Distr. GENERAL

8 October 2012

WP. 5
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Seminar on New Frontiers for Statistical Data Collection
(Geneva, Switzerland, 31 October-2 November 2012)

Topic (i): New data sources

OPTIMAL SURVEY STRATEGIES IN THE MULTIVARIATE MULTI-DOMAIN CONTEXT WITH MULTIPLE SOURCES OF ADMINISTRATIVE INFORMATION COVERING DIFFERENT POPULATION SUBSETS

Contributed Paper

Prepared by Piero Demetrio Falorsi, Stefano Falorsi and Paolo Righi, Italian National Institute of Statistics (ISTAT)

I. Introduction

1. Large scale surveys in official statistics commonly produce a huge number of estimates related both to different parameters of interest and to highly detailed estimation domains. These domains generally define non-nested partitions of the target population. When the domain indicator variables are available for each sampling unit at the sampling framework level, the survey sampling designer could attempt to select a sample which covers each domain. In so doing, direct estimates can be obtained for each domain and sampling errors at a domain level would be controlled.

2. In the paper we present an unified and general approach for defining an optimal sampling design for one stage sampling design when the domain membership variables are known at the design stage. This case may represent the most frequent situation for establishment surveys and for other survey contexts (e.g. the social surveys or the agricultural surveys) if the domains are of geographical type (e.g.: type of municipality, region, province, etc.). However, the increasing development of data integration among administrative registers and survey frames will increase the applicability of the approach herein presented. Furthermore, we consider the challenging new environment in which different administrative data sources may be linked to some population subsets and thus the sampling frame may be partitioned according to the different number of auxiliary variables, deriving from several administrative and statistical sources.

3. Our paper faces this complex and challenging situation and propose a coherent a general survey strategy which allows to face efficiently this new context and allows to fully exploit the use of ads.

4. The approach is easy to implement and flexible covering as particular cases most both of the optimal solutions described in literature and the actual one stage sampling designs. The most relevant elements which characterize the generalization framework are:

(a) the balanced sampling design (Deville and Tillé, 2004) which, according to the different definitions of the balancing equations, represents a wide-ranging sampling schema.

(b) The concept of planned domains (see section IV) which are subsets of the estimation domains within which the sampling sizes are defined in advance and allow to define in a simply and natural way either the traditional stratified (one-way) sampling designs (Cochran, 1977) or the multi-way stratified (or incomplete stratified) sampling designs (Jessen, 1970; Lu and Sitter, 2002).

(c) The superpopulation model for predicting the unknown values of the variables of interest allowing to obtain either equal or unequal probability sampling designs, within specific subset of population units.

(d) A general form of regression estimator which allows the use of different functional forms of working superpopulation models linking the survey and the administrative variables. The estimates may be characterized by the following aspects: (a) are unbiased and efficient considering jointly the design and the model; (b) are calibrated on the existing administrative information; and (c) are consistent among different variables and different levels of aggregations. For some functional forms of the working superpopulation models, population synthetic data files of imputed data may be constructed; model and design unbiased estimates may be then obtained by a simple aggregation of the imputed data of the synthetic data files.

5. The balanced sampling, the superpopulation models and the regression estimators are well known in literature; within the framework here proposed, they play an instrumental role and can be usefully adopted as useful tools for the generalizations. The main and the new contribution of the paper is a clear definition of the informative context and the illustration of how well-known statistical tools may be properly adapted for defining a unified and generalized framework. The sampling solution is based on some recent, yet unpublished results (Falorsi and Righi, 2012), related to the algorithm for the definition of the optimal inclusion probabilities that consider the more realistic case in which the variables of interest are not known (and must be properly estimated) and takes into account the fact that the measure of accuracy is an implicit function of the inclusion probabilities.

6. The paper is organized as follows. Section II states the informative context. Section III considers the entire chain of the statistical production process in the context here considered where different administrative data cover subsets of the target population. Section IV describes the proposed (unified and general) survey strategy. A focus on the estimation in developed in section V. Some brief conclusions are eventually given in section VI.

II. Informative context

A. Parameters of interest

7. Let *U* be the reference population of *N* elements and let U_d (*d*=1, …, *D*) be a ***Domain of Interest***, (DI), a generic sub-population of *U* with N_d elements. Let y_{rk} denote the value of the *r*-th (*r* =1, …, *R*) variable of interest in the *k*-th population and let and γ_{dk} denote the DI membership indicator, being $\gamma_{dk} = 1$ if $k \in U_d$ and $\gamma_{dk} = 0$ otherwise. Suppose firstly that the γ_{dk} values are known, and available in the sampling frame, for all units in the population. The parameters of interest are the $Q = D \times R$ domains totals

$$t_{(dr)} = \sum\nolimits_{k \in U} y_{rk}\gamma_{dk} = \sum\nolimits_{k \in U_d} y_{rk} \qquad (r = 1,\dots,R\,;\, d=1,\dots, D). \tag{2.1}$$

8. The expression (2.1) defines a complex *multivariate multi-domain* problem since there are *R* variables (*multivariate* aspect) and *D*>1 domains (*multi-domain* aspect) of interest.

B. Auxiliary information

9. Suppose that *B* (with B>1) administrative data sources are available and that each source covers a particular subset of *population register* which, for all practical purposes, identifies operationally the population *U*. Let us further assume that it is possible to link the units belonging to the *b*-th (*b*=1,…,*B*) administrative data source with those included into the population register and that the linkage may be realized without error; this framework, although simplified, characterizes many actual information contexts (as for instance that of the business surveys) where a unique identifier high quality code (e.g. the Vat code) is available. At the conclusion of the linkage operations, the subpopulation ${}_bU \subseteq U$ (*b*=1,…,*B*), for which the variables from the *b*-th administrative data source are available, may be distinguished in the population register. Therefore, for the generic unit *k* of ${}_bU$, it is possible to create the vector ${}_b\mathbf{x}_k$ of auxiliary variables extracted from the source *b*. The population register (denoted in the following with the suffix *b*=0) may be considered as a particular source covering all the population units; the domain of interest membership indicators, γ_{dk} , may be derived from the variables available in the population register; let us further assume that the γ_{dk} values are observed without measurement error. The subpopulations ${}_bU$ may overlap and so it is possible to partition the population register into subpopulations, ${}_{(a)}U$ (*a*=1,…,*A*) -denoted in the following as ***Population Information Profiles*** (PIPs)- characterized by a diverse amount of auxiliary information. For each unit *k* in the PIP, ${}_{(a)}U$, it is then possible to build up a vector ${}_{(a)}\mathbf{x}_k$ of auxiliary variables, by merging the vectors ${}_b\mathbf{x}_k$ (0,1,…,*B*) of administrative variables available for the unit. In order to illustrate this aspect, consider the example, shown in the schema 1 below, in which *B*=2. In this situation, partitioning *U* into *A*=4 PIPs is possible. For the first PIP, ${}_{(1)}U$, neither of the two administrative sources are available, so for the generic unit *k* belonging to ${}_{(1)}U$, the only auxiliary information available is that in the population register which implies ${}_{(1)}\mathbf{x}_k = {}_0\mathbf{x}_k$. For the second subpopulation, ${}_{(2)}U$, only the first additional administrative data source is available, which involves ${}_{(2)}\mathbf{x}_k = ({}_0\mathbf{x}'_k, {}_1\mathbf{x}'_k)'$. For the third PIP, ${}_{(3)}U$, both the additional administrative data sources are available, entailing ${}_{(3)}\mathbf{x}_k = ({}_0\mathbf{x}'_k, {}_1\mathbf{x}'_k, {}_2\mathbf{x}'_k)'$. Finally, for the fourth PIP, ${}_{(4)}U$, only the second administrative source is available, implying ${}_{(4)}\mathbf{x}_k = ({}_1\mathbf{x}'_k, {}_2\mathbf{x}'_k)$.

Schema 1. Example of the partition of the population register into *Population Profiles* with two administrative data sources(*)

	Auxiliary variable								**Subpopulation** $_{(a)}U$
	Register Variables Source=0 $_0\mathbf{x}_k$	Additional data sources							
		Source b=1 $_1\mathbf{x}_k$			Source b=2 $_2\mathbf{x}_k$				
Units	x								$_{(1)}U$: no additional administrative source available. $_{(1)}\mathbf{x}_k = {}_0\mathbf{x}_k$
	x								
	x	x	x	x					$_{(2)}U$: only the administ rative source 1 is available. $_{(2)}\mathbf{x}_k = ({}_0\mathbf{x}'_k, {}_1\mathbf{x}'_k)'$
	x	x	x	x					
	x	x	x	x					
	x	x	x	x	x	x	x	x	$_{(3)}U$ both administrative sources 1 and 2 are available. $_{(3)}\mathbf{x}_k = ({}_0\mathbf{x}'_k, {}_1\mathbf{x}'_k, {}_2\mathbf{x}'_k)'$
	x	x	x	x	x	x	x	x	
	x				x	x	x	x	$_{(4)}U$: only the administrative source 2 is available. $_{(4)}\mathbf{x}_k = ({}_1\mathbf{x}'_k , {}_2\mathbf{x}'_k)$
	x				x	x	x	x	
	x				x	x	x	x	
	x				x	x	x	x	

(*) An x in a given cell denotes that the variable (in *column*) is available for the unit (in *row*)

Note that each DI may be obtained by the *union* of the different *subpopulation intersections* $_{(a)}U_d = U_d \cap {}_{(a)}U$. For each intersection the vector of totals of auxiliary variables

$$_{(a)}\mathbf{t}_{dx} = \sum\nolimits_{k \in {}_{(a)}U_d} {}_{(a)}\mathbf{x}_k$$

may be considered as known.

C. The working model

10. The following *working* superpopulation model may be defined for the units belonging to $_{(a)}U$ (a=1,…,A):

$$\begin{cases} E_M(y_{rk} \mid {}_{(a)}\mathbf{x}_k) = {}_{(a)}f({}_{(a)}\mathbf{x}_k, {}_{(a)}\boldsymbol{\theta}_r) \equiv {}_{(a)}\tilde{y}_{rk} \\ E_M({}_{(a)}u_{rk} = y_{rk} - {}_{(a)}\tilde{y}_{rk}) = 0;\ E_M({}_{(a)}u^2_{rk}) = {}_{(a)}\sigma^2_{rk};\ E_M(u_{rk}, u_{rl}) = 0 \quad \forall k \neq l, \end{cases} \quad \text{for} k,l \in {}_{(a)}U \qquad (2.2a)$$

in which $E_M(\cdot)$ denotes the model expectation operator, $_{(a)}f(\cdot)$ is a function depending on the unknown vector $_{(a)}\boldsymbol{\theta}_r = ({}_{(a)}\theta_{1r},\ldots,{}_{(a)}\theta_{ir},\ldots,{}_{(a)}\theta_{{}_{(a)}Gr})'$ of $_{(a)}G$ parameters, $_{(a)}u_{rk}$ is the error term; furthermore we assume

$$_{(a)}\sigma^2_{rk} = {}_{(a)}\sigma^2_r \ {}_{(a)}v_k^{{}^{(a)}\tau} \qquad (2.2b)$$

where $_{(a)}v_k$ is a *known* auxiliary variable which in general may be expressed as function of the vector $_{(a)}\mathbf{x}_k$, $_{(a)}\sigma_r$ and $_{(a)}\tau$ are scalar parameters.

11. We note that the model (2.2) could assume diverse functional ${}_{(a)}f\,({}_{(a)}\mathbf{x}_k\,,{}_{(a)}\boldsymbol{\theta}_r)$ forms in the different PIPs. Furthermore, we remark that in some cases, a given variable of interest, say the $\ddot{r}-\mathrm{th}$, could coincide with an element, say ${}_{(a)}x_{gk}$, of the vector ${}_{(a)}\mathbf{x}_k$. In this situation, the following relation holds

$$y_{\ddot{r}k} = {}_{(a)}\tilde{y}_{\ddot{r}k} = {}_{(a)}x_{gk} \text{ and } {}_{(a)}u_{\ddot{r}k} = 0 . \qquad (2.2c)$$

III. An overview of the overall statistical process

12. The overall statistical process which starts from the collection of raw survey data and finishes with the publication of the estimates $\hat{t}_{(dr)}$ of the target totals $t_{(dr)}$ could be roughly represented as the *chain* of the following process steps:

(a) pre-processing

(b) sampling design and selection

(c) data collection

(d) throughput

(e) estimation.

13. In this paper, the attention is mainly paid to the sampling and estimation steps, respectively illustrated in sections IV and V below; however, it is worthwhile to note that the availability of administrative data sources represents a strong input for revisiting all the production process steps, in order to define a coherent and self-consistent strategy aiming at a fully exploit the informative context above described. Some remarks will be developed below.

A. Pre-processing

14. The *frame* illustrated in schema 1, in which the ads are linked to the population register, has to be built up in this phase. In most of the actual survey contexts, this job could be successfully done only involving in a joint work people with different competencies: statistical methods, thematic knowledge and expertise on the specific contents of ads.

15. Theoretically, the frame should be built specifically for each survey; but, nowadays most of the statistical organizations are facing budget cuts and thus a more realistic solution would be that of building a multipurpose frame which could be used for different surveys. The frame represents an *enabling infrastructure*. It plays a central role in the whole process chain and it is the result of a complex statistical process which is characterized by the following main actions:

(a) The selection of the ads to take on board for building the frame.

(b) The pre-treatment of the administrative data sources, for (*i*) identifying the statistical units from the administrative ones; correcting inconsistencies in the administrative variables; etc.

(c) The record linkage in which each *reconstructed unit* of a selected ads has to be *linked* to the corresponding unit in the population register.

(d) The choice of the administrative variables to be included in the frame. These must be properly chosen.

16. The second and the third actions represent complex statistical and technical operations often based on statistical methods which are well known and experienced within the statistical organizations. The latter action (and somehow the first) seems to be the most complex; since apart from the methodologists, also people with conceptual and thematic competencies should be involved for its completion: indeed, in the real survey contexts, there are a lot of variables to be considered in each administrative data source. A preliminary step should be that of performing a *quality assessment* of the overall ads and more specifically of some of the administrative variables; this assessment could be useful also for the subsequent process steps (as for instance the data collection) (Costanzo *et al*. 2011; Costanzo, 2012). After the quality assessment, the individuation of the subset of the *core* survey variables which may be considered as the most relevant for the survey objectives should be realized. An administrative variable could be included in the frame either if it coincides with a survey variable or if it is predictive of some of the core variables. The techniques of data mining (Cabena et al. 1997; Dulli et al. 2009) could be usefully adopted for performing both of the above actions (the individuation of the *core survey variables*, and the selection of a predictive subset of administrative variables). For a given survey occasion, the assessment of the *prediction capacity* of an administrative variable with respect to some of the *core* survey variables could be performed either on previous administrative and survey data or studying the relationship with the current administrative data with the current survey data collected with a pilot survey.

B. Sampling

17. The sample is designed and selected in this phase. In section IV, is described in detail a sampling method, based on balanced sampling, practical and easy to implement, which may represent a general and unified approach for *optimal sampling* in the context herein considered and allows to take into account the different information patterns defined by the PIPs. Furthermore, the method allows to take into account in the design phase the complex problem of the total *non-response*; some remarks on this topics are developed in section IV.

C. Data collection

18. The availability of the frame could represent a challenging resource during the data collection phase, especially in the case in which a computer assisted data collection technique (Web, CATI, or CAPI) is implemented. As for instance:

- consider the case where an administrative variable, which coincides with a survey variable, is available of for a given sampling unit; in this case, the survey question could be omitted from the questionnaire which has to be filled out from the unit. We underline that the quality assessment of the overall ads and more specifically of the administrative variables, which coincide with the survey variables, become strictly necessary when a subset of the survey variables could be collected directly from the administrative data source. The implementation of the system, here briefly described, would be strongly facilitated by the development of metadata driven information systems. In the Business surveys context, the recent development of standard, as XBRL (XBRL, 2012), may represent a challenging enhancement.
- The ${}_{(a)}\tilde{y}_{rk}$ model predicted values with the related variance models ${}_{(a)}\sigma^2_{rk}$ could be used for checking the acceptability of the value y_{rk} collected from a respondent unit. Note that at the beginning of the survey operations, the predicted values can only be computed by applying the parameters models fitted on data collected from either a pilot survey or a previous survey. But after some time, the data collected from respondents could be used for fitting new model parameters; in so doing it is important to check carefully the effect of the early respondents since, often, they may represent a segregated sub-populations (Falorsi *et al*. 2005).

D. Throughput

19. A sufficient large sample size for each PIP, ${}_{(a)}U$, could help both in the edit and in the imputation, since the related methods could fully exploit the information deriving from the ads. As for instance, consider the *donor imputation* technique, and suppose that a given unit, say the *k*-th, should be imputed; a reasonable strategy for finding the donor could be the selection of the unit which has the minimum distance, according to some metric, with the auxiliary vector, ${}_{(a)}\mathbf{x}_k$, of the sampling unit which has to be imputed.

E. Estimation

20. In section V an estimation strategy which allows to fully exploit in the estimation the existing administrative information. The estimates are calibrated, with respect to the available auxiliary information for each population sub-set ${}_{(a)}U_d$. Furthermore the estimates are consistent among variables and among different levels of aggregation.

IV. An optimal sampling strategy

A. Sampling design and sampling selection schema

21. A single stage random sample without replacement, *s*, of fixed size *n*, is selected from the population *U*, by a sampling design where the units are included in the sample according to the $\boldsymbol{\pi} = (\pi_1,...,\pi_k,...,\pi_N)'$ vector of inclusion probabilities which is built so as to assure predefined *expected* sample sizes, say n_h (where n_h are **integer numbers**) for given subpopulations U_h (*h*=1, …, *H*) of size N_h , hereinafter denoted as ***planned domains*** (PD), and therefore

$$\sum\nolimits_{k\in U} \pi_k \, \delta_{hk} = \sum\nolimits_{k\in U_h} \pi_k = n_h \qquad (h\text{=}1,...,H) \qquad (4.1)$$

in which δ_{hk} is a PD membership variable available in the sampling frame for all units in the population, being $\delta_{hk} = 1$ if $k \in U_h$ and $\delta_{hk} = 0$, otherwise. Without loss in generality, we assume that the PD are defined either as subsets of the DIs or of the PIPs; so that both the DIs and the PIPs may be obtained as aggregation of entire planned domains. The above assures that the *expected* sample size of a given DI may be obtained as simple sum of the expected sample sizes of the planned domain which are included in it. The same statement holds for a given PIP. It is clear that the planned domains may overlap.

22. The sample selection is realized by the *cube* algorithm (Deville and Tillé, 2004) which selects a random sample such that the below balancing equations are satisfied for any selection

$$\sum\nolimits_{k\in s} \boldsymbol{\delta}_k = \mathbf{n} \qquad (4.2)$$

with $\boldsymbol{\delta}'_k = (\delta_{1k},...,\delta_{Hk})$, being $\mathbf{n} = (n_1,...,n_h,...,n_H)'$ a vector of **integer** numbers containing the sample sizes of the planned domains. The (4.2) assures that in each possible sample selection, the realized sample size for the planned domain U_h is fixed and equal to the *expected one*, n_h. Generally, the balancing is only approximate. In our case the balancing equations (4.2) are always exactly satisfied since the sum of the inclusion probabilities for each planned domain is an integer; the (Deville & Tillé, 2000; Deville and Tillé, 2004; pag. 905 section 7.3).

23. Let us note, that the sample design above illustrated represent a general schema and allows to define in a simply and natural way either the traditional stratified (*one-way*) sampling designs (Cochran, 1977) or the *multi-way* stratified (or incomplete stratified) sampling designs (Jessen, 1970; Lu and Sitter, 2002). The Simple Random Sampling Without Replacement design with fixed sample size (SRSWOR) and the Stratified Simple Random Sampling Without Replacement design with fixed sample sizes in each stratum (SSRSWOR) are special cases of balanced sampling; the proof are given in Deville and Tillé (2004, p.895 and p. 905) and in Deville and Tillé (2005, p. 577). In our framework the SRSWOR design is implemented with H=1 and π_k =n/N; the SSRSWOR is realized when the following two condition hold $\sum_{h=1}^{H} \delta_{hk} = 1$, $\pi_k = n_h / N_h$ for all $k \in U_h$. A Multi-way sampling design may be obtained for instance by defining $U_h = U_d \ \forall h$.

B. Estimation

24. Under the model (2.2), the estimator of the totals of interest may be expressed under the generic form (Lehtonen et al., 2003):

$$_{gen}\hat{t}_{dr} = \sum_{a=1}^{A} \sum_{k \in {}_{(a)}U} {}_{(a)}\hat{\tilde{y}}_{rk} \gamma_{dk} + \sum_{a=1}^{A} \sum_{k \in {}_{(a)}s} \gamma_{dk} \, {}_{(a)}\hat{u}_{rk} / \pi_k \tag{4.3}$$

where: ${}_{(a)}\hat{\tilde{y}}_{rk} = {}_{(a)}f({}_{(a)}\mathbf{x}_k, {}_{(a)}\hat{\boldsymbol{\theta}}_r)$ is the sample estimate of the prediction ${}_{(a)}\tilde{y}_{rk}$; ${}_{(a)}\hat{u}_{rk} = (y_{rk} - {}_{(a)}\hat{\tilde{y}}_{rk})$ represents the sample estimate of the residual ${}_{(a)}\hat{u}_{rk}$; ${}_{(a)}s = s \cap {}_{(a)}U$.

25. The estimator (4.3) is obtained as the sum of two addenda. The first addendum , $\sum_{a=1}^{A} \sum_{k \in {}_{(a)}U} {}_{(a)}\hat{\tilde{y}}_{rk} \gamma_{dk}$, is the *synthetic* component of the estimator and represents its dominant component. The second addendum, $\sum_{a=1}^{A} \sum_{k \in {}_{(a)}s} \gamma_{dk} \, {}_{(a)}\hat{u}_{rk} / \pi_k$, represents the *bias correction term* and is the *minor* part of the estimate being roughly equal to 0. The properties of estimator (4.3) are examined thoroughly in section V and therein are defined the conditions according to which the bias correction term equals zero; here we note that the estimator (4.3) may be computed if the sample size in each subset ${}_{(a)}s$ is sufficient to build the predictions. The domain estimates for the generic DI U_d may be computed also in absence of sample in the domain, since the synthetic component may always be computed. Eventually, we note that the predictions ${}_{(a)}\hat{\tilde{y}}_{rk}$ differ from one model specification to another, depending on the functional form and from the choice of the auxiliary variables.

26. Starting from the results given in Deville and Tillé (2004 and 2005), Falorsi and Righi (2008) propose the following variance approximation of the estimates $_{gen}\hat{t}_{dr}$:

$$\mathrm{E}_p({}_{gen}\hat{t}_{dr} - t_{dr})^2 \cong \dot{V}_p(\hat{t}_{(dr)} \,|\, \boldsymbol{\pi}, \textstyle\sum_{k \in s} \boldsymbol{\delta}_k = \mathbf{n}) = N/(N-H)[\sum_{a=1}^{A} \sum_{k \in {}_{(a)}U} (1/\pi_k - 1) \, {}_{(a)}\eta_{(dr)k}^2] \tag{4.4}$$

where E_p and V_p denote the expectation and the variance over repeated sampling, being

$$ {}_{(a)}\eta_{(dr)k} = {}_{(a)}u_{rk}\,\gamma_{dk} - \pi_k g_{(dr)k} \text{ and } g_{(dr)k} = \boldsymbol{\delta}'_k \mathbf{B}_{(dr)}, \tag{4.4a} $$

being

$$ \mathbf{B}_{(dr)} = \mathbf{A}^{-1} \sum_{a=1}^{A} \sum_{k \in {}_{(a)}U} \pi_k \boldsymbol{\delta}_k \gamma_{dk}\ {}_{(a)}u_{rk}\,(1/\pi_k - 1) $$

and

$$ \mathbf{A} = \sum_{k \in U} \pi_k^2 \boldsymbol{\delta}_k \boldsymbol{\delta}'_k (1/\pi_k - 1) $$

.

27. More recently, an alternative approximation of the sampling variance was considered in Breidt and Chauvet (2011) which is shown to account for the whole variance better than the approximation (4.4) when the balanced equations are not exactly satisfied and the cube algorithm executes the landing phase; but the same authors state that the approximation (4.4) is approximately unbiased in the case, herein considered, in which the balanced equations are exactly satisfied.

28. Starting from (4.4) it is straightforward to build up a plug-in sampling estimate of the sampling variance

$$ \hat{V}_p(\hat{t}_{(dr)} \,|\, \boldsymbol{\pi}, \sum_{k \in s} \boldsymbol{\delta}_k = \mathbf{n}) = N/(N-H)[\sum_{a=1}^{A} \sum_{k \in {}_{(a)}s} (1-\pi_k/\pi_k^2)\ {}_{(a)}\hat{\eta}_{(dr)k}^2], \tag{4.5} $$

in which ${}_{(a)}\hat{\eta}_{(dr)k} = \hat{u}_{rk}\gamma_{dk} - \boldsymbol{\delta}'_k \hat{\mathbf{B}}_{(dr)}$ and $\hat{\mathbf{B}}_{(dr)}$ represents the sampling estimate of the population vector $\mathbf{B}_{(dr)}$.

C. Definition of the optimal inclusion probabilities

29. The *inclusion probabilities vector* should built up so as to achieve the following ***requirements***:

(a) to achieve the minimum cost solution;

(b) to assure a *sufficient* accuracy of the domain estimates;

(c) to assure a *sufficient* accuracy of the model parameters ${}_{(a)}\boldsymbol{\theta}_r$;

(d) to guarantee a fixed sample size for each DI and for each PIP, in a way to compute model assisted unbiased and efficient direct estimates.

30. As far as concerns the specific *measure of accuracy*, we underline that at design phase, the accuracy may be measured by the *Anticipated Variance* (AV, Isaki and Fuller, 1982; Nedyalkova and Tillé, 2008). In our context, the AV is defined as:

$$ AV(\hat{t}_{(dr)} \,|\, \boldsymbol{\pi}, \sum_{k \in s} \boldsymbol{\delta}_k = \mathbf{n}) = E_M E_p((\hat{t}_{(dr)} - t_{(dr)})^2 \,|\, \boldsymbol{\pi}, \sum_{k \in s} \boldsymbol{\delta}_k = \mathbf{n}) $$

$$\cong E_M(\dot{V}_p(\hat{t}_{(dr)} \,|\, \boldsymbol{\pi}, \textstyle\sum_{k\in s} \boldsymbol{\delta}_k = \mathbf{n})) = N/(N-H)\sum_{a=1}^{A}\sum_{k\in {}_{(a)}U}(\frac{1}{\pi_k}-1)E_M({}_{(a)}\eta^2_{(dr)k}) \quad . \qquad (4.6)$$

31. The explicit expression of (4.6) is derived in the appendix A1; looking at the explicit expression, we note that for its calculation in the design phase it is necessary to assume as known the scalars parameters ${}_{(a)}\sigma_r$ and ${}_{(a)}\tau$ (see Appendix A1). In practice the scalar parameters have to be estimated from pilot or previous survey data. The most influential parameter is the scalar σ_r^2; the sample size of the pilot (or previous) survey should be sufficient to guarantee that the estimates of scalars σ_r^2 have a sufficient precision in order to obtain that there is only a limited impact on the overall sample size, *n*, by solving the below problem (4.7) considering either the upper or lower bounds of the confidence intervals of σ_r^2. Finally, we note that in case of perfect model fit: that is $y_{rk} = \tilde{y}_{rk}$, then the anticipated variance coincides with the design variance; so the anticipated variance may be viewed as a general measure of the sampling accuracy.

32. The fundamental results for the derivation of the AV of the estimates ${}_{(a)}\hat{\theta}_{ir}$ are given in appendix A2. Here we note that for some functional expressions ${}_{(a)}f$, the calculus of the AV may necessitate the availability of an estimate, ${}_{(a)}\tilde{\boldsymbol{\theta}}_r$, of ${}_{(a)}\hat{\boldsymbol{\theta}}_r$. In practice these model parameters have to be estimated from pilot or previous survey data; the same considerations developed for the accuracy of σ_r^2 hold in this case.

33. To take into account the previous requirements, the $\boldsymbol{\pi}$ vector values are determined by solving the following problem:

$$\begin{cases} Min(\sum_{k\in U}\pi_k\, c_k) \\ AV({}_{gen}\hat{t}_{(dr)} \,|\, \boldsymbol{\pi}, \sum_{k\in s}\boldsymbol{\delta}_k = \mathbf{n}) \le \overline{V}_{(dr)} \quad (d=1,...,D; r=1,...,R) \\ AV({}_{(a)}\hat{\theta}_{ir} \,|\, \boldsymbol{\pi}, \sum_{k\in s}\boldsymbol{\delta}_k = \mathbf{n}) =\le {}_{(a)}\overline{V}_{ir} \quad (a=1,...,A; r=1,...,R; i=1,...,{}_{(a)}G) \\ 0<\pi_k\le 1 \qquad (k=1,...,N) \end{cases}, \qquad (4.7)$$

where; $\overline{V}_{(dr)}$ is a fixed quantity which defines the anticipated variance threshold of the total of *r*-th variable in the domain U_d of interest; ${}_{(a)}\overline{V}_{ir}$ is a fixed quantity which defines the anticipated variance threshold for the estimate ${}_{(a)}\hat{\theta}_{ir}$.

34. The solution of the problem (4.7) fulfils the expected minimum cost assuring the respect of $(D\times R)+(\sum_{a=1}^{A}{}_{(a)}G\times R)$ constraints. The first set of *D*x*R* constraints in (4.7) are strictly related to the survey objectives and assure that the planned measures of accuracy of the domain sampling estimates are lower than given thresholds which are commonly defined on the basis of the relevance of the survey objectives and taking into account the survey budget. The second set $(\sum_{a=1}^{A}{}_{(a)}G\times R)$ of constraints are of instrumental type and are aiming to assure that in each PIP there is sufficient sample size for allowing the calculus of estimator (4.3), by computing reliable estimates of the superpopulation model parameters. The technical details for the solution of the problem (4.7) are illustrated in Righi and Falorsi (2011) and in Falorsi and Righi (2008; 2011 and 2012); in the latter work the authors, starting from the algorithms developed for the optimal allocation in stratified sampling (Bethel, 1989; Chromy, 1987), find a more general solution that consider the more realistic case in which the variables of interest are not known (and must be

properly estimated) and takes into account the fact that the measure of accuracy is an implicit function of the inclusion probabilities. The recent paper by Chauvet et al. (2011) treats the problem of finding the optimal inclusion probabilities in balanced sampling; the author propose the adoption of the fixed point algorithm; another relevant contribution on the same issue may be found in Tillé and Favre (2005). Both papers do not deal with the case in which the balancing variables depend on the inclusion probabilities and represent only a partial solution to the problem related to the fact that the sampling variance is an implicit function of the inclusion probabilities. However we note that both papers suggest to use the same variance approximation adopted in this manuscript.

Calibration phase

35. After the optimization phase, in which the $\boldsymbol{\pi}$ vector is defined as solution of the problem (4.7), a *calibration phase* is performed (Falorsi and Righi, 2008) which changes as little as possible the *optimal* inclusion probabilities in such a way that summing up on each *planned domain* the calibrated inclusion probabilities gives an integer. The use of the Generalized Iterative Proportional Fitting algorithm (GIPF; Dykstra and Wollan 1987) assures that all the resulting calibrated inclusion probabilities are on the in the (0,1] interval. Although the calibration phase is essential, because it assures that the balanced equations are exactly satisfied (see section IV.A), in this paper it is not described since it actually represents a minor part of the overall strategy.

V. Further notes on estimation

A. Estimating the superpopulation parameters

36. On a sample basis, the estimates of the parameters ${}_{(a)}\hat{\boldsymbol{\theta}}_r$ may be computed by solving the following system of estimating equations (Kim and Rao, 2012):

$$\sum\nolimits_{k\in {}_{(a)}s}(y_{rk}-{}_{(a)}f({}_{(a)}\mathbf{x}_k,{}_{(a)}\hat{\boldsymbol{\theta}}_r))\,{}_{(a)}\mathbf{h}_{rk}/\pi_k=\mathbf{0} \quad , \qquad (5.1)$$

where in which

${}_{(a)}\mathbf{h}_{rk}=(\delta_{(a)}f({}_{(a)}\mathbf{x}_k,{}_{(a)}\boldsymbol{\theta}_r)/\delta\,{}_{(a)}\boldsymbol{\theta}_r)/{}_{(a)}v_k^{(a)\tau}=({}_{(a)}h_{1rk},\ldots,{}_{(a)}h_{irk},\ldots,{}_{(a)}h_{{}_{(a)}Grk})'$ and $\mathbf{0}$ is a column vector of zeroes. The resulting estimates are *optimal* in the sense they are *design unbiased and the joint mean squared error with respect to the sampling design and the model is minimal* (Godambe and Thompson, 1986; 2010).

B. Estimator in a weighted form

37. For a lot of functional forms, ${}_{(a)}f$, the estimator ${}_{gen}\hat{t}_{dr}$ may be expressed as an usually sample weighted estimator:

$${}_{gen}\hat{t}_{(dr)}=\sum\nolimits_{a=1}^{A}\sum\nolimits_{k\in {}_{(a)}s} y_{rk}\,w_{dk} \qquad (r=1,\ldots,R\,;\; d=1,\ldots,D), \qquad (5.2)$$

where the weights w_{dk} are *domain dependent* and can be obtained by solving the following calibration problem (Singh and Mohl, 1996)

$$\begin{cases} Min(\sum_{k\in {}_{(a)}s} D(\gamma_{dk}/\pi_k, w_{dk}) \\ \sum_{k\in {}_{(a)}s} w_{dk}\, {}_{(a)}\mathbf{x}_k = {}_{(a)}\mathbf{t}_{dx} \end{cases} \qquad (5.3)$$

being, $D(\gamma_{dk}/\pi_k, w_{dk})$ a distance function between the *direct weights*, γ_{dk}/π_k, and the *final weights* w_{dk}.

38. The weights corresponding to a linear relationship

$${}_{(a)}f\left({}_{(a)}\mathbf{x}_k, {}_{(a)}\boldsymbol{\theta}_r\right) \equiv {}_{(a)}\tilde{y}_{rk} = {}_{(a)}\mathbf{x}'_k\, {}_{(a)}\boldsymbol{\theta}_r \qquad (5.4)$$

may be obtained using the chi-squared distance $D(\gamma_{dk}/\pi_k, w_{dk}) = {}_{(a)}v_k^{(a)\tau}\,\pi_k\,[(\gamma_{dk}/\pi_k) - w_{dk}]^2$. In this case the well-known *mgreg* estimator is obtained (Rao, 2002, pag.20; Särndal, Swensson,Wretman, 1992) and the explicit form of the weight is given by:

$$w_{dk} = {}_{(a)}\kappa_{dk}/\pi_k \quad \text{for } k \in {}_{(a)}U$$

$${}_{(a)}\kappa_{dk} = \gamma_{dk} + (\sum_{j\in {}_{(a)}U_d} {}_{(a)}\mathbf{x}_j \gamma_{dk} - \sum_{j\in {}_{(a)}s_d} \gamma_{dk(a)}\mathbf{x}_j/\pi_j)'$$

$$(\sum_{j\in_{(a)}s} {}_{(a)}\mathbf{x}_j\, {}_{(a)}\mathbf{x}'_j/\pi_j\, {}_{(a)}v_j^{(a)\tau})^{-1}\, {}_{(a)}\mathbf{x}_k / {}_{(a)}v_j^{(a)\tau}\,.$$

39. Other forms of distance may be defined (Singh and Mohl, 1996) allowing to bound weights in given ranges $L\gamma_{dk}/\pi_k < w_{dk} < U\gamma_{dk}/\pi_k$, being *L* and *U* respectively the lower and upper bounds of the factors correcting the weights. The weights could also be defined on the basis of the ridge regression techniques (Chatterjee, and Hadi, 2006), smoothing the influence of extreme influential weighted values.

C. Conditions for the construction of a *synthetic* file of imputed data having good statistical quality

40. Following Kim and Rao (2011), here below we examine the conditions under which (*i*) the bias correction term of estimator (4.3) is equal 0 or (*ii*) the asymptotic bias of the *synthetic part* of the estimator (4.3), ${}_{gen,syn}\hat{t}_{dr} = \sum_{a=1}^{A}\sum_{k\in {}_{(a)}U} {}_{(a)}\hat{\tilde{y}}_{rk}\gamma_{dk}$, is negligible.

41. The condition (*i*) is fulfilled if

$$\sum_{a=1}^{A}\sum_{k\in {}_{(a)}s} \gamma_{dk}\, {}_{(a)}\hat{u}_{rk}/\pi_k = 0\,. \qquad (5.5)$$

42. The above is essentially the congeniality condition of Meng (1994) used in the context of multiple imputation. The (5.5) is respected if and only if exists a vector $\boldsymbol{\alpha}$ for which $\gamma_{dk} = \boldsymbol{\alpha}'\, {}_{(a)}\mathbf{x}_k$. In the special cases of linear or logistic augmented regression working models, the (5.5) is respected if the vector of the DI indicators γ_{dk} is in the column space of the matrix ${}_{(a)}\mathbf{X} = \{{}_{(a)}\mathbf{x}_k \;\; k = 1,\ldots {}_{(a)}N\}$. In the case etheroscedastic linear or logistic augmented regression working models, the above condition may be reformulated as $\gamma_{dk}\, {}_{(a)}v_j^{(a)\tau} = \boldsymbol{\alpha}'\, {}_{(a)}\mathbf{x}_k$.

43. Turning now to the case (*ii*) the asymptotic *relative bias* of the synthetic part of estimator (4.3), is

$$RB({}_{gen,syn}\hat{t}_{dr}) = -\frac{\operatorname{cov}(\gamma_{dk}, {}_{(a)}u_{rk})}{(\sum_{k=1}^{N}\gamma_{dk}/N)(t_{dr}/N_d)}, \tag{5.6}$$

where $\operatorname{cov}(\gamma_{dk}, {}_{(a)}u_{rk})$ is the population covariance among the DI membership indicators γ_{dk} and the residuals ${}_{(a)}u_{rk}$ (for $k=1,...,{}_{(a)}N; a=1,...,A)$. It follows from (5.6) that the relative bias of the synthetic part of estimator (4.3) is negligible if the γ_{dk} indicators are approximately unrelated to the model residuals ${}_{(a)}u_{rk}$. This will be the case if the working model (2.2) is correctly specified.

44. The expressions (5.5) and (5.6) clarify the theoretical conditions which justify the necessity of planning the sample size for each DI and for each PIP; furthermore they define the conditions for the constructions of *synthetic files* of imputed data ${}_{(a)}\hat{\tilde{y}}_{rk}$ for each unit in the population. The estimates of interest may be obtained as a simple aggregation over the population of imputed data. These estimates are unbiased (or nearly unbiased) either or if the conditions (5.5) is respected or the relative bias $RB({}_{gen}\hat{t}_{dr})$, as expressed by (5.6) is near 0.

D. Synthesis of the properties of the estimator

45. The main properties of the ${}_{gen}\hat{t}_{dr}$ estimator are listed below.

(a) The estimates are *design* and *model* unbiased.

(b) The estimates are *efficient* with respect to both the design and the model (see section V.A).

(c) The estimates are calibrated for each subset ${}_{(a)}U_d$. Thus, the sample estimates of the total auxiliary variables ${}_{(a)}\mathbf{x}_k$ reproduces the total known ${}_{(a)}\mathbf{t}_{dx}$ at domain level. Thus, for each intersection ${}_{(a)}U_d$, if one auxiliary variable coincides with a variable of interest, its estimate coincides with the known total.

(d) If the same functional forms ${}_{(a)}f$ are used for the estimation of the different *R* variables, then the estimates of the aggregates of the different variables are *consistent* in the sense they respect the relationship existing among the variables at unit level. This may be easily understood by considering the estimator in its sample weighted form (see section V.B). This result is essential for the construction of *synthetic data file* (see section V.C).

(e) The estimates are consistent at the different levels of aggregation. The sum of the estimates ${}_{gen}\hat{t}_{(hr)}$ over subsets, U_h, which represent a partition of a DI U_d, coincides with the estimate defined at the DI level: $\sum_{h \in d} {}_{gen}\hat{t}_{(hr)} = {}_{gen}\hat{t}_{(dr)}$. Therefore, the estimates are consistent at *population* level, so to say that the sum of the domain estimates which represent a partition of the population *U always* reproduces the same estimate of the total referred to the population *U*. This result is relevant if the publication of the survey results is based on a corporate data ware-house.

VI. Concluding remarks

46. The paper discusses a survey strategy, based both on balanced sampling and on a generalized form of regression estimator, which may represent a general and unified approach for defining an optimal survey strategy 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 which define two or more partitions of the target population; furthermore many administrative data sources may be linked to the population register defining population subsets with different information patterns. This is the common situation for large scale surveys in official statistics which normally produce a huge number of estimates related both to different parameters of interest and to highly detailed estimation domains.

47. The main contributions of the paper focus on (*i*) a clear definition of the informative context; (*ii*) the definition of the optimal inclusion probabilities; (*iii*) the proposal of an estimation technique that fully exploit the existing administrative data sources. The algorithm implements the allocation for a general multi-way sampling design in which the standard approach (one-way stratification) is a special case. Moreover, the allocation is multi-domain and multivariate: either the costs or the sample size is minimized guaranteeing that the sampling variances of the target estimates of several variables related to the different planned domains be lower than prefixed level of accuracy thresholds. The estimation method has good qualities with respect both to the *model* and to the *design*; it is defined in a generalized and unified framework which allows to easily incorporate different form of the working superpopulation models linking the variables of interest with the auxiliary variables deriving from administrative data sources. Furthermore, the estimates are calibrated with respect to all the existing administrative data sources and the estimates are consistent among both different variables and different levels of aggregation. In order to investigate the empirical properties of the proposed sampling strategy, some experiments (not illustrated herein, see Falorsi and Righi, 2012) have been carried out on real and simulated data sets. All the experiments produced coherent and satisfactory results.

References

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

Breidt, F.J., and Chauvet, G. (2011). Improved Variance Estimation for Balanced Samples Drawn via the Cube Method. *Journal of Statistical Planning and Inference*, 141, 479–487.

Cabena P.; Hadjinian P.; Stadler R.; Verhees P.; Zanasi P., (1997), *Discovering data mining from concept to implementation*, Prentice Hall PTR 1997.

Chatterjee S.and Hadi A. (2006), *Regression Analysis by Example*, Wiley, New York.

Chauvet, G., Bonnéry, D., and Deville, J.-C. (2011). Optimal Inclusion Probabilities for Balanced Sampling. *Journal of Statistical Planning and Inference*, 141, 984–994.

Chromy J. (1987). Design Optimization with Multiple Objectives, *Proceedings of the Survey Research Methods Section. American Statistical Association*, 194-199.

Cochran W.G. (1977). *Sampling Techniques*. Wiley. New York.

Costanzo L., Di Bella G., Leonardi P.;Vignola M., (2011); Main findings of the Information Collection on the Use of Admin Data for Business Statistics in EU and EFTA Countries; *http://essnet.admindata.eu/Document/GetFile?objectId=5291*.

Costanzo L., (2012); Legal barriers, quality issues: what really hampers a wider use of administrative data in business statistics?; *http://www.q2012.gr/articlefiles/sessions/20.1_ESSnet%20admin%20data_Costanzo.pdf*.

Dykstra R., Wollan P. (1987). Finding I-Projections Subject to a Finite Set of Linear Inequality Constraints, Applied Statistics, 36, 377-383.

Dulli S.; Furini S.; Peron E. (2009), *Data Mining*, Springer Verlag, 2009.

Deville J.-C., Tillé Y. (2004) Efficient Balanced Sampling: the Cube Method, *Biometrika*, 91, 893-912.

Deville J.-C., Tillé Y. (2005) Variance approximation under balanced sampling, *Journal of Statistical Planning and Inference*, 128, 569-591.

Falorsi P.D., Alleva G., Bacchini F., Iannaccone R., (2005), Estimates based on preliminary data from a specific subsample and from respondents not included in the subsample, *Statistical Methods and Application* , Number 1. No 1, 2005.

Falorsi P. D., Righi P. (2008) A Balanced Sampling Approach for Multi-way Stratification Designs for Small Area Estimation, *Survey Methodology*, 34, 223-234.

Falorsi P. D., Righi P. (2012) A Balanced Sampling Approach for Multi-way Stratification Designs for Small Area Estimation, *Paper submitted to the Journal of Official Statistics for pubblication*.

Godambe V. P. and Thompson M.E. (1986), Parameters of Superpopulation and Survey Population: Their Relationships and Estimation. *International Statistical Review*, **54**, pp.127-138.

Godambe V. P. and Thompson M.E. (2009), Estimating Functions and Survey Sampling. *Handbook of Statistics* **29B** (Edited by Pfefferman D. and Rao.) pp.83-101.

Isaki C.T., Fuller W.A. (1982) Survey design under a regression superpopulation model, *Journal of the American Statistical Association*, 77, 89-96.

Jessen R. J. (1970). Probability Sampling with Marginal Constraints, *Journal American Statistical Society*, 65: 776-795.

Kim J. K.and Rao J. N.K. (2012), Combining data from two independent surveys: a model assisted approach. Biometrika, **99**, 1, pp. 85-100.

Lehtonen R., Särndal C. E, Veijanen A. (2003). The effect of Model Choice in Estimation for Domains, Including Small Domains, *Survey Methodology*, 1: 33-44.

Lu W., Sitter R. R. (2002). Multi-way Stratification by Linear Programming Made Practical, *Survey Methodology*, 2: 199-207.

Nedyalkova, D., and Tillé, Y. (2008). Optimal sampling and estimation strategies under the linear model. *Biometrika*, 95, 521-537.

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

Righi P., Falorsi P. D., (2011) Optimal Allocation Algorithm for a Multi-Way Stratification Design, *Proceedings of the Second ITACOSM Conference*, 27-29 June 2011, Pisa, 49-52.

Särndal, C.E., Swensson, B., Wretman, J., (1992). *Model Assisted Survey Sampling*, Springer-Verlag.

Särndal, C.E., Lundström, S. (2005). *Estimation in Surveys with Nonresponse*, Wiley, New York.

Singh A., Mohl C. (1996), Understanding calibration estimators in survey sampling, *Survey Methodology*, vol. **22**. N.2

Tillé, Y. and Favre, A.-C. (2005). Optimal Allocation in Balanced Sampling. *Statistics and Probability Letters*, 74, 31–37.

XBRL (2012); http://www.xbrl.org/.

Appendices

A. Av of the estimate ${}_{gen}\hat{t}_{(dr)}$

1. Let us derive the model expectation of each of the three terms involved in the squared value of (4.4a). We have $E_M\,({}_{(a)}u_{rk})^2\,\gamma_{dk} = ({}_{(a)}\sigma_{rk}^2)\gamma_{dk}$. Consider now the term $E_M\,(\pi_k^2\,g_{(dr)k}^2)$. Since $E_M\,({}_{(a)}u_{jk}\,{}_{(a')}u_{lk}) = 0,$ we have

$$E_M\,(\pi_k^2\,g_{(dr)k}^2) = \pi_k^2\boldsymbol{\delta}'_k\,\mathbf{A}^{-1}[\textstyle\sum_{j\in U}\,\boldsymbol{\delta}_j\boldsymbol{\delta}'_j\;{}_{(a)}\sigma_{rj}^2\;\gamma_{dj}(1-\pi_j)^2]\mathbf{A}^{-1}\boldsymbol{\delta}_k\;.$$

Finally, let us analyse the third element.

Being $E_M\,({}_{(a)}u_{rk}\sum_{a=1}^{A}\sum_{k\in{}_{(A)}U}\,\boldsymbol{\delta}_k\;{}_{(a)}u_{rk}\;\gamma_{dk}(1-\pi_k)) = \boldsymbol{\delta}_k\;{}_{(a)}\sigma_{rk}^2\;\gamma_{dk}(1-\pi_k)$,

then $E_M\,(2\pi_k\;{}_{(a)}u_{rk}\gamma_{dk}\,g_{(dr)k}) = 2\pi_k\boldsymbol{\delta}'_k\,\mathbf{A}^{-1}[\boldsymbol{\delta}_k\;{}_{(a)}\sigma_{rk}^2\gamma_{dk}\,(1-\pi_k)]$.

2. Therefore Taking into account the model expectations (2.2), in our context the AV of the estimates ${}_{gen}\hat{t}_{(dr)}$ may be defined as:

$$AV({}_{gen}\hat{t}_{(dr)}\,|\boldsymbol{\pi},\textstyle\sum_{k\in s}\boldsymbol{\delta}_k=\mathbf{n})=f\left[\sum_{k\in U}\frac{\omega_{(dr)k}}{\pi_k}-\sum_{k\in U}\left(\varphi_{(dr)k}+\sum_{i=0}^{2}\pi_k^i\;C_{i(dr)k}(\boldsymbol{\pi})\right)\right], \qquad \text{(A1.1a)}$$

where:

$$\begin{aligned}
&\omega_{drk}=(\textstyle\sum_{a=1}^{A}{}_{(a)}\lambda_k\;{}_{(a)}\sigma_r^2\;{}_{(a)}v_k^{(a)^{\tau}})\gamma_{dk}\,,\ \varphi_{(dr)k}=\omega_{(dr)k}\,,\\
&C_{0(dr)k}(\boldsymbol{\pi})=\ 2\,\boldsymbol{\delta}_k'\,\mathbf{A}^{-1}[(\textstyle\sum_{a=1}^{A}{}_{(a)}\lambda_k\;{}_{(a)}\sigma_r^2\;{}_{(a)}v_k^{(a)^{\tau}})\boldsymbol{\delta}_k\,\gamma_{dk}(1-\pi_k)]\,,\\
&C_{1(dr)k}(\boldsymbol{\pi})=-[2\,\boldsymbol{\delta}_k'\,\mathbf{A}^{-1}[(\textstyle\sum_{a=1}^{A}{}_{(a)}\lambda_k\;{}_{(a)}\sigma_r^2\;{}_{(a)}v_k^{(a)^{\tau}})\boldsymbol{\delta}_k\,\gamma_{dk}(1-\pi_k)]+\\
&\qquad\qquad+\boldsymbol{\delta}_k'\,\mathbf{A}^{-1}[\textstyle\sum_{j\in U}\boldsymbol{\delta}_j\,\boldsymbol{\delta}_j'\,\gamma_{dj}(1-\pi_j)^2(\textstyle\sum_{a=1}^{A}{}_{(a)}\lambda_j\;{}_{(a)}\sigma_r^2\;{}_{(a)}v_j^{(a)^{\tau}})\,]\mathbf{A}^{-1}\boldsymbol{\delta}_k],\\
&C_{2(dr)k}(\boldsymbol{\pi})=\ \boldsymbol{\delta}_k'\,\mathbf{A}^{-1}[\textstyle\sum_{j\in U}\boldsymbol{\delta}_j\,\boldsymbol{\delta}_j'\,\gamma_{dj}(1-\pi_j)^2(\textstyle\sum_{a=1}^{A}{}_{(a)}\lambda_j\;{}_{(a)}\sigma_r^2\;{}_{(a)}v_j^{(a)^{\tau}})\,]\mathbf{A}^{-1}\boldsymbol{\delta}_k\,,
\end{aligned} \qquad \text{(A1.1b)}$$

being ${}_{(a)}\lambda_k=1$ if $k\in{}_{(a)}U$ and ${}_{(a)}\lambda_k=0$ otherwise.

B. Av of the estimates of the superpopulation model parameters

3. On a Census basis, the super-population parameter vector ${}_{(a)}\boldsymbol{\theta}_r$ may be estimated on the population obtaining the population parameter, ${}_{(a)}\ddot{\boldsymbol{\theta}}_r$, by solving the following homogeneous system of estimating equations (Godambe and Thompson, 2009; Kim and Rao, 2011) :

$${}_{(a)}\mathbf{D}_r({}_{(a)}\ddot{\boldsymbol{\theta}}_r)=\sum_{k\in{}_{(a)}U}{}_{(a)}\mathbf{d}_{rk}=\sum_{k\in{}_{(a)}U}(y_{rk}-{}_{(a)}f({}_{(a)}\mathbf{x}_k,{}_{(a)}\ddot{\boldsymbol{\theta}}_r))\,{}_{(a)}\mathbf{h}_{rk}=\mathbf{0} \qquad \text{(A2.1)}$$

4. On a sample basis, the estimates may be computed by solving the following system:

$${}_{(a)}\hat{\mathbf{D}}_r({}_{(a)}\hat{\boldsymbol{\theta}}_r)=\sum_{k\in{}_{(a)}s}{}_{(a)}\mathbf{d}_{rk}/\pi_k=\sum_{k\in{}_{(a)}s}(y_{rk}-{}_{(a)}f({}_{(a)}\mathbf{x}_k,{}_{(a)}\hat{\boldsymbol{\theta}}_r))\,{}_{(a)}\mathbf{h}_{rk}/\pi_k=\mathbf{0}\,. \qquad \text{(A2.2)}$$

5. Using the first order terms of the Taylor series linearization technique, it is then possible to derive the *sandwich* sampling variance of the estimates ${}_{(a)}\hat{\boldsymbol{\theta}}_r$:

$$\mathbf{0}={}_{(a)}\hat{\mathbf{D}}_r({}_{(a)}\hat{\boldsymbol{\theta}}_r)\cong{}_{(a)}\hat{\mathbf{D}}_r({}_{(a)}\ddot{\boldsymbol{\theta}}_r)+{}_{(a)}\mathbf{G}_r\,({}_{(a)}\hat{\boldsymbol{\theta}}_r-{}_{(a)}\ddot{\boldsymbol{\theta}}_r) \qquad \text{(A2.3)}$$

being ${}_{(a)}\mathbf{G}_r=\left[\dfrac{\delta\,{}_{(a)}\hat{\mathbf{D}}_r({}_{(a)}\ddot{\boldsymbol{\theta}}_r)}{\delta\,{}_{(a)}\ddot{\boldsymbol{\theta}}_r}\right]=\displaystyle\sum_{k\in{}_{(a)}s}\frac{\delta(y_{rk}-{}_{(a)}f({}_{(a)}\mathbf{x}_k,{}_{(a)}\ddot{\boldsymbol{\theta}}_r))\,{}_{(a)}\mathbf{h}_{rk}/\pi_k}{\delta\,{}_{(a)}\ddot{\boldsymbol{\theta}}_r}.$

6. For defining the sample design, for all practical purposes the matrix ${}_{(a)}\mathbf{G}_r$ may be approximated by its population value

$$ {}_{(a)}\mathbf{G}_r \cong \sum_{k\in U_{(a)}} \frac{\delta(y_{rk} - {}_{(a)}f({}_{(a)}\mathbf{x}_k, {}_{(a)}\ddot{\boldsymbol{\theta}}_r))\,{}_{(a)}\mathbf{h}_{rk}}{\delta\,{}_{(a)}\ddot{\boldsymbol{\theta}}_r} . $$

7. In the design phase, for some functional forms ${}_{(a)}f$, the calculus of the elements of the above matrix may necessitate the availability of an estimation, ${}_{(a)}\tilde{\boldsymbol{\theta}}_r$, of ${}_{(a)}\ddot{\boldsymbol{\theta}}_r$. This may be obtained from a pilot or from a previous survey. From (A2.3) it is then possible to derive the *sandwich* variance of ${}_{(a)}\hat{\boldsymbol{\theta}}_r$

$$ \mathbf{V}({}_{(a)}\ddot{\boldsymbol{\theta}}_r - {}_{(a)}\hat{\boldsymbol{\theta}}_r \mid \boldsymbol{\pi}, \textstyle\sum_{k\in s} \boldsymbol{\delta}_k = \mathbf{n}) = {}_{(a)}\mathbf{G}_r^{-1}\,\mathbf{V}({}_{(a)}\hat{\mathbf{D}}_r({}_{(a)}\ddot{\boldsymbol{\theta}}_r) \mid \boldsymbol{\pi}, \textstyle\sum_{k\in s} \boldsymbol{\delta}_k = \mathbf{n})\,{}_{(a)}\mathbf{G}_r^{-1} , \qquad \text{(A2.4)} $$

where

$$ \mathbf{V}_p({}_{(a)}\hat{\mathbf{D}}_r({}_{(a)}\ddot{\boldsymbol{\theta}}_r) \mid \boldsymbol{\pi}, \textstyle\sum_{k\in s} \boldsymbol{\delta}_k = \mathbf{n}) = V_p(\textstyle\sum_{k\in {}_{(a)}s} {}_{(a)}u_{rk}\,{}_{(a)}\mathbf{h}_{rk}/\pi_k \mid \boldsymbol{\pi}, \textstyle\sum_{k\in s} \boldsymbol{\delta}_k = \mathbf{n}) . \qquad \text{(A2.5)} $$

Therefore, starting from expressions (A2.4) and (A2.5) it is straightforward to derive the following

$$ V_p({}_{(a)}\hat{\boldsymbol{\theta}}_r \mid \boldsymbol{\pi}, \textstyle\sum_{k\in s} \boldsymbol{\delta}_k = \mathbf{n}) = V(\textstyle\sum_{k\in_{(a)} s} {}_{(a)}\mathbf{G}_r^{-1}\,{}_{(a)}\mathbf{h}_{rk}\,{}_{(a)}u_{rk} \mid \boldsymbol{\pi}, \textstyle\sum_{k\in s} \boldsymbol{\delta}_k = \mathbf{n}) . \qquad \text{(A2.6)} $$

The sampling variance of the estimate ${}_{(a)}\hat{\theta}_{ir}$ is the *i*-th element on the main diagonal of the matrix $V_p({}_{(a)}\hat{\boldsymbol{\theta}}_r \mid \boldsymbol{\pi}, \sum_{k\in s} \boldsymbol{\delta}_k = \mathbf{n})$, and therefore

$$ V_p({}_{(a)}\theta_{ir} \mid \boldsymbol{\pi}, \textstyle\sum_{k\in s} \boldsymbol{\delta}_k = \mathbf{n}) = V(\textstyle\sum_{k\in_{(a)} s} {}_a z_{irk}/\pi_k \mid \boldsymbol{\pi}, \textstyle\sum_{k\in s} \boldsymbol{\delta}_k = \mathbf{n}) \qquad \text{(A2.7)} $$

being

$$ {}_a z_{irk} = {}_{(a)}u_{rk}\,{}_{(a)}\omega_{irk} . \qquad \text{(A2.8)} $$

in which ${}_{(a)}\omega_{irk}$ is the squared root of the *i*-th element of the main diagonal of ${}_{(a)}\mathbf{G}_r^{-1}\,{}_{(a)}\mathbf{h}_{rk}\,{}_{(a)}\mathbf{h}'_{rk}\,{}_{(a)}\mathbf{G}_r^{-1}$. By consequence, the approximated sampling variance of the estimate ${}_{(a)}\hat{\theta}_{ir}$ may be obtained by expressions (4.4a) except for the substitution of the terms ${}_{(a)}u_{rk}\,\gamma_{dk}$ with ${}_a z_{irk}$. The derivation of the anticipated variance is straightforward.