PL02

Optimal Sampling for the Integrated Observation of Different Populations

P. D. Falorsi Istat | Rome | Italy
P. Righi Istat | Rome | Italy

DOI: 10.1481/icasVII.2016.pl2b

ABSTRACT

This work deepens the problem of defining an optimal sampling in the multivariate case where the variables of interest are related to different target populations. In order to get insight to the underlying phenomena, the observation has to be carried out in an integrated way, implying that units of a given population have to be observed jointly with the related units of the other population. Indirect sampling provides a natural framework for this setting since the units belonging to a population that is the object of a given survey can become carriers of information on another statistical population. This problem is studied with respect to the different contexts which characterize the available information in the sampling design phase, ranging from very well organized situations in which the links among the different units are known in the design phase to a situation in which the available information is very poor. Empirical studies on agricultural data of two developing countries are developed. These show that controlling in the design phaseis effective since by not doing so, the errors of the indirectly observed population can become very high. Furthermore, the need of having good models for predicting the unknown variables or the links is stressed.

Keywords: integrated surveys, sample allocation, indirect sampling

PAPER

1. Introduction

The need of observing in an integrated way different statistical populations related to each other is often encountered in survey sampling. The underlying relationships among the populations are regulated by formal rules, contingent dependencies or relationships created for the pursuit of common purposes. For instance, agricultural surveys often refer to statistical units such as rural households, farms and land parcels that are related to each other. The integrated observation of such populations allows to measure global phenomena of the agricultural sector: for example, the education level of a farm holder and the farm size can affect the productivity of land parcels and thus the risk of malnutrition of rural households. The observation of such units in an integrated way can be recommended to get insights into the agricultural system of a country.

Indirect sampling (Lavallée, 2007) provides a natural framework for the estimation of the parameters of two target populations that are related to each other since the units belonging to a population that is the object of a given survey can become carriers of information on another statistical population, through the type of relationship between the entities. Furthermore, indirect sampling is suitable for producing statistics of populations for which there is no sampling frame. In such context, the sampling procedure assumes a population U^A related with the population of interest U^B, and for whichthe sampling frame of U^A is available. Then, a sample is selected from U^A, and using the existing links between the two populations the units of U^B are observed.

This work deepens the problem of sampling allocation when an indirect sampling design is implemented. The allocation problem for the direct sampling setting has been dealt with in several papers. When one target parameter is to be estimated for the overall population, the optimal allocation in stratified sampling can be performed (Cochran, 1977). When more than one target parameter is to be estimated the problem aleads to a compromise allocation method (Khan et al., 2010), with a loss of precision compared to the individual optimal allocations. Several authors have discussed various criteria for obtaining a feasible compromise allocation: see, for example, Kokan and Khan (1967), Chromy (1987), Bethel (1989) and Choudhry et al. (2012).

Falorsi and Righi (2015) provide a general framework for multivariate and multi-domain surveys. This paper offers a further generalization of the framework proposed by Falorsi and Righi (2015) to the case of integrated observation of two or more populations. Different scenarios related to the level of knowledge of the existing links are examined. Section 2 introduces the background and symbols. Sections 3 and 4 illustrate the basic allocation problem and how it is declined in the different informative scenarios.

2. Background

Let s^A, M^A, m^A be a selected sample from U^A without replacement and with fixed sample size, thenumber of units in U^A and the number of units in s^A, respectively. We use π_j^A to represent the inclusion probability of the jth unit in U^A with $\pi_j^A > 0$ and $\sum_{j \in U^A} \pi_j^A = m^A$. We denote with $y_{j,v}$ the value of the vth (v=1,…,V) characteristic on unit j and the total of all $y_{j,v}$'s by Y_v^A. We estimate the total Y_v^A according to the Horvitz-Thompson (HT) estimator, $\hat{Y}_v^A = \sum_{j \in s^A} w_j^A y_{j,v}$, where $w_j^A = (1/\pi_j^A)$. Many practical sampling designs define planned domains that are sub-populations in which the sample sizes are fixed before selecting the sample. Denote by U_h^A (h=1,…,H) the planned domain of size $M_h^A = \sum_{j \in U^A} d_{j(h)}$ where $d_{j(h)} = 1$ if $j \in U_h^A$ and $d_{j(h)} = 0$ otherwise. Let us suppose that the $d_{j(h)}$ values are known, and available in the sampling frame, for all population units. Fixed sizes sampling designs are those satisfying $\sum_{j \in s^A} \mathbf{d}_j = \mathbf{m}^A$,where $\mathbf{d}_j = (d_{j(1)},...,d_{j(h)},...,d_{j(H)})'$ and $\mathbf{m}^A = (m_1^A,...,m_h^A,...,m_H^A)'$ isthe vector of integer numbers defining the sample sizes fixed at the design stage, with $\sum_{j \in U^A} d_{j(h)} \pi_j^A = m_h^A$. In our setting, the planned domains can overlap; therefore, the unit j may have more than one value $d_{j(h)} = 1$ (for $h = 1,..., H$). Several customary fixed size sampling designs may be considered as special cases. A well-known example is the stratified sampling design where strata are the planned domains and the $\mathbf{d}_j$ vector has H-1 elements equal to zero, and one element equal to 1. We suppose that the $M^A \times H$ matrix $\mathbf{D} = (\mathbf{d}_1,..,\mathbf{d}_j,...,\mathbf{d}_{M^A})'$ is non-singular. According to this general sampling design framework, Deville and Tillé (2005) proposed an approximated expression of the variancebased on the Poisson sampling theory for $\hat{Y}_v^A$ given by $V(\hat{Y}_v^A \mid \mathbf{m}^A) \cong \sum_{j \in U^A} [(1/\pi_j^A) - 1] \eta_{j,v}^2$, where

$$\eta_{j,v} = y_{j,v} - \pi_j^A \mathbf{d}_j' \boldsymbol{\beta}_v \quad \text{and} \quad \boldsymbol{\beta}_v = \boldsymbol{\Delta}^{-1} \sum_{l \in U^A} \pi_l^A (1/\pi_l^A - 1) \mathbf{d}_l y_{l,v} \text{ ,with}$$

$\boldsymbol{\Delta} = \sum_{j \in U^A} \mathbf{d}_j \mathbf{d}_j' \pi_j^A (1 - \pi_j^A)$.

We also make use of the notation: M^B, N^B, U_i^B and M_i^B to be the number of units in U^B, thenumber of clusters in U^B, the ith cluster of U^B with $\bigcup_{i=1}^N U_i^B = U^B$ and the number of units in the$i$th cluster U_i^B. We indicate by $y_{ik,r}$ the value of the rth (r =1,…,R) characteristic for the kth unit of the ith cluster of U^B and the total of all $y_{ik,r}$'s by $Y_r^B = \sum_{i=1}^{N^B} \sum_{k=1}^{M_i^B} y_{ik,r}$.

We define $l_{j,ik}$ as an indicator variable of link existence: $l_{j,ik} = 1$ indicates that there is a link between jth unit in U^A and kth unit in U_i^B, while $l_{j,ik} = 0$ indicates otherwise.

Let us suppose that we carry out an indirect sampling process in which if the unit $j \in U^A$ is included in $j \in s^A$, then all the clusters U_i^B, for which $L_{j,i}^B = \sum_{k=1}^{M_i^B} l_{j,ik} > 0$, are observed in the indirect sample of population U^B. Let n^B be the cluster sample size of population U^B obtained after the indirect sampling process. We estimate Y_r^B according to the Horvitz-Thompson estimator based onthe theory of the Generalized Weight Share Method (GWSM):

$$\hat{Y}_r^B = \sum_{i=1}^{n^B} w_i^B y_{i,r} \tag{2.1}$$

where $y_{i,r} = \sum_{k=1}^{M_i^B} y_{ik,r}$ and $w_i^B = \sum_{j \in s^A} w_j^A \tilde{L}_{j,i}^B$, with $\tilde{L}_{j,i}^B = L_{j,i}^B / L_i^B$ and $L_i^B = \sum_{j=1}^{M^A} L_{j,i}^B$.

Theorem in Section 3 of Lavallée (2007) states that (2.1) offers an unbiased estimator for Y_r^B provided all links $l_{j,ik}$ can be correctly identified and $L_i^B > 0$ for all $i \in U^B$. By defining $z_{j,r} = \sum_{i=1}^{N^B} \tilde{L}_{j,i}^B y_{i,r}$, the estimator (2.1) can be expressed as an usual Horvitz-Thompson (HT)

estimator on the z values referring to the U^A population, $\hat{Y}_r^B = \sum_{j \in s^A} w_j^A z_{j,r}$ .Therefore, the variance, $V(\hat{Y}_r^B)$, of $\hat{Y}_r^B$ maybe expressed as the variance of the HT estimator on the U^A population.The approximate variance of $\hat{Y}_r^B$ implementing fixed sizes sampling designs isgiven by $V(\hat{Y}_r^B \mid \mathbf{m}^A) \cong \sum_{j \in U^A} [(1/\pi_j^A) - 1] \eta_{j,r}^2$,where $\eta_{j,r} = z_{j,r} - \pi_j^A \mathbf{d}'_j \boldsymbol{\beta}_r$ with $\boldsymbol{\beta}_r = \Delta^{-1} \sum_{l \in U^A} \pi_l^A (1/\pi_l^A - 1) \mathbf{d}_j z_{l,r}$.

3. Problem

Given the above framework, we are interested in finding the vector $\boldsymbol{\pi}^A = (\pi_1^A, ..., \pi_j^A, ..., \pi_{M^A}^A)$ of inclusion probabilities that minimizes the expected survey cost bounding the sampling variances, $V(\hat{Y}_v^A)$ (v=1, ... ,V) and $V(\hat{Y}_r^B)$ (r=1,... ,R) under given variance thresholds:

$$\begin{cases} \min \sum_{j \in U^A} c_j \pi_j^A \\ V(\hat{Y}_v^A \mid \mathbf{m}^A) \le V_v^* \quad \forall v = 1,...,V \\ V(\hat{Y}_r^B \mid \mathbf{m}^A) \le V_r^* \quad \forall r = 1,...,R \\ 0 < \pi_j^A \le 1 \end{cases}, \tag{3.1}$$

where V_v^* (v=1,...,V) and V_r^* (r=1,...,R) are the variance thresholds fixed by the sampling designer and c_j is the variable cost for observing the unit j in the population U^A and the linked $L_j^A = \sum_{i=1}^{N^B} L_{j,i}^B$ units in the population U^B. A reasonable expression of c_j is $c_j = f_c(L_j^A; C^B)$,where f_c is a known monotone non-decreasing function and C^B is the per unit cost for observing a cluster in the population U^B .Brewer and Gregoire (2009) propose an extensive analysis of different forms of costs functions. The minimization problem (3.1) is a generalization of the univariate precision constrained optimization approach (Cochran, 1977).The problem (3.1) assumes that all the values $y_{j,v}, y_{i,r}, y_{ik,r}, L_j^A$, $L_{j,i}^B$ and L_i^B are known as alsothe vectors $\boldsymbol{\beta}_v$ and $\boldsymbol{\beta}_r$, although they depend on the vector $\boldsymbol{\pi}^A$. In this case, problem (3.1) becomes a classical Linear Convex Separate Problem (LCSP; Boyd and Vanderberg, 2004) and can be solved by the algorithm proposed in Falorsi and Righi (2015). The algorithm represents a slight modification of the algorithm of Chromy (1987), originally developed for multivariate optimal allocation in Stratified Simple Random Sampling Without Replacement (SSRSWOR) designs and implemented in standard software tools[1]. Alternatively, the LCSP can be dealt with by the SAS procedure NLP as suggested by Choudhry *et al.*(2012).

4. Informative contexts and optimization problem

Optimization problem as presented in (3.1) is quite theoretical since one needs to know the values of the variables of interest in both populations and the values of actual links among the unitsof the two populations. From now on, we introduce three more concrete informative contexts in successive steps. We start from two contexts in which the information is very rich, whereas the third context considers a casein which the information is very poor. Thelatter context is the most common, altough the incresing availability of administrative registers and statistical software tools for data integration increase the plausibility of the first two contexts.

Context 1. The sampling frames for U^A and U^B are available. All the values L_j^A $L_{j,i}^B$ and L_i^B are known and the values of $y_{j,v}, y_{i,r}$ are unknownbut can be predicted by suitable superpopulation models.

This context may be realistic in countries, like the Nordic ones, having well established register based systems (Wallgren and Wallgren, 2014) in which the units ofagiven statistical register have unique identifiers of good quality, which allow to identify the same unit in the whole systems of registers.The*working*models that we studycan be expressed under the following forms:

[1] See for example the Mauss-Rsoftware available at: http://www3.istat.it/strumenti/metodi/software/campione/mauss_r/.

$$\begin{cases} y_{j,v} = \tilde{y}_{j,v} + u_{j,v} = f_v(\mathbf{x}_j;\boldsymbol{\varphi}_v) + u_{j,v} \\ E_{M_v}(u_{j,v}) = 0 \quad \forall j \\ E_{M_v}(u_{j,v}^2) = \sigma_{j,v}^2 \\ E_{M_v}(u_{j,v}, u_{l,v}) = 0 \ \forall j \neq l \end{cases}, \begin{cases} y_{i,r} = \tilde{y}_{i,r} + u_{i,r} = f_r(\mathbf{x}_i;\boldsymbol{\varphi}_r) + u_{i,r} \\ E_{M_r}(u_{i,r}) = 0, \\ E_{M_r}(u_{i,r}^2) = \sigma_{i,r}^2 \\ E_{M_r}(u_{i,r}, u_{i',r}) = 0 \ \forall i \neq i' \end{cases} \tag{4.1}$$

where, omitting the subscripts for the sake of brevity, **x**are vectors of predictors (available in the two sampling frames), $\boldsymbol{\varphi}$ are the vectors of regression coefficients and $f(\mathbf{x};\boldsymbol{\varphi})$ are known functions, u are the error terms, $\tilde{y}$ are the predicted values and $E_M(\cdot)$ denote the expectations under the models. We assume that the parameters of the models to be known, although in practice they are usually estimated. The right-hand side superpopulation model of (4.1) can be defined starting from an elementary unit level model. We do not dealt with this second modelin this paper. The model expectations at cluster level on the right hand side of (4.1) can be easily derived as:

$E_{M_r}(y_{i,r}) = \tilde{y}_{i,r} = \sum_{k=1}^{M_i^B} \tilde{y}_{ik,r}$; $V_{M_r}(y_{i,r}) = \sigma_{i,r}^2 = M_i^B \sigma_r^2 [1 + (M_i^B - 1)\rho_r]$; $Cov_{M_r}(y_{i,r}, y_{i',r}) = 0$ for $i \neq i'$,

where V_{M_r} and Cov_{M_r} denote respectively the model variance and covariance.

Note that the *working* models (4.1) are variable specific. They are introduced as useful tools for defining the sampling design but they are not necessarily representing exactly the real models generating the data.

According to (4.1),the model predictions and the variancesof the z variables are given by

$E_{M_r}(z_{j,r}) = \tilde{z}_{j,r} = \sum_{i=1}^{N^B} \tilde{L}_{j,i}^B \tilde{y}_{i,r}$ and $V_{M_r}(z_{j,r}) = \sigma_{j,zr}^2 = \sum_{i=1}^{N^B} (\tilde{L}_{j,i}^B)^2 \sigma_{i,r}^2$.

Thus, in the optimization problem,the variance terms $V(\hat{Y}_v^A \mid \mathbf{m}^A)$ and $V(\hat{Y}_r^B \mid \mathbf{m}^A)$ are replaced by the Anticipated Variances. Denoting with E the expectation under the sampling design, the anticipated variance of the HT estimator $\hat{Y}_v^A$ is $E_{M_v} E(\hat{Y}_v^A - Y_v^A)^2 = E_{M_v} V(\hat{Y}_v^A - Y_v^A)$., with $V(\hat{Y}_v^A - Y_v^A) = V(\hat{Y}_v^A \mid \mathbf{m}^A) \cong \sum_{j \in U^A} [(1/\pi_j^A) - 1] \eta_{j,v}^2$ The same result may be derived for the estimate $\hat{Y}_r^B$. Thus,we obtain the following expressions: $E_{M_v}[V(\hat{Y}_v^A \mid \mathbf{m}^A)] = \sum_{j \in U^A} [(1/\pi_j^A) - 1] E_{M_v}(\eta_{j,v}^2)$, $E_{M_r}[V(\hat{Y}_r^B \mid \mathbf{m}^A)] = \sum_{j \in U^A} [(1/\pi_j^A) - 1] E_{M_r}(\eta_{j,r}^2)$,where $E_{M_v}(\eta_{j,v}^2)$ and $E_{M_v}(\eta_{j,r}^2)$ arenot given here for the sake of brevity.The problem (3.1) is then reformulatedusing $E_{M_v} V(\hat{Y}_v^A \mid \mathbf{m}^A)$and $E_{M_r} V(\hat{Y}_r^B \mid \mathbf{m}^A)$.

Remark 4.1: Falorsi and Righi (2015) propose an upward approximation of the anticipated variances that simplified the optimization problem. This conservative approximationis a safe choice in this setting, since they prevent from the risk of defining an insufficient sample size for the expected accuracies.

Remark 4.2: Lavallée and Labelle-Blanchet (2013) deal with the problem of indirect sampling applied to skewed populations by suggesting eight alternative methods for modifying the links, $l_{j,ik}$, to reduce the variance of the estimates.

Context 2.Suppose that in the sample design phase, the links $l_{j,ik}$ are not known with certainty but the probabilities of existing links, $Pr(l_{j,ik} = 1) = \lambda_{j,ik}$, are available.

To include the linkage uncertainty in the optimal allocation, we assume the links follow a Bernoulli model $M_l, l_{j,ik} \sim \mathrm{B}(\lambda_{j,ik})$, where $E_{M_l}(l_{j,ik}) = \lambda_{j,ik}$ and $V_{M_l}(l_{j,ik}) = \lambda_{j,ik}(1 - \lambda_{j,ik})$.We assume the parameters $\lambda_{j,ik}$ to be known, although in practice they are usually estimated with probabilistic record linkage procedures(Lavallée and Caron, 2001).

In this framework, the anticipated variance has to take into account both the models M_l and M_r .Since $E_{M_l} E_{M_r} E(\hat{Y}_r^A - Y_r^A)^2 = E_{M_l} E_{M_r} V(\hat{Y}_r^A - Y_r^A) + E_{M_l} V_{M_r} E(\hat{Y}_r^A - Y_r^A) + V_{M_l} V_{M_r} E(\hat{Y}_r^A - Y_r^A)$ and $E(\hat{Y}_v^A - Y_v^A) = 0$, the problem (3.1) can be reformulated replacing the function to be minimized by $\min \sum_{j \in U^A} E_{M_l}(c_j) \pi_j^A$ and the two set of variances respectively with $E_{M_v} V(\hat{Y}_v^A \mid \mathbf{m}^A)$ and $E_{M_l} E_{M_r} V(\hat{Y}_r^B \mid \mathbf{m}^A)$ where $E_{M_l} E_{M_r}[V(\hat{Y}_r^B \mid \mathbf{m}^A)] = \sum_{j \in U^A} [(1/\pi_j^A) - 1] E_{M_l} E_{M_r}(\eta_{j,r}^2)$. The main results for the derivation of the expression of $E_{M_l} E_{M_r} V(\hat{Y}_r^B \mid \mathbf{m}^A)$ are not given in the paper. They are based on the Taylor' series approximation and making some reasonable assumption on the

links. The predicted $\tilde{z}_{j,r}$ values are obtained as $\tilde{z}_{j,r} = \sum_{i=1}^{N^*} \tilde{\Lambda}^B_{j,i}\, \tilde{y}_{i,r}$,where $\tilde{\Lambda}^B_{j,i} = (\Lambda^B_{j,i} / \Lambda^B_i)$,with $\Lambda^B_{j,i} = \sum_{k=1}^{M_i^B} \lambda_{j,ik}$ and $\Lambda^B_i = \sum_{j=1}^{M^A} \Lambda^B_{j,i}$.

Remark4.3:the uncertainty on total survey costs, which depends both on the selected sample and the model uncertainty on costs, obliges to consider in the optimization problem the *expected costs* $E_{M_1}(c_j) = f_c(\Lambda^A_j; C^B)$ where $\Lambda^A_j = \sum_{i=1}^{M^B} \sum_{k=1}^{M_i^B} \lambda_{j,ik}$.

Context 3.Data integration is not possible because the record linkage process does not provide good linkages, or simply because the frame of population U^B does not exist.

This is the most usualcontext in developing countries. However, it may also characterize specific survey contexts in developed countries, for instance in case of hard-to-reach populations.

In this case, the optimizazion problem can be dealt byusing all the available information, even if of poor quality. For instance, if a size variables x related to the variable y_r is known from the frame for the unitsof population U^A, and totals or estimated totals $\tilde{Y}^B_{r(q)}$ of U^B are available at certain domain level q (q=1,…,Q),then the predicted z variables can be determined as:

$$\tilde{z}_{j,r} = \frac{x_j}{\sum_{l \in U^A_q} x_l} \tilde{Y}^B_{r(q)} \quad \text{for } j \in U^A_q. \tag{4.2}$$

Examples of building the z values are illustrated in Section 5.3.2 of *Guidelines on Integrated Survey Framework* (FAO, 2015). Here, an example is reported. Population U^A is given by the farm register of a country. From the register, we known the region q where agiven farm belongs. The population U^B is defined by the rural households. Suppose furthermore that we know (from the Census data or from a previous survey) the total $Y^D_{r(q)}$ for a given variable of interest, e.g., "revenue", for the domain q. In particular, consider the farm j of region q with 50 workers living in the region q. In this region, the total number of workers of the farms (estimated or known) is 330,000, and we havethe total $\tilde{Y}^B_{r(q)} = 100,000$. The predicted revenue $\tilde{z}_{j,r}$ for the farm is given by $\tilde{z}_{j,r} = (50/330{,}000)100{,}000 = 15.15$ .Note that this kind of prediction corresponds to the hypothesisof uniformity of the links in the qth domain. In absence of unit level information for U^B, this hypothesisseems tobe reasonable.In this context, it is necessary to try to model directly the z-value with a model of the type:

$$\begin{cases} z_{j,r} = \tilde{z}_{j,r} + u_{j,zr} = f_{zr}(\mathbf{x}_j; \varphi_{zr}) + u_{j,zr} \\ E_{M_{zr}}(u_{j,zr}) = 0 \quad \forall j; \\ E_{M_{zr}}(u^2_{j,zr}) = \sigma^2_{j,zr}; \\ E_{M_{zr}}(u_{j,zr}, u_{l,zr}) = 0 \quad \forall j \neq l \end{cases} \tag{4.3}$$

where $\mathbf{x}_j$ is a vector of variables related to the size of unit j. For building plausible predictions on the variance $\sigma^2_{j,zr}$,it may be necessary to carry out a pilot survey. However, in some cases, it may be very difficult to implement such an effort. Making reasonable assumption on the relationship of the squared predictions $\tilde{z}^2_{j,r}$ with the variances $\sigma^2_{j,zr}$,the optimization problem could be carried out with considering the variances of the predictions.

It is also necessary to build a model for assessing the survey costs on the total links L^A_j :

$$\begin{cases} L^A_j = \Lambda^A_j + u_{j,\Lambda} = f_\Lambda(\mathbf{x}_j; \boldsymbol{\varphi}_\Lambda) + u_{j,\Lambda} \\ E_{M_\Lambda}(u_{j,\Lambda}) = 0 \quad \forall u_{j,\Lambda}; \\ E_{M_\Lambda}(u^2_{j,\Lambda}) = \sigma^2_{j,\Lambda}; \\ E_{M_\Lambda}(u_{j,\Lambda}, u_{l,\Lambda}) = 0 \quad \forall j \neq l \end{cases} \tag{4.4}$$

The predictions Λ^A_j needto be positive. A useful model is the logaritmic one:

$\log(\Lambda^A_j) = \mathbf{x}'_j \boldsymbol{\varphi}_\Lambda$. The model (4.4) allowsthe prediction of the total number of links Λ^A_j of the unit j, thus defining the expected cost survey cost attached to it.

Accordingly to the models (4.3) and (4.4), the problem (3.1) can be reformulated as follows:

$$\begin{cases} \min \sum_{j\in U^A} E_{M_\Lambda}(c_j)\pi_j^A \\ E_{M_v} V(\hat{Y}_v^A \mid \mathbf{m}^A) \le V_v^* \quad \forall v = 1,\ldots,V \\ E_{M_{zr}} V(\hat{Y}_r^B \mid \mathbf{m}^A) \le V_r^* \quad \forall r = 1,\ldots,R \\ 0 < \pi_j^A \le 1 \end{cases}, 0 \tag{4.5}$$

where $E_{M_\Lambda}(c_j) = f(\Lambda_j^A; C^B)$. The solution algorithm is identical to the one that solves the problem defined in context 2 except that the models for predicting the z-values and the expected costs are less specific, which results in a higher model uncertainty.

Remark 4.4:In some situations, the model variances $\sigma_{j,zr}^2$ are not known, and it is not feasible (for organizational or cost constraints) to carry out a pilot study for assessing them, while the predictions $\tilde{z}_{j,r}$ can be assessed with a super-simplified model as in (4.2). In order to find a realistic sampling solution, it may be reasonable to assume that the following relations hold: $\tilde{z}_{j,r}^2 + \sigma_{j,zr}^2 \cong \kappa \tilde{z}_{j,r}^2$,where $\kappa > 1$.The sample designer may find a *quasi-optimal* sampling solution by running the problem with alternative reasonable choices of the κ value (e.g., κ=2, 3 or 4), and studying the sensitivity of the different solutions.

Remark 4.5: a good strategy which allows to be robust against model failure is to select a balanced sample with respect to the auxiliary variables $\mathbf{x}_j$. In this case, the auxiliary variables $\mathbf{d}_j$ of the balancing equations are replaced by the augmented variables $\mathbf{d}_j^* = (\mathbf{d}_j', \mathbf{x}_j' / \pi_j^A)'$. For the calculation of the variances, the residuals $\eta_{j,v}$ are substituted by the modified residuals $\eta_{j,v}^* = y_{j,v} - \pi_j^A (\mathbf{d}_j^*)' \boldsymbol{\beta}_v^*$, where $\boldsymbol{\beta}_v^* = (\boldsymbol{\Delta}^*)^{-1} \sum_{l\in U^A} \pi_l^A (1/\pi_l^A - 1)\, \mathbf{d}_j^*\, y_{l,v}$ with $\boldsymbol{\Delta}^* = \sum_{j\in U^A} \mathbf{d}_j^* (\mathbf{d}_j^*)' \pi_j^A (1-\pi_j^A)$. For the modified residuals $\eta_{j,r}^*$, similar expressions are used.

REFERENCES

Boyd, S., and Vanderberg, L. (2004).Convex Optimization.Cambridge University Press.

Brewer, K. R. W. and T. G. Gregoire.(2009). Introduction to survey sampling.In Handbook of Statistics – Vol. 29A. Sample Surveys: Design, Methods and Applications. (D. Pfeffermann and C. R. Rao, eds.) Elsevier B. V. pp. 9-37.

Choudhry, G. H., Rao, J. N. K., and Hidiroglou, M. A. (2012) On sample allocation for efficient domain estimation,Survey Methodology, 18, 23-29.

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

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

FAO (2015) Guidelines on Integrated Survey Framework.GUIDELINES & HANDBOOKShttp://gsars.org/en/guidelines-for-the-integrated-survey-framework/ . Accessed on August 2016.

Falorsi, P. D. and Righi, P. (2015),Generalized Framework for Defining the Optimal Inclusion Probabilities of One-Stage Sampling Designs for Multivariate and Multi-domain Surveys, Survey Methodology, 41, 215-236.

Lavallée, P (2007) Indirect sampling.Springer, New York.

Lavallée, P., Caron, P. (2001) Estimation Using the Generalised Weight Share Method: The Case of Record Linkage.Survey Methodology, 27, 155-169.

Lavallée, P., Labelle-Blanchet, S. (2013) Indirect sampling applied to skewed populations », Survey Methodology, 39, 183-215.

Khan, M. G. M., Mati, T., and Ahsan, M. J. (2010) An optimal Multivariate stratified sampling design using auxiliary information: An integer solution using goal programming approach. Journal of Official Statistics, 26, 695-708.

Kokan, A. and Khan, S. (1967) Optimum allocation in multivariate surveys: An analytical solution. Journal of the Royal Statistical Society, Series B, 29, 115-125.

Wallgren A. andWallgren B. (2014) Register-based Statistics: Administrative Data for Statistical Purposes. John Wiley & Sons, Chichester, UK. ISBN: ISBN 978-1-119-94213-9.