A Unified Approach for Defining Optimal Multivariate and Multi-Domains Sampling Designs

Piero Demetrio Falorsi and Paolo Righi[1]

Abstract The present paper illustrates a sampling method, based on balanced sampling, practical and easy to implement, which may represent a general and unified approach for defining the optimal inclusion probabilities and the related domain sampling sizes in many different survey contexts characterized by the need of disseminating survey estimates of prefixed accuracy for a multiplicity both of variables and of domains of interest. The method, depending on how it is parameterized, can define a standard cross-classified or a multi-way stratified design. The sampling algorithm defines an optimal solution - by minimizing either the costs or the sampling sizes - which guarantees: (*i*) lower sampling errors of the domain estimates than given thresholds and (*ii*) that in each sampling selection the sampling sizes for all the domains of interest are fixed and equal to the planned ones. It is supposed that, at the moment of designing the sample strategy, the domain membership variables are known and available in the sampling frame and that the target variables are unknown but can be predicted with suitable superpopulation models.

Introduction

A *unified approach* (UI), which is practical and easy to implement, for defining *optimal multivariate multi-domain sampling* is introduced below.
Some parts of this approach have been described with more details in the papers of Falorsi and Righi (2008) and of Righi and Falorsi (2011).

[1] Piero Demetrio Falorsi, Istat; falorsi@istat.it
Paolo Righi, Istat; parighi@istat.it

1. The parameters of interest are *R*x*D* totals, the generic of which, $t_{(dr)} = \sum_{k \in U} y_{rk}\gamma_{dk} = \sum_{k \in U_d} y_{rk}$, represents the total of the variable r (r = 1,…,R) in the *Domain of Interest* (DI) U_d (d=1,…, D) which is a subpopulation (of size N_d) of the population U. The symbols y_{rk} and γ_{dk} denote respectively the value of the r-th (r =1, …, R) variable of interest of the k-th population unit and the domain membership indicator being $\gamma_{dk} = 1$ if $k \in U_d$ and $\gamma_{dk} = 0$ otherwise. The γ_{dk} values are known, and available in the sampling frame.
2. In addition to the DIs, the other subpopulations relevant in the approach are the *Planned Domains* (PDs), U_h (h=1,…, H), which are subpopulations for which the sample designer want to plan and to fix in advance the sample sizes so as to control the accuracy of the domain estimates. The PDs are in general defined as subpopulations of the DIs. As described below in section 2, the definition of the PDs allows to implement different sampling designs.
3. The random selection of the sample s is implemented with the cube algorithm (Deville and Tillè, 2004) respecting the following *balancing equations*: $\sum_{k \in s} \boldsymbol{\delta}_k = \sum_{k \in U} \pi_k \boldsymbol{\delta}_k$ in which, with reference to the unit k, π_k is the inclusion probability and $\boldsymbol{\delta}'_k = (\delta_{1k},...,\delta_{Hk})$ is a vector of indicator variables, available in the sampling frame, being $\delta_{hk} = 1$ if $k \in U_h$ and $\delta_{hk} = 0$, otherwise. The above equations guarantee that in each possible sample selection, the realized sample sizes for the planned domains U_h are fixed and equal to the expected ones. Since the PDs are defined as subpopulations of the DIs, also the latter have planned sample sizes.
4. The unknown y_{rk} values are predicted with a simple *working model*, *M*, $y_{rk} = \tilde{y}_{rk} + u_{rk}$ in which, $\tilde{y}_{rk}$ and u_{rk} (k=1,…,N) denote respectively the predictions and the random residuals which have the following model expectations:

$$E_M(u_{rk}) = 0 \,\forall k;\; E_M(u_{rk}^2) = \sigma_{rk}^2;\; E_M(u_{rk}, u_{rl}) = 0 \;\forall k \neq l\,, \qquad (1.1)$$

further we assume $\sigma_{rk}^2 = \sigma_r^2 v_k^\tau$ where v_k is an auxiliary variable, σ_r^2 and τ are scalar parameters which we assume as known when planning the sampling design. In practice the scalar parameters have to be estimated from pilot or previous survey data.
5. According to Deville and Tillè (2005), an approximation of the Measure of the Accuracy (MA) (eg. the sampling variance or the anticipated variance) of the balanced sampling may be defined as implicit function of the inclusion probabilities and of the squared residual of a generalized linear regression model linking an appropriate transformation of the target variable (which may be known or predicted) to the auxiliary variables involved in the balancing equations. Taking into account the Horvitz Thompson (HT) estimator, $\hat{t}_{(dr)} = \sum_{k \in U} y_{rk}\gamma_{dk} / \pi_k$ of the

totals $t_{(dr)}$ and considering the Anticipated Variance (Isaki and Fuller, 1982) as measure of accuracy , the MA may be expressed by:

$$MA(\hat{t}_{(dr)}) = AV(\hat{t}_{(dr)} \mid \boldsymbol{\pi}) \cong E_M E_p(\hat{t}_{(dr)} - t_{(dr)} \mid \boldsymbol{\pi})^2$$
$$= f \sum_{k \in U} (\frac{1}{\pi_k} - 1) E_M(\eta^2_{(dr)k}) \tag{1.2}$$

where: E_p denotes the expectation over repeated sampling, $\boldsymbol{\pi}$ is the vector of the inclusion probabilities, $\eta_{(dr)k} = (\tilde{y}_{rk} + u_{rk})\gamma_{dk} - \pi_k g_{(dr)k}$, $f = N/(N-H)$, $g_{(dr)k} = \boldsymbol{\delta}'_k \mathbf{A}^{-1} \sum_{k \in U} \boldsymbol{\delta}_k (\tilde{y}_{rk} + u_{rk})\gamma_{dk}(1-\pi_k)$, being $\mathbf{A} = \sum_{k \in U} \boldsymbol{\delta}_k \boldsymbol{\delta}'_k \pi_k (1-\pi_k)$.

6. The MA may be expressed with a ***general expression*** based on ***stable generic terms*** assuming different forms, according to the chosen MA and to the sampling context

$$MA(\hat{t}_{(dr)}) = f\left[\sum_{k \in U} \frac{\omega_{(dr)k}}{\pi_k} - \sum_{k \in U}\left(\varphi_{(dr)k} + \sum_{i=0}^{2} \pi_k^i \, C_{i(dr)k}(\boldsymbol{\pi})\right)\right] \tag{1.3}$$

The stable generic terms $\omega_{(dr)k}$ and $\varphi_{(dr)k}$ are fixed quantities (which may be known or predicted) and the $C_{i(dr)}(\boldsymbol{\pi})$ (i=0,1,2) are functions of the vector $\boldsymbol{\pi}$. For instance, the stable generic terms in the case of (1.2) are given by

$\omega_{drk} = (\tilde{y}^2_{rk} + \sigma^2_{rk})\gamma_{dk}$, $\varphi_{(dr)k} = \omega_{(dr)k}$,

$C_{0(dr)k}(\boldsymbol{\pi}) = = 2\,\boldsymbol{\delta}'_k \mathbf{A}^{-1}[\tilde{y}_{rk}\gamma_{dk}\mathbf{b}_{\tilde{y}(dr)} + \boldsymbol{\delta}_k \sigma^2_{rk}\gamma_{dk}(1-\pi_k)]$,

$C_{1(dr)k}(\boldsymbol{\pi}) = -[2\boldsymbol{\delta}'_k \mathbf{A}^{-1}[\tilde{y}_{rk}\gamma_{dk}\mathbf{b}_{\tilde{y}(dr)} + \boldsymbol{\delta}_k \sigma^2_{rk}\gamma_{dk}(1-\pi_k)] +$

$+ \boldsymbol{\delta}'_k \mathbf{A}^{-1}[\mathbf{b}_{\tilde{y}(dr)}\mathbf{b}'_{\tilde{y}(dr)} + \sum_{j \in U} \boldsymbol{\delta}_j \boldsymbol{\delta}'_j \gamma_{dj}(1-\pi_j)^2 \sigma^2_{rj}]\mathbf{A}^{-1}\boldsymbol{\delta}_k]$,

$C_{2(dr)k}(\boldsymbol{\pi}) = \boldsymbol{\delta}'_k \mathbf{A}^{-1}[\mathbf{b}_{\tilde{y}(dr)}\mathbf{b}'_{\tilde{y}(dr)} + \sum_{j \in U} \boldsymbol{\delta}_j \boldsymbol{\delta}'_j \gamma_{dj}(1-\pi_j)^2 \sigma^2_{rj}]\mathbf{A}^{-1}\boldsymbol{\delta}_k$,

being $\mathbf{b}_{\tilde{y}(dr)} = \sum_{k \in U} \boldsymbol{\delta}_k \tilde{y}_{rk} \gamma_{dk}(1-\pi_k)$.

The expression (1.3) is suitable for an automated spreadsheet of the algorithm (see below) defining the optimal inclusion probabilities.

7. The *inclusion probabilities* are defined as a solution of the following optimization problem which guarantees lower sampling errors of the domain estimates

$$\begin{cases} Min(\sum_{k \in U} \pi_k c_k) \\ MA(\hat{t}_{(dr)}) \le \overline{V}_{(dr)} \quad (d = 1,...,D; r = 1,...,R) \\ 0 < \pi_k \le 1 \qquad\qquad (k = 1,...,N) \end{cases} \tag{1.4}$$

where: $MA(\hat{t}_{(dr)})$ is defined according to (1.3), $\overline{V}_{(dr)}$ is a fixed quantity which defines the threshold of the measure of accuracy of the estimate $\hat{t}_{(dr)}$ and c_k is the cost for collecting information from the unit k. The dominant term in the formula (1.3), is $\sum_{k \in U} \omega_{(dr)k} / \pi_k$ while the other addenda give a minor contribution. The algorithm for solving the problem (1.4) consists of three nested calculation loops. The outer loop fixes the values of the functions $C_{i(dr)k}(\boldsymbol{\pi})$. The inner loop defines the π_k^i values which appear as multiplying factor of the functions $C_{i(dr)k}(\boldsymbol{\pi})$ and then the innermost loop is a modified Chromy algorithm (Falorsi and Righi; 2008) which finds the solution to the minimum constrained problem (1.4) for given values of $\sum_{i=0}^{2} \pi_k^i \; C_{i(dr)k}(\boldsymbol{\pi})$.

2. Some examples

As a general rule, in order to define the *optimal inclusion probabilities* for a given sampling strategy, the following operations have to be done:

1. Define the DIs and the related PDs.
2. Define the estimator. The HT estimator is considered in the above section; the generalized regression estimator is introduced in section 3.
3. Define the form of the model (1.1) for predicting the unknown y_{rk} values.
4. Define the form of the MA (eg. expression 1.2) and reformulate it according to the general expression (1.3) which is suitable for the automation of the algorithm for finding optimal inclusion probabilities.

The theory here illustrated is developed for single stage sampling; however, the approach could be easily extended to consider the case of multistage sampling designs.
Some examples are given below in order to demonstrate how the proposed sampling design could represent a way to generalize in a unified framework some well-known sampling designs. In the following the anticipated variance is taken into account as measure of the accuracy. Consider first the univariate and single-domain case in which $R = D = 1$.

Example 1. Optimal Stratified sampling
Assume that the PDs define a single partition of the population *U,* so that each PD coincides with a stratum, and suppose that the predicted values of the variable *r* ($r=R=1$) of interest are constant in each stratum with uniform stratum variance, e.g. $\tilde{y}_{rk} = \overline{Y}_{rh}$ and $\sigma_{rk}^2 = \sigma_{rh}^2$ (for $k \in U_h$). In this context the UI defines a Stratified Simple Random Sampling WithOut Replacement (SSRSWOR) design. If the costs c_k are uniform in each planned domain, that is $c_k = c_h$ for $k \in U_h$, then the stratum sample sizes are computed according to the optimal allocation (Cochran, 1977, section 5.5) in which $n_h \approx N_h \sigma_{rh} / \sqrt{c_h}$. If the costs c_k are uniform for all the units in the

population, then the well-known Neyman's allocation is realized with $n_h \approx N_h \sigma_{rh}$. Eventually, if the variances are constant over strata, that is $\sigma_{rh} = \overline{\sigma}_r$, then the proportional allocation is implemented, resulting $n_h \approx N_h$.

Example 2. Optimal pps sampling
Assume that there is a single planned domain that coincides with the population *U* and define the stable terms in (1.3) as $\omega_{drk} = \sigma^2_{rk}$, $\varphi_{(dr)k} = \omega_{(dr)k}$, $C_{i(dr)k}(\boldsymbol{\pi}) = 0$ (i=0,1,2). Then, according to the results given in Särndal *et al.* (1992, ch 12), the UI defines optimal inclusion probabilities proportional to the squared roots of the measures of the heteroscedasticity : $\pi_k \approx \sqrt{x_k}$.

Let us consider now the multivariate multi-domain case and suppose that the sampling estimates have to be calculated for the domains of three domain types T_l (l=1, .., 3) each of which defines a partition of the population of U of cardinality D_l being $D = D_1 + D_2 + D_3$. A demonstration of how the sample size of the interest domains may be obtained by different sampling designs is shown below.

Example 3
The standard approach, here denoted as *cross-classified* or *one-way stratified design*, defines the strata by cross-classifying the modalities of the three domain types.
We can obtain the *one-way stratified design* with the UI by assuring that the U_h coincide with the strata of the one-way stratified design. Then: $H = D_1 \times D_2 \times D_3$. The vectors $\boldsymbol{\delta}'_k$ are defined as (0,..,1,...,0) vectors and each U_h can be defined by a specific intersection of the populations of three domains of interest, one for each domain type.
Furthermore if, for every variable *r* of interest, the predicted values are constant in each stratum with uniform stratum variance, e.g. $\tilde{y}_{rk} = \overline{Y}_{rh}$ and $\sigma^2_{rk} = \sigma^2_{rh}$ (for $k \in U_h$), then a SSRSWOR design is implemented. After some algebra the (1.2) becomes

$$AV(\hat{t}_{(dr)} \mid \boldsymbol{\pi}) = f \sum_{h \in \Gamma_d} \sigma^2_{rh} \sum_{k \in U_h} (1/\pi_k - 1) = f \sum_{d=1}^{D} \sum_{h \in \Gamma_d} \sigma^2_{rh} N_h (N_h / n_h - 1),$$

since the terms $\tilde{y}_{rk}$ disappear and $\pi_k = \pi_h$ (for $k \in U_h$).

Example 4
Consider the previous situation in which the U_h coincide with the strata of the one-way stratified design and the predicted values are constant in each stratum. If the model variances are proportional to a known values of some auxiliary variable, eg. $\sigma^2_{rk} = \sigma^2_r v_k$, then a stratified random sampling without replacement with varying inclusion probabilities design is implemented.

Example 5
The PDs U_h are defined combining all the couples of the domains of the domain types; then $H = (D_1 \times D_2) + (D_1 \times D_3) + (D_2 \times D_3)$ and the $\boldsymbol{\delta}'_k$ are defined as vectors with three values equal to one, each in correspondence of one of the three above couples, e.g. (0,...,0,1,0,...,0,1,0,...,0,1,0,...,0).

Example 6
Some PDs U_h agree with the domains of one population partitions, for instance T_1, and the others U_h are defined combining couples of the remaining domain types T_2 and T_3. Then: $H = D_1 + (D_2 \times D_3)$ and the $\boldsymbol{\delta}'_k$ are defined as vectors having two values equal to one, the first in correspondence to the domains of the partition T_1 and the second in correspondence to the couple of the partitions T_2 and T_3, e.g. (0,...,0,1,0,...,0,1,0,...,0).

Example 7
The PDs U_h agree with the domains of interest; then $H = D_1 + D_2 + D_3$ and the $\boldsymbol{\delta}'_k$ are defined as vectors with three ones each in correspondence to one of the three domain types.

The examples 3 and 4 describe *one-way* (or *standard*) *stratified designs*, while the remaining examples (5,6,7) refer to a *multi-way stratified design*. The choice of the sampling design depends on theoretical and operative reasons. From the operative view point the implementation of the one-way stratified design belongs to the current culture of the survey practitioners and it's implementation is uncomplicated, while the *multi-way* design is seldom adopted for defining the sampling strategies of the actual surveys; however these kinds of designs allow to face a lot of empirical contexts in which the traditional approach fails to achieve the target objectives.

3 Remarks on the regression estimator

Consider the case in which for producing the sampling estimates, vectors of auxiliary variables are available for all the population units and suppose that the predictions based on this auxiliary information are those given in model (1.1). In this context, the estimates of interest may be computed with a generalized modified regression estimator, which may be expressed as (Rao, 2003, pag. 20):

$$_{greg}\hat{t}_{(dr)} = \sum_{k \in U} \tilde{y}_{rk}\, \gamma_{dk} + \sum_{k \in s} u_{rk}\, \gamma_{dk} / \pi_k \;\; (r = 1,\ldots,R \;;\; d = 1,\ldots, D). \qquad (3.1)$$

An approximation of the Anticipated Variance of the estimator (3.1) under balanced sampling is

$$AV(_{greg}\hat{t}_{(dr)} \mid \boldsymbol{\pi}) = E_M[f \sum_{k \in U} (1/\pi_k - 1)\; _{greg}\eta^2_{(dr)k}],$$

being $_{greg}\eta_{(dr)k} = u_{rk}\, \gamma_{dk} - \pi_k \boldsymbol{\delta}_k [\sum_{j \in U} \boldsymbol{\delta}_j \boldsymbol{\delta}'_j\, \pi_j (1-\pi_j)]^{-1} \sum_{k \in U} \boldsymbol{\delta}_j\, u_{rj}\, \gamma_{dj}\, (1-\pi_j)$.
The expression of the residuals $_{greg}\eta_{(dr)k}$ is equivalent to the expression $\eta_{(dr)k}$ given

in formula (1.2), except for the substitution of the terms $(\tilde{y}_{rk} + u_{rk})\gamma_{dk}$ with $u_{rk}\gamma_{dk}$. The derivation of the expression of stable generic terms of (1.3) is straightforward.

4 Remarks on nonresponse

Suppose that, for different causes, it is impossible to collect the survey variables from some sample units. Only to make the things simple, let us further hypothesize that: (*i*) the phenomenon of nonresponse is substantially different among the PDs U_h (h=1,...,H); (*ii*) the response propensities, θ_k, are roughly constant for the units belonging to the subpopulation U_h, that is $\theta_k \cong \theta_h$ for $k \in U_h$; (*iii*) when planning the sample design, a quite reliable estimate, say $\tilde{\theta}_h$, of the response propensity of the units belonging to U_h may be obtained from the previous surveys. According to the strategy proposed in Särndal and Lundström (2005, expression 6.4), the estimator of the totals of interest is calculated with the *calibration estimator*:

$$ {}_{cal}\hat{t}_{(dr)} = \sum_{k \in s^*} y_{rk}\gamma_{dk}\,\lambda_k / \hat{\theta}_k \pi_k \qquad (r = 1,\ldots,R\,;\, d=1,\ldots, D), \tag{4.1}$$

where: s^* is the sample of respondents; $\hat{\theta}_k$ is the sample estimate of the response probability; $\lambda_k = 1+[\sum_{j\in U}\boldsymbol{\delta}_j - \sum_{j\in s^*}(\pi_j\hat{\theta}_j)^{-1}\boldsymbol{\delta}_j]'[\sum_{j\in s^*}(\pi_j\hat{\theta}_j)^{-1}\boldsymbol{\delta}_j\boldsymbol{\delta}'_j]^{-1}\boldsymbol{\delta}_k$.

In the context here described, the response probabilities may be estimated by $\hat{\theta}_k = m_h / n_h$ for $k \in s_h^* = s^* \cap U_h$ being m_h the sample size of s_h^*. Let us note that the stratum response probabilities have been introduced with two different symbols, $\tilde{\theta}_h$ and $\hat{\theta}_h$, since the first is an estimate available when planning the sample design, and the latter is estimated from the current survey data. Under the hypothesis that by calibrating in each PD, the nonresponse bias becomes negligible and considering the response phenomenon as a second phase of sampling, then the MA of (4.1) may be computed by (Särndal and Lundström; 2005, pag. 150):

$$MA({}_{cal}\hat{t}_{(dr)}) = AV({}_{cal}\hat{t}_{(dr)} \mid \boldsymbol{\pi}) = AV_{sam} + AV_{NR}$$

in which $AV_{sam} = E_m E_p(\sum_{k\in s} y_{rk}\gamma_{dk}\nu_k / \pi_k \mid \boldsymbol{\pi})$ is the anticipated variance of the calibrated estimator in the absence of nonresponse and $AV_{NR} = E_m E_p V_q(\sum_{k\in s^*} y_{rk}\gamma_{dk}\nu_k / \hat{\theta}_k\pi_k \mid \boldsymbol{\pi})$ represents the additional part of variability due to the phenomenon of non-response, denoting with $V_q(\cdot)$ the variance of (4.1) over different sets of respondents.

Let $e_{rk} = y_{rk} - \boldsymbol{\delta}'_k (\sum_{j \in U} \boldsymbol{\delta}_j \boldsymbol{\delta}'_j)^{-1} \sum_{j \in U} \boldsymbol{\delta}_j y_{rj}$ denote the residual with respect to the regression model in which the variables of interest y_{rk} are regressed with respect to the auxiliary vectors $\boldsymbol{\delta}'_k$ and let ${}_e\sigma^2_{rk}$ indicate the model variance of e_{rk}. By adopting, in the phase of planning the sampling design, the reasonable approximations $(e_{rk}\, \gamma_{dk} - \pi_k \boldsymbol{\delta}_k \mathbf{A}^{-1} \sum_{j \in U} \pi_k\, \boldsymbol{\delta}_j\, \gamma_{dk} e_{rk}\, (1/\pi_k - 1))^2 \cong e^2_{rk}\, \gamma_{dk}$, and $f \cong 1$, the anticipated variance of (4.1) may be approximated by $AV({}_{cal}\hat{t}_{(dr)} \,|\boldsymbol{\pi}) = \sum_{k \in U} 1/\pi_k\, ({}_e\sigma^2_{rk}\, \gamma_{dk} / \tilde{\theta}_k) - \sum_{k \in U} {}_e\sigma^2_{rk}\, \gamma_{dk})$, being $\tilde{\theta}_k = \tilde{\theta}_h$ for $k \in U_h$. Thus, having reliable estimates of the response propensities $\tilde{\theta}_k$ and of the model variances ${}_e\sigma^2_{rk}$, it is possible to define the inclusion probabilities that individuate the minimum cost solution, taking into account the additional part of the variance deriving from the expected non response. After some simple algebra, in this context, the terms of the general expression (1.3) of the MA are given by: $\omega_{(dr)k} = {}_e\sigma^2_{rk}\, \gamma_{dk} / \tilde{\theta}_k$, $\varphi_{(dr)k} = {}_e\sigma^2_{rk}\, \gamma_{dk}$, $C_{i(dr)k}(\boldsymbol{\pi}) = 0$ (for i=0,1,2). Let us note that, if the model variance is constant in each PD h (that is ${}_e\sigma^2_{rk} = {}_e\sigma^2_{rh}$ for $k \in U_h$), then $\pi_k = n_h / N_h$ and then the MA may be reformulated according to the sound expression (Särndal and Lundström; 2005, pag. 171-172)

$$AV({}_{cal}\hat{t}_{(dr)} \,|\boldsymbol{\pi}) = \sum_{h=1}^{H} N_h (N_h / \tilde{m}_h - 1)\, {}_e\sigma^2_{rk}\,,$$

being $\tilde{m}_h = \tilde{\theta}_h\, n_h$ the expected number of respondents in U_h.

References

1. Cochran, W.G.: Sampling Techniques. Wiley, New York (1977)
2. Deville, J.-C., Tillé, Y.: Efficient balanced sampling: the cube method, Biometrika **91**, 893-912 (2004)
3. Deville, J.-C., Tillé, Y.: Variance approximation under balanced sampling. Journal of Statistical Planning and Inference **128**, 569-591 (2005)
4. Falorsi, P.D., Righi, P.: A balanced sampling approach for multi-way stratification designs for small area estimation. Survey Methodology **34**, 223-234 (2008)
5. Isaki, C.T., Fuller, W.A.: Survey design under a regression superpopulation model. Journal of the American Statistical Association **77**, 89-96 (1982)
6. Särndal, C.-E., Swensson, B., Wretman, J.: Model Assisted Survey Sampling. Springer-Verlag, Berlin (1992).
7. Särndal, C.-E., Lundström, S.: Estimation in Surveys with Nonresponse. Wiley, New York (2005)
8. Rao, J.N.K.: Small Area Estimation, Wiley, New York (2003)
9. Righi, P., Falorsi, P.D.: Optimal allocation algorithm for a multi-way stratification design. Proceedings of the Second ITACOSM Conference, 27-29 June 2011, Pisa, 49-52 (2011)