta}_{\text{dir}}(\hat{Y}_d)$, being $\omega \equiv \text{dir}$;

(2) direct estimators, $\hat{\rho}'_{y_d, sr, h}$ and $\hat{\rho}'_{y_d, nsr, h}$, or indirect estimators, $\tilde{\rho}'_{y_d, sr, h}$ and $\tilde{\rho}'_{y_d, nsr, h}$, of approximated expression (17) (or expression (18) if the target variable is binary) as described in par. 2.4. The resulting estimator of $\delta(\hat{Y}_d)$ is denoted as $\tilde{\delta}_{\text{apx}}(\hat{Y}_d)$, i.e. $\tilde{\delta}_{\omega}(\hat{Y}_d) = \tilde{\delta}_{\text{apx}}(\hat{Y}_d)$, being $\omega \equiv \text{apx}$;

(3) *dre*, $\hat{\rho}''_{y_d, sr, h}$ and $\hat{\rho}''_{y_d, nsr, h}$, or *ire*, $\tilde{\rho}''_{y_d, sr, h}$ and $\tilde{\rho}''_{y_d, nsr, h}$, under model (24) (or model (25) if the target variable is binary) as described in par. 2.5. The resulting estimators of $\delta(\hat{Y}_d)$ are denoted as $\tilde{\delta}_{\text{dre}}(\hat{Y}_d)$ and $\tilde{\delta}_{\text{ire}}(\hat{Y}_d)$ respectively; i.e. $\tilde{\delta}_{\omega}(\hat{Y}_d) = \tilde{\delta}_{\text{dre}}(\hat{Y}_d)$, being $\omega \equiv \text{apx}$ and $\tilde{\delta}_{\omega}(\hat{Y}_d) = \tilde{\delta}_{\text{ire}}(\hat{Y}_d)$, being $\omega \equiv \text{ire}$.

5. Application of proposed methodology to labor force surveys

5.1. *Introduction*

This section will present the application of the proposed methodology for the estimation in the context of the sampling design used in ISTAT surveys on households, carried out via direct interview.

With this in mind, and for a better understanding of the proposed methodology, it we give a preliminary description of the sample design used in these surveys. They are based on the complex type of design applied with reference to each of the *minimum planned territorial domains*. This implies that the sample is designed in such a way that the most disaggregated territorial domains (obtained by aggregating complete strata) within which a predetermined sample size may be guaranteed if reliable estimations of the parameters of interest are to be produced. Theses domains are made up of provinces, with the *Labour Force Survey* (*LFS*). The design applied for each minimum territorial domain requires a decreasing ranking of municipalities, on the basis of demographic size according to resident population. Once the ranking is determined, the demographically larger municipalities are automatically included in the sample, each one forming a stratum to itself; the territorial domain made up of these municipalities therefore is called the self representative domain. The remaining municipalities, identified as not self representative, are subdivided into strata of roughly constant size (in terms of resident population) and from each one of them a predetermined number of sample municipalities are selected, with *ppswor* sampling. From each municipality in *sr* and from each sample municipality in *nsr*, households are selected with equal probability and without replacement, by means of a systematic selection from official registers; all family components are interviewed. To determine the number of families to be selected from the strata of each minimum territorial domain, the criterion of self-weighting is used. From this it derives that in the *sr* domain strata, the number of sample households is larger according to municipality size, in terms of resident population, while in the *nsr* domain strata, the number of households selected is more or less constant.

To sum up, the sampling design applied in each planned minimum territorial domain requires the use of two different selection designs for, respectively, *sr* and *nsr*. For the former, a one stage stratified cluster design is used: municipalities coincide with strata and households are clusters of individuals; while for the *nsr* domain municipalities, the design is of a two stage stratified type – municipalities represent the primary unit and households are the secondary unit, made up of clusters of individuals.

In this context for the estimation of design effect formula (43) must be considered. To this aim it is useful to note that, for the complex sampling design adopted in ISTAT's surveys on households, above described, $\bar{e}_{sr,dh}$ represents the average household size for *h*-th stratum in *sr* domain $(h = 1,...,H_{sr})$ while $\bar{e}_{nsr,dh}$

is the average number of individuals selected for *h*-th stratum in *nsr* domain $(h = 1,..., H_{nsr})$. Because of U_d $(d = 1,..., D)$ is a planned domain population (i.e. not cutting across strata) obtained as aggregation of complete strata then $\bar{e}_{sr,dh} = \bar{e}_{sr,h}$ and $\bar{e}_{nsr,dh} = \bar{e}_{nsr,h}$.

5.2. *The empirical study*

The following application is intended as a ckeck of the validity of the proposed regression methodology for estimating intra-class coefficients and the corresponding sample design effect. To this end, a comparative analysis between the indirect regression estimator method, $\tilde{\delta}_{\text{ire}}(\hat{Y}_d)$ $(d = 1,..., D)$, proposed in par. 2.5 and the direct estimator of design effect, $\tilde{\delta}_{\text{dir}}(\hat{Y}_d)$ $(d = 1,..., D)$, has been carried out, based on the Monte Carlo simulations. In this way it is possible evaluating the empirical properties of the different methods in terms of mean-square error and bias, calculated in the space of simulated samples.

Basic data from the general population census for 1991 (C91) and for 2001 (C01) referred to region Lazio were used for our analysis. The following variables, taken from the census, were considered for each individual: identification codes for province, municipality and household, and professional status according to the two categories of *employed* and *job-seeking*. Then two new dichotomous variables y_1 and y_2 were built: the first is equal to 1 if the individual is employed but otherwise equal to zero, while the second is equal to 1 if the individual is job-seeking but otherwise equal to zero.

R=500 samples (known as *replications*) were selected from the basic Cen01 data, on the basis of the complex sample design (as described in par. 5.1) adopted by *LFS* adopting the same first and second stage sizes.

For each replication *r* (*r*=1,…,*R*), and for each province[3] *d* (*d*=1,…,*D*) of Lazio estimations, ${}_r\hat{Y}_d$, of the totals, Y_d, for employed and job-seeking individuals, were calculated. Furthermore the sample design effect, ${}_r\tilde{\delta}_{\omega}({}_r\hat{Y}_d)$ of estimation ${}_r\hat{Y}_d$ was calculated, both using $\tilde{\delta}_{\text{ire}}(\hat{Y}_d)$ and $\tilde{\delta}_{\text{dir}}(\hat{Y}_d)$. The *true* value of $\delta(\hat{Y}_d)$ was calculated using C01 data, as described for formula (42). In building the proposed estimators, ${}_r\tilde{\delta}_{\omega}({}_r\hat{Y}_d)$ $(d = 1,..., D)$ for $\omega \equiv ire$, intra-class correlation coefficients $\rho_{y_d, sr, h}$ $(h = 1,..., H_{sr})$ and $\rho_{y_d, nsr, h}$ $(h = 1,..., H_{nsr})$ was calculated by means of the model (25) using the C01 data (Case 1). Another situation was considered (Case 2) in which data of C91 was utilized for the calculation of $\sigma'_{E,h}$ and $P_{y,h}$. Case 2 was considered in order to evaluate whether or not the methodology under study remains valid when using data that had not been

[3] Provinces are the smallest planned domains for *LFS*.

updated. The two alternative estimators are indicated as $\tilde{\delta}_{\text{ire,C01}}(\hat{Y}_d)$ and $\tilde{\delta}_{\text{ire,C91}}(\hat{Y}_d)$. For the construction of $\tilde{\delta}_{\text{ire}}(\hat{Y}_d)$, when calculating estimates $\hat{P}_{y,h}$ mean synthetic estimations were used, based on data from the entire Lazio (to which the provinces belongs), with the aim of stabilizing estimations relative to each single stratum h $(h=1,...,H)$.

For each province *d*, the properties of each estimators $\tilde{\delta}_{\omega}({}_r\hat{Y}_d)$ under study are generally evaluated, in terms of bias and variability, on the basis of the assumed values of the following statistics:

$$\text{RB}_d\left\{\tilde{\delta}_{\omega}(\hat{Y}_d)\right\}=\frac{1}{R}\left[\sum_{r=1}^{R}\frac{\tilde{\delta}_{\omega}({}_r\hat{Y}_d)-\delta(\hat{Y}_d)}{\delta(\hat{Y}_d)}\right]\times 100 \tag{44}$$

$$\text{RRMSE}_d\left\{\tilde{\delta}_{\omega}(\hat{Y}_d)\right\}=\sqrt{\frac{1}{R}\left(\sum_{r=1}^{R}\left[\frac{\tilde{\delta}_{\omega}({}_r\hat{Y}_d)-\delta(\hat{Y}_d)}{\delta(\hat{Y}_d)}\right]^2\right)}\times 100\,. \tag{45}$$

The above-mentioned evaluation criteria, expressed as percentages, respectively measure Relative Bias and the Root of the Relative Mean-Square Error. By calculating the average of all *d* domains (*d*=1,…,*D*) of the absolute values $\text{DR}_d\left\{\tilde{\delta}_{\omega}(\hat{Y}_d)\right\}$ e $\text{REQMR}_d\left\{\tilde{\delta}_{\omega}(\hat{Y}_d)\right\}$ statistics, evaluation criteria are obtained, given by:

$$\overline{\text{RB}}=\frac{1}{D}\sum_{d=1}^{D}\left|\ \text{RB}_d\left\{\tilde{\delta}_{\omega}(\hat{Y}_d)\right\}\right|\ , \tag{46}$$

$$\overline{\text{RRMSE}}=\frac{1}{\text{D}}\sum_{\text{d=1}}^{\text{D}}\text{RRMSE}_d\left\{\tilde{\delta}_{\omega}(\hat{Y}_d)\right\}., \tag{47}$$

In Table 1 are presented the results of the simulation study. In particular, the global evaluation indices, $\overline{\text{RB}}$ and $\overline{\text{RRMSE}}$, for each of the estimators in Case 1 and Case 2 are presented. The analysis of both cases shows that the $\tilde{\delta}_{\text{ire}}(\hat{Y}_d)$ estimator is superior both in terms of bias and mean squared error, although in Case 1, as expected, the estimator gives better results than those of Case 2, in which the intra-class correlation coefficient were estimated on the basis of data (for the calculation of $\sigma'_{E,h}$ and $P_{y,h}$) that had not been updated.

6. Conclusion

The results of the empirical analysis show that the proposed estimation technique, based on indirect regression estimator of intra-class correlation coefficient, improves the quality of the estimation of *deff* with respect to the standard direct method. Then this methodology may be exploited in many phases of the statistical data production proces in which an efficient estimation of design effect is needed.
One of the most important way of utilization of *deff* is related to the allocation of sample sizes into strata and, more in general, for planned domains when complex sampling plans, based on multistage stratified selections of units, are adopted. In particular to overcome the complexity of allocation problem in multipurpose large scale surveys multivariate and multi-domain allocation methodologies are applied. In this case the availability of coefficient estimations of design effects may produce large gains in the quality of the estimates of the target parameters.
Another phase of statistical process of data production in which the proposed methodology may be usefully utilized is related to the estimation of sampling variances for small domains. In that case the standard estimator of sampling variances may be very unstable due small planned and/or observed sample sizes inside each domain.

Table 1: $\overline{RB}$ and $\overline{RRMSE}$ of $\tilde{\tilde{\delta}}_{ire/C01}(\hat{Y}_d)$, $\tilde{\tilde{\delta}}_{ire/C91}(\hat{Y}_d)$ and $\tilde{\tilde{\delta}}_{dir}(\hat{Y}_d)$ for *Employees* and *People looking for a job*

Estimator	$\overline{RB}$	$\overline{RRMSE}$	$\overline{RB}$	$\overline{RRMSE}$
	Employees		*People looking for a job*	
$\tilde{\tilde{\delta}}_{dir}(\hat{Y}_d)$	6.83	1.44	15.18	7.35
$\tilde{\tilde{\delta}}_{ire/C01}(\hat{Y}_d)$	-1.63	0.717	-1.23	1.18
$\tilde{\tilde{\delta}}_{ire/C91}(\hat{Y}_d)$	5.63	0.73	-8.86	0.02

References

Bethel J. (1989), Sample Allocation in Multivariate Surveys, *Survey Methodology*, Vol. 15, 47-57.

Cicchitelli G., A. Herzel, Montanari G.E. (1992), *Il campionamento statistico*, Bologna, Il Mulino.

Falorsi S., Russo A. (2001) Il disegno di rilevazione per le indagini panel sulle famiglie, *Rivista di Statistica Ufficiale*, N. 3/2001, Franco Angeli, 55-90.

G. Kalton (1994) Comment on the paper by M.P.Singh, J.Gambino, H.J.Mantel, *Issues and Strategies for Small Area Data*, Survey Methodology, year 1994, 20, pp.3-22.

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

Rao J. N. K. (2003), *Small Area Estimation*, New York, Wiley.

Further comments on the research topic were presented in Biffignandi S. discussion, The Use of Small Area Estimation in Economic Policy and Social Science Research: Perspectives and Problems, discussion of special issue of Rivista Internazionale di Scienze Sociali, n. CXVI, 4 Ottobre-Dicembre 2008