Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Cost optimal sampling for the integrated observation of different populations

by Piero Demetrio Falorsi, Paolo Righi and Pierre Lavallée

Release date: December 17, 2019

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at STATCAN.infostats-infostats.STATCAN@canada.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version **is also available.**

Cette publication est aussi disponible en français.

Survey Methodology, December 2019
Vol. 45, No. 3, pp. 485-511
Statistics Canada, Catalogue No. 12-001-X

Cost optimal sampling for the integrated observation of different populations

Piero Demetrio Falorsi, Paolo Righi and Pierre Lavallée[1]

Abstract

Social or economic studies often need to have a global view of society. For example, in agricultural studies, the characteristics of farms can be linked to the social activities of individuals. Hence, studies of a given phenomenon should be done by considering variables of interest referring to different target populations that are related to each other. In order to get an insight into an underlying phenomenon, the observations must be carried out in an integrated way, in which the units of a given population have to be observed jointly with related units of the other population. In the agricultural example, this means that a sample of rural households should be selected that have some relationship with the farm sample to be used for the study.

There are several ways to select integrated samples. This paper studies the problem of defining an optimal sampling strategy for this situation: the solution proposed minimizes the sampling cost, ensuring a predefined estimation precision for the variables of interest (of either one or both populations) describing the phenomenon. Indirect sampling provides a natural framework for this setting since the units belonging to a population can become carriers of information on another population that is the object of a given survey.

The problem is studied for different contexts which characterize the information concerning the links available in the sampling design phase, ranging from situations in which the links among the different units are known in the design phase to a situation in which the available information on links is very poor. An empirical study of agricultural data for a developing country is presented. It shows how controlling the inclusion probabilities at the design phase using the available information (namely the links) is effective, can significantly reduce the errors of the estimates for the indirectly observed population. The need for good models for predicting the unknown variables or the links is also demonstrated.

Key Words: Integrated surveys; Sample allocation; Indirect sampling.

1 Introduction

The need to observe together different populations related to each other is often encountered in social or economic studies. For example, in agricultural studies, the characteristics and behavior of farms can be linked to phenomena not only related to the farms themselves, but also to the social activities of individuals. This requires the study of the population of rural households, in addition to the study of the population of farms, in some integrated way. That is, to get an insight into an underlying phenomenon, the observations must be carried out in an integrated way, implying that the units of a given population have to be observed jointly with the related units of the other population. In the agricultural example, this means that a sample of rural households should be selected that have some relationship with the farm sample to be used for the study.

The integrated observation of two populations implies that if we observe the variables of the unit j of the first population, U^A, we need to observe the variables of all the units in the second population, U^B, which are linked with the j^{th} unit of U^A. The links among the units of the two populations are regulated by formal rules, contingent dependencies or relationships created for these purposes. Continuing with the

1. Piero Demetrio Falorsi and Paolo Righi, Italian National Institute of Statistics. E-mail: falorsi@istat.it, parighi@istat.it; Pierre Lavallée, E-mail: plavall1962@gmail.com.

agricultural example, these studies often refer to different statistical populations such as farms, rural households and land parcels, the units of which are linked to each other. The people of a given household may be the workers of a specific farm and those workers represent the links between the household and the farm. Furthermore, a given farm comprises specific land parcels which represent the links between that farm and the population of land parcels. The integrated observation of such populations allows the measurement global phenomena of the agricultural sector. Consider a given farm: the education level of the farm holder and the farm size, which are variables related to the population of farms, can affect the productivity of the land (a variable related to the statistical population of land parcels) which belongs to the farm. This productivity may have an impact on the risk of malnutrition of the households (population of rural households) in which the workers of the farm live. Thus, the observation of such different units in an integrated way provide insights into the relationships which link the level of education, the land productivity and the risk of malnutrition. If only aggregates are examined, then the advantage of integrated sampling is that it allows sampling from population U^B without having a frame available for it.

Another concrete example where the methodology may be of use is for firm-establishment-employee studies. For instance, the wellness of the households of people employed in firms which have a well-defined policy of social responsibility may be different from that of other types of households and the success in their children's schooling can be higher. In this case, the integrated observation allows the study the behavior of different sub-classes of households defined by a variable observable in the population of firms.

Other examples can be found in socio-demographic studies. For instance, the phenomenon of children who spend time in two households can be studied with the integrated observations of the population U^A of households and the population U^B of children.

Generally speaking, integrated observation may be of use for studying phenomena that involve variables which are correlated but belong to different statistical populations. Integrated observation allows the study of the relationships among all the variables of interest for the given phenomena, even if they belong to different populations. The independent observation of such populations would not allow the observation of the set of all the related variables of interest and hence it would not be possible to study the relationships among all the variables describing the phenomenon.

Indirect sampling (Lavallée, 2002, 2007) provides a natural framework for the estimation of the parameters of two target populations that are related to each other. In the indirect sampling framework, the units belonging to a population that are selected for a given survey can enable the collection of information on another population, through the relationship between the units in the two populations. Furthermore, indirect sampling is suitable for producing statistics of populations for which there is no sampling frame. In such a context, the sampling procedure assumes that population U^A is related to the population of interest U^B, but only the sampling frame of U^A is available. Then, a sample is selected from U^A, and using the links between the two populations, a sample of units of U^B is observed.

This paper studies the problem of sampling design for integrated observation of different populations. For this, an indirect sampling design is implemented. In particular, the focus is on the determination of the

inclusion probabilities. Since the sum of these probabilities define the expected sample size, we roughly define the problem as a *sampling allocation problem*. In fact, the two problems (determination of the inclusion probabilities and sampling allocation) coincide in stratified sampling. The allocation problem for the usual (direct) sampling setting has been dealt with in several books and papers. When one target parameter is to be estimated for the overall population, the optimal allocation in stratified sampling can be performed (Cochran, 1977, and Särndal, Swensson and Wretman, 1992). In particular, the optimal sample allocation minimizes the variance of the estimated total, subject to a given budget or, reversing the problem, a sample allocation that minimizes costs can be performed, subject to a given sampling error constraint. In multivariate cases, where more than one characteristic of each sampled unit must be measured, the optimal allocation for individual characteristics are of little practical use, unless the characteristics under study are highly correlated. This is because an allocation that is optimal for one characteristic can be far from optimal for others. The multidimensionality of the problem also leads to a compromise allocation method (Khan, Mati and Ahsan, 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, Rao and Hidiroglou (2012).

Falorsi and Righi (2015) provide a general framework for sample design in multivariate and multi-domain surveys. This paper offers a further generalization of this framework to the case of integrated observation of two populations. Different scenarios related to the level of knowledge of the links are examined: the first scenario assumes the links between the populations are known in the design phase; the second scenario assumes the links between U^A and U^B are estimated in the design phase; in the third scenario, no links between U^A and U^B are available, but auxiliary variables on U^A can provide useful information on U^B.

Section 2 introduces the background and symbols. Section 3 and Section 4 illustrate the basic optimization problem and how it is applied in the different scenarios. Empirical evidence is shown in Section 5.

2 Background

Let U^A and U^B denote two related target populations, where U^A is the population with the available sampling frame, and U^B the survey population for which a sampling frame may or may not be available. For the agricultural example, U^A is the population of farms and U^B the population of rural households. Let s^A be a sample selected from U^A without replacement and with fixed sample size m^A, where U^A contains M^A units. Let π_j^A represent the inclusion probability of the j^{th} unit in U^A with $\pi_j^A > 0$ and $\sum_{j \in U^A} \pi_j^A = m^A$ with $\boldsymbol{\pi}^A = \left(\pi_1^A, \ldots, \pi_j^A, \ldots, \pi_{M^A}^A\right)'$. We denote by $y_{j,v}$ the value of the v^{th} $(v = 1, \ldots, V)$ characteristic on unit j and their total by Y_v^A.

We estimate the total Y_v^A according to the Horvitz-Thompson (HT) estimator,

$$\hat{Y}_v^A = \sum_{j \in s^A} w_j^A \, y_{j,v}, \tag{2.1}$$

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 $U_h^A\,(h = 1, \ldots, H)$ 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 size sampling designs are those satisfying

$$\sum_{j \in s^A} \mathbf{d}_j = \mathbf{m}^A,$$

where $\mathbf{d}_j = \left(d_{j(1)}, \ldots, d_{j(h)}, \ldots, d_{j(H)}\right)'$ and $\mathbf{m}^A = (m_1^A, \ldots, m_h^A, \ldots, m_H^A)'$ is the 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 $h = 1, \ldots, H$). Several customary fixed size sampling designs may be considered as particular cases. A well-known example is the Stratified Simple Random Sampling WithOut Replacement (SSRSWOR) design where strata are the planned domains and each $\mathbf{d}_j$ vector has $H - 1$ elements equal to zero, and one element equal to 1, which implies that each unit j can belong to one and only one planned domain. Furthermore, in this design all the units in the stratum U_h^A have a uniform inclusion probability given by $\pi_j^A = \frac{m_h^A}{M_h^A}$ for $j \in U_h^A$. If each $\mathbf{d}_j$ vector has $H - 1$ elements equal to zero and one element equal to 1, and the π_j^A values can be different in the stratum, we have a stratified sampling design, without replacement with fixed sample sizes and varying probabilities in each stratum. On the basis of the Winkler's definition (2001), if $\sum_{h=1}^{H} d_{j(h)} > 1$, we have an Incomplete Multi-Way Stratified Sampling design.

We suppose that the $M^A \times H$ matrix $\mathbf{D} = \left(\mathbf{d}_1, \ldots, \mathbf{d}_j, \ldots, \mathbf{d}_{M^A}\right)'$ is non-singular. According to this general sampling design framework, Deville and Tillé (2005) proposed an approximated expression of the variance for $\hat{Y}_v^A$ based on the Poisson sampling theory given by

$$V\left(\hat{Y}_v^A \,\middle|\, \mathbf{m}^A\right) \cong \left[M^A/(M^A - H)\right] \sum_{j \in U^A} \left(\frac{1}{\pi_j^A} - 1\right) \eta_{j,v}^2, \tag{2.2}$$

where $\hat{Y}_v^A \,\middle|\, \mathbf{m}^A$ is the HT estimator based on a general fixed sample size design with $\mathbf{m}^A$ units related to the vector $\boldsymbol{\pi}^A$,

$$\eta_{j,v} = y_{j,v} - \pi_j^A \mathbf{d}_j' \boldsymbol{\beta}_v, \tag{2.3}$$

and

$$\boldsymbol{\beta}_v = \boldsymbol{\Delta}^{-1} \sum_{j \in U^A} \pi_j^A \left(\frac{1}{\pi_j^A} - 1\right) \mathbf{d}_j y_{j,v} \tag{2.4}$$

with

$$\mathbf{\Delta} = \sum_{j \in U^A} \mathbf{d}_j \mathbf{d}_j' \pi_j^A \left(1 - \pi_j^A\right). \tag{2.5}$$

The variance (2.2) resembles the variance expression of the HT estimator under a Poisson sampling design, but it uses the residuals $\eta_{j,v}$, instead of the original value $y_{j,v}$. In practice, when $H = 1$ this is the variance approximation of the Conditional Poisson Sampling design (CPS, as introduced in Deville and Tillé, 2005). CPS selects samples by means of a Poisson sampling design without replacement until a given sample size is obtained.

To clarify the degree of approximation of (2.2), consider the SSRSWOR design. According to expression (2.2), we have

$$V\left(\hat{Y}_v^A \,\middle|\, \mathbf{m}^A\right) = \left[M^A/(M^A - H)\right] \sum_{h=1}^{H} \sigma_{v,h}^2 M_h^A \left(\frac{M_h^A}{m_h^A} - 1\right),$$

where $\sigma_{v,h}^2$ is the design variance of the $y_{j,v}$ values in stratum U_h^A (see Appendix 4 of Falorsi and Righi, 2015). The above approximation works well when the number of domains H remains small compared to the overall population size M^A.

Let M^B, N^B, U_i^B and M_i^B be the number of units in U^B, the number of clusters in U^B, the i^{th} cluster of U^B with $\bigcup_{i=1}^{N^B} U_i^B = U^B$ and the number of units in the i^{th} cluster U_i^B, respectively. We denote by $y_{ik,r}$ the value of the r^{th} $(r = 1, \ldots, R)$ characteristic for the k^{th} unit of the i^{th} cluster of U^B and the population total of all $y_{ik,r}$'s by

$$Y_r^B = \sum_{i=1}^{N^B} \sum_{k=1}^{M_i^B} y_{ik,r}.$$

Let $l_{j,ik}$ be an indicator variable of link existence: $l_{j,ik} = 1$ indicates that there is a link between j^{th} unit in U^A and k^{th} unit in U_i^B, while $l_{j,ik} = 0$ indicates otherwise.

Suppose that we carry out an indirect sampling process: if the unit $j \in U^A$ is included 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 (i.e., $y_{ik,r}$) in the indirect sample of population U^B. Let n^B be the size of the sample of clusters in population U^B obtained after the indirect sampling process. We estimate Y_r^B according to the estimator based on the theory of the Generalized Weight Share Method (GWSM) of Lavallée (2002, 2007):

$$\hat{Y}_r^B = \sum_{i=1}^{n^B} y_{i,r} w_i^B, \tag{2.6}$$

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

$$\tilde{L}_{j,i}^B = \frac{L_{j,i}^B}{L_i^B}$$

and

$$L_i^B = \sum_{j=1}^{M^A} L_{j,i}^B .$$

Theorem in Section 3 of Lavallée (2002, 2007) states that (2.6) provides 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

$$z_{j,r} = \sum_{i=1}^{N^B} \tilde{L}_{j,i}^B \, y_{i,r} , \tag{2.7}$$

the estimator (2.6) can be expressed as a usual Horvitz-Thompson (HT) estimator on the z values referring to the U^A population,

$$\hat{Y}_r^B = \sum_{s^A} z_{j,r} w_j^A . \tag{2.8}$$

Therefore, the variance $V\left(\hat{Y}_r^B\right)$ of $\hat{Y}_r^B$ may be expressed as the variance of the HT estimator on the U^A population. The approximate variance of $\hat{Y}_r^B$ for fixed size sampling designs is given by

$$V\left(\hat{Y}_r^B \,\middle|\, \mathbf{m}^A\right) \cong \left[M^A/(M^A - H)\right] \sum_{j \in U^A} \left(\frac{1}{\pi_j^A} - 1\right) \eta_{j,r}^2 , \tag{2.9}$$

where $\hat{Y}_r^B \,\middle|\, \mathbf{m}^A$ is the HT estimator based on a general fixed sample size design with $\mathbf{m}^A$ units and the related vector $\boldsymbol{\pi}^A$,

$$\eta_{j,r} = z_{j,r} - \pi_j^A \mathbf{d}_j' \boldsymbol{\beta}_r$$

with

$$\boldsymbol{\beta}_r = \boldsymbol{\Delta}^{-1} \sum_{j \in U^A} \pi_j^A \left(\frac{1}{\pi_j^A} - 1\right) \mathbf{d}_j \, z_{j,r} .$$

Remark 2.1. An interesting extension of the above framework, useful in case of integrated studies, is the case of a total derived from a cross tabulation of a variable of the population U^A with a variable of the population U^B. In order to illustrate this extension, let y_v be a variable of U^A with C modalities and let $y_{j,v(c)}$ denote a dichotomous variable where $y_{j,v(c)} = 1$ if the unit j is characterized by the modality

$c\,(c = 1, \ldots, C)$ of y_v, and $y_{j,v(c)} = 0$ otherwise. Furthermore, let y_r be a variable of U^B with G modalities and let $y_{ik,r(g)}$ denote a dichotomous variable where $y_{ik,r(g)} = 1$ if the unit k of the cluster i is characterized by the modality $g\,(g = 1, \ldots, G)$ of y_r, and $y_{ik,r(g)} = 0$ otherwise. The total number of units of U^B characterized by the modality $g\,(g = 1, \ldots, G)$ of y_r and linked with units of population U^A characterized by the modality c of the variable y_v, can be defined as

$$Y^B_{(c,g)} = \sum_{j=1}^{N^A} \sum_{i=1}^{N^B} \sum_{k=1}^{M_i^B} l_{j,ik}\, y_{j,v(c)}\, y_{ik,r(g)} = \sum_{i=1}^{N^B} y_{i,(c,g)},$$

where $y_{i,(c,g)} = \sum_{j=1}^{N^A} \sum_{k=1}^{M_i^B} l_{j,ik}\, y_{j,v(c)} y_{ik,r(g)}$.

As an example, let us consider the case, illustrated in the introduction, of an integrated analysis examining the productivity of farms and the malnutrition of households, and suppose that $Y^B_{(c,g)}$ represents the total of persons with a malnutrition problem in the households of workers of farms characterized by an high productivity. In this case $y_{j,v(c)}$ has value 1 if the productivity of the farm j is high and $y_{ik,r(g)}$ has value 1 if the person k of the household i has a problem of malnutrition.

The GWSM estimator of $Y^B_{(c,g)}$ can be obtained directly from expression (2.8) using the transformed variable $z_{j,r} = \sum_{i=1}^{N^B} \tilde{L}^B_{j,i}\, y_{i,(c,g)}$.

3 Problem

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

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

where $\mathbf{m}^A$ is given by the $\boldsymbol{\pi}^A$ vector minimizing the cost function, $V_v^*\,(v = 1, \ldots, V)$ and V_r^* $(r = 1, \ldots, R)$ 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 $L_j^A = \sum_{i=1}^{N^B} L_{j,i}^B$ linked units in the population U^B. In other words, we want to obtain the optimal selection probabilities that will minimize the variance of estimates obtained for both U^A and U^B. For the agricultural example, this would translate to developing optimal selection probabilities that will lead to estimates for the population of farms, as well as the population of rural households, with specified precision.

A reasonable expression of c_j is

$$c_j = f_c\left(C^A, L_j^A; C^B\right), \tag{3.2}$$

where f_c is a known monotone non-decreasing function, C^A is the per unit cost for observing a unit in the population U^A and C^B is the cost for observing the elementary unit 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 $y_{j,v}$, $y_{i,r}$, L_j^A, $L_{j,i}^B$, L_i^B, $\boldsymbol{\beta}_v$ and $\boldsymbol{\beta}_r$ are known. In this case, problem (3.1) becomes a classical Linear Convex Separate Problem (LCSP) (Boyd and Vanderberg, 2004) and it can be solved by the algorithm proposed Chromy (1987), originally developed for multivariate optimal allocation in an SSRSWOR design and implemented in standard software tools. (See for example the Mauss-R software available at: http://www3.istat.it/strumenti/metodi/software/campione/mauss_r/.) Alternatively, the LCSP can be dealt with by the SAS procedure NLP as suggested by Choudhry et al. (2012). The vectors $\boldsymbol{\beta}_v$ and $\boldsymbol{\beta}_r$ depend on the vector $\boldsymbol{\pi}^A$. Falorsi and Righi (2015) define a new algorithm which finds the optimal solution taking into account the dependence between $\boldsymbol{\beta}_v$ and $\boldsymbol{\beta}_r$ with the optimal vector $\boldsymbol{\pi}^A$.

4 Informative contexts and optimization problem

Optimization problems as presented in (3.1) are quite theoretical since one needs to know the values of the variables of interest in both populations U^A and U^B, and the values of actual links among the units of the two populations. We now present three more concrete contexts involving various amount of information. We start from two contexts in which the information is very rich, whereas the third context considers a case in which the information is very poor. The latter context is the most common, although the growing availability of administrative registers and statistical software tools for data integration increases 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 unknown but can be predicted by suitable superpopulation models.

This context may be realistic in countries, such as the Nordic ones, having well established register-based systems (Wallgren and Wallgren, 2014) in which the units of a given statistical register have unique identifiers of good quality, which allows identification of the same unit in the whole systems of registers. For the agricultural example, this means that one can link each farm to one or more rural households, and each rural household to one or more farms.

The *working* models that we study can be expressed under the following forms:

Unit level **Cluster level**

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

where, omitting the subscripts for sake of brevity, $\mathbf{x}$ 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. The predictors $\mathbf{x}$ in the unit and cluster level models can be different. We assume that the parameters of the models are known, although in practice they are usually estimated.

Even if the model $f_r(\cdot)$ is not known, the model expectations at cluster level for the population U^B can be derived from a model defined at *elementary* unit level, indicated with $f_{re}(\cdot)$. The elementary unit level model can be stated as $y_{ik,r} = \tilde{y}_{ik,r} + u_{ik,r} = f_{re}(\mathbf{x}_{ik}; \boldsymbol{\varphi}_r) + u_{ik,r}$; $E_{M_{re}}(u_{ik,r}) = 0$; $E_{M_{re}}(u_{ik,r}^2) = \sigma_r^2$; $E_{M_{re}}(u_{ik,r}, u_{ik',r}) = \sigma_r^2 \ \rho_r \ \ \forall k \neq k'$; $E_{M_{re}}(u_{ik,r}, u_{i'k',r}) = 0 \ \ \forall i \neq i'$; where ρ_r is the intra-cluster correlation.

The model expectations at cluster level on the right-hand side of (4.1) can be easily derived as:

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

Note that the *working* models (4.1) are variable specific. They are introduced as useful tools for developing 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 variances of the z variables are given by

$$E_{M_r}(z_{j,r}) = \tilde{z}_{j,r} = \sum_{i=1}^{N^B} \tilde{L}_{j,i}^B \, \tilde{y}_{i,r} \ \text{ and } \ V_{M_r}(z_{j,r}) = \sigma_{j,zr}^2 = \sum_{i=1}^{N^B} \left(\tilde{L}_{j,i}^B\right)^2 \ \sigma_{i,r}^2. \tag{4.2}$$

Thus, in the optimization problem (3.1), the variance terms, $V(\hat{Y}_v^A \,|\, \mathbf{m}^A)$ and $V(\hat{Y}_r^B \,|\, \mathbf{m}^A)$, are replaced by the Anticipated Variances. Denoting with $E(\cdot)$ the expectation under the sampling design, the anticipated variance (AV) of $\hat{Y}_v^A$ may be reformulated as follows:

$$\text{AV}(\hat{Y}_v^A) = E_{M_v} E(\hat{Y}_v^A - Y_v^A)^2 = E_{M_v} V(\hat{Y}_v^A - Y_v^A) + V_{M_v} E(\hat{Y}_v^A - Y_v^A).$$

We have

$$E(\hat{Y}_v^A - Y_v^A) = 0,$$

and

$$V(\hat{Y}_v^A - Y_v^A) = V(\hat{Y}_v^A \,|\, \mathbf{m}^A) \ \cong \ \sum_{j \in U^A} \left(\frac{1}{\pi_j^A} - 1\right) \eta_{j,v}^2.$$

The same result may be derived for the estimate $\hat{Y}_r^B$. Thus, we obtain the following expressions:

$$\text{AV}(\hat{Y}_v^A) = E_{M_v} V(\hat{Y}_v^A \,|\, \mathbf{m}^A) \ \cong \ \sum_{j \in U^A} \left(\frac{1}{\pi_j^A} - 1\right) E_{M_v}(\eta_{j,v}^2) \tag{4.3}$$

$$\mathrm{AV}\left(\hat{Y}_r^B\right) = E_{M_r} V\left(\hat{Y}_r^B \,\middle|\, \mathbf{m}^A\right) \cong \sum_{j \in U^A} \left(\frac{1}{\pi_j^A} - 1\right) E_{M_r}\left(\eta_{j,r}^2\right) \tag{4.4}$$

where $E_{M_v}\left(\eta_{j,v}^2\right)$ and $E_{M_r}\left(\eta_{j,r}^2\right)$ are given by expressions (A.2) and (B.2) of Appendices A and B.

The problem (3.1) for searching the optimal $\boldsymbol{\pi}^A$ vector is then reformulated as follows:

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

Remark 4.1. The anticipated variances in (4.5) have cumbersome formulae. A conservative simplified expression of $E_{M_v} V\left(\hat{Y}_v^A \,\middle|\, \mathbf{m}^A\right)$ is given in Remark 4.1 of Falorsi and Righi (2015). More simplified conservative approximations of both $E_{M_v} V\left(\hat{Y}_v^A \,\middle|\, \mathbf{m}^A\right)$ and $E_{M_r} V\left(\hat{Y}_r^B \,\middle|\, \mathbf{m}^A\right)$ are obtained by approximating the sampling design variance with the Poisson sampling variance. We then have

$$E_{M_v} V\left(\hat{Y}_v^A \,\middle|\, \mathbf{m}^A\right) \le \sum_{j \in U^A} \left(\frac{1}{\pi_j^A} - 1\right) E_{M_v}\left(y_{j,v}^2\right),\ E_{M_r} V\left(\hat{Y}_r^B \,\middle|\, \mathbf{m}^A\right) \le \sum_{j \in U^A} \left(\frac{1}{\pi_j^A} - 1\right) E_{M_r}\left(z_{j,r}^2\right),$$

replacing $\eta_{j,\upsilon}$ and $\eta_{j,r}$ by $y_{j,\upsilon}$ and $z_{j,r}$, respectively, where $E_{M_v}\left(y_{j,v}^2\right) = \tilde{y}_{j,v}^2 + \sigma_{j,v}^2$ and $E_{M_r}(z_{j,r}^2) = \tilde{z}_{j,r}^2 + \sigma_{j,zr}^2$ (see Appendix B). Conservative approximations are a safe choice in this setting, since they eliminate 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 in the presence of skewed populations, while keeping estimation unbiased. Using the methods 2 and 3 proposed by these authors, the algorithm can run by simply replacing the links $l_{j,ik}$ by weighted links, $\theta_{j,ik}$, in $E_{M_r} V\left(\hat{Y}_r^B \,\middle|\, \mathbf{m}^A\right)$.

Context 2. The links $l_{j,ik}$ are not known with certainty but the probabilities of links existing, $\Pr\left(l_{j,ik} = 1\right) = \lambda_{j,ik}$, are available.

To include the linkage uncertainty in the optimization, we assume the links follow a Bernoulli model M_l, $l_{j,ik} \sim \mathrm{B}\left(\lambda_{j,ik}\right)$, where $E_{M_l}\left(l_{j,ik}\right) = \lambda_{j,ik}$ and $V_{M_l}\left(l_{j,ik}\right) = \lambda_{j,ik}\left(1 - \lambda_{j,ik}\right)$. 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). For the agricultural example, such a situation would occur when, for instance, the population of farms is linked to the population of rural households using probabilistic record linkage because no common identifier exists. In this framework, the anticipated variance must take into account both models M_l and M_r. Since

$$E_{M_l}E_{M_r}E\left(\hat{Y}_r^B - Y_r^B\right)^2 = E_{M_l}E_{M_r}V\left(\hat{Y}_r^B - Y_r^B\right) + E_{M_l}V_{M_r}E\left(\hat{Y}_r^B - Y_r^B\right) + V_{M_l}E_{M_r}E\left(\hat{Y}_r^B - Y_r^B\right)$$

and $E\left(\hat{Y}_r^B - Y_r^B\right) = 0,$ the problem (4.5) can be reformulated as follows:

$$\begin{cases} \min \sum_{j\in U^A} E_{M_l}\left(c_j\right)\ \pi_j^A \\ E_{M_v}V\left(\hat{Y}_v^A \,\middle|\, \mathbf{m}^A\right) \le V_v^* \quad \forall v = 1, \ldots, V \\ E_{M_l}E_{M_r}V\left(\hat{Y}_r^B \,\middle|\, \mathbf{m}^A\right) \le V_r^* \quad \forall r = 1, \ldots, R \\ 0 < \pi_j^A \le 1 \quad \forall j = 1, \ldots, M^A \end{cases} \tag{4.6}$$

where

$$E_{M_l}E_{M_r}V\left(\hat{Y}_r^B \,\middle|\, \mathbf{m}^A\right) \cong \sum_{j\in U^A}\left(\frac{1}{\pi_j^A} - 1\right)E_{M_l}E_{M_r}\left(\eta_{j,r}^2\right), \tag{4.7}$$

$$E_{M_l}\left(c_j\right) = f_c\left(\Lambda_j^A; C^B\right),$$

with $\Lambda_j^A = \sum_{i=1}^{N^B}\Lambda_{j,i}^B$ and $\Lambda_{j,i}^B = \sum_{k=1}^{M_i^B}\lambda_{j,ik}.$

The main results for the derivation of the expression of $E_{M_l}E_{M_r}V\left(\hat{Y}_r^B \,\middle|\, \mathbf{m}^A\right)$ are given in Appendix C. These are derived using Taylor series approximation and postulating the independence of the process which generates the links $l_{j,ik}$ with the one that creates the variables of interest $y_{i,r}$. Under these approximations, the predicted values $\tilde{z}_{j,r}$ are obtained as

$$\tilde{z}_{j,r} \cong \sum_{i=1}^{N^B}\tilde{\Lambda}_{j,i}^B\tilde{y}_{i,r} \tag{4.8}$$

where

$$\tilde{\Lambda}_{j,i}^B = \frac{\Lambda_{j,i}^B}{\Lambda_i^B}$$

with

$$\Lambda_i^B = \sum_{j=1}^{M^A}\Lambda_{j,i}^B. \tag{4.9}$$

The uncertainty on total survey costs, which depends both on the selected sample and the model uncertainty on costs, obliges us to consider the expected costs $E_{M_l}\left(c_j\right)$ in the optimization problem. Steel and Clark (2014) show how the uncertainty on the expected costs can affect the accuracy of the sample design.

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 common context in developing countries. It may also characterize specific survey contexts in developed countries, for instance in the case of hard-to-reach populations. Returning to the agricultural example, this would mean that one might have a list of farms, but not a list of rural households.

In this case, the problem of optimal integrated sampling can be solved by using all the available information, even if of poor quality. In the following, three options for dealing with the optimization problem are illustrated starting from the option which requires the minimum of information to those which need more information that could be expensive to obtain.

Option 3.1. *Building the predictions of the z variables and decreasing the variance thresholds V_r^* by a scale factor.* Suppose that from the frame of population U^A, it is possible to know the values of a size variable γ related to the total links L_j^A of the units j. For instance, if the population U^A is a population of farms and the population U^B is a population of households, then the number of workers in the farms (variable γ_j) can represent a good approximation of the total number of links, L_j^A, of the farm. Suppose further that the totals or the estimated totals, $\tilde{Y}_{r(q)}^B$, are available at certain domain level, $U_{(q)}^B$ $(q = 1, \ldots, Q)$, defined at geographic level, with $U^B = \bigcup_{q=1}^{Q} U_{(q)}^B$ and $U_{(q)}^B \cap U_{(q')}^B = \varnothing$ for $q \neq q'$. Then the predicted z variables can be defined as:

$$\tilde{z}_{j,r} = \frac{\gamma_j}{\sum_{l \in U_{(q)}^A} \gamma_l} \tilde{Y}_{r(q)}^B \quad \text{for} \quad j \in U_{(q)}^A, \tag{4.10}$$

where $U_{(q)}^A$ denotes the geographic domain q for the population U^A. In practice, the ratio approach in (4.10) assumes that unit j can be given a share of the total $\tilde{Y}_{r(q)}^B$ proportional to the size of the unit itself. Other examples of building the predictions of the z values are illustrated in Section 5.3.2 of *Guidelines on Integrated Survey Framework* (FAO, 2015).

Having determined the predictions, $\tilde{z}_{j,r}$, it may be reasonable to assume that the following relationship holds:

$$E_{M_{zr}}\left(z_{j,r}^2\right) = \tilde{z}_{j,r}^2 + \sigma_{j,zr}^2 \cong k_r \tilde{z}_{j,r}^2, \tag{4.11}$$

where $k_r > 1$. Under (4.11), it is straightforward to show that

$$E_{M_{zr}} V\left(\hat{Y}_r^B \,\middle|\, \mathbf{m}^A\right) \cong k_r V\left(\hat{\tilde{Y}}_r^B \,\middle|\, \mathbf{m}^A\right),$$

where $\hat{\tilde{Y}}_r^B = \sum_{j \in s^A} w_j^A \tilde{z}_{j,r}$. The sampling variance $V\left(\hat{\tilde{Y}}_r^B \,\middle|\, \mathbf{m}^A\right)$ may be computed using expressions (2.2), (2.3), (2.4) and (2.5) by substituting the variable $y_{j,v}$ the prediction $\tilde{z}_{j,r}$. The optimization problem for searching for the optimal $\boldsymbol{\pi}^A$ vector can then be reformulated as:

$$\begin{cases} \min \sum_{j \in U^A} E_{M_\Lambda}\left(c_j\right) \pi_j^A \\ E_{M_v} V\left(\hat{Y}_v^A \,\middle|\, \mathbf{m}^A\right) \leq V_v^* \quad \forall v = 1, \ldots, V \\ V\left(\hat{\tilde{Y}}_r^B \,\middle|\, \mathbf{m}^A\right) \leq V_r^* / k_r \quad \forall r = 1, \ldots, R \\ 0 < \pi_j^A \leq 1 \quad \forall j = 1, \ldots, M^A. \end{cases} \tag{4.12}$$

The sample designer may find the solution by running the optimization problem (4.12) with alternative reasonable choices of the k_r value (e.g., $k_r = 2$, 3 or 4), and studying the sensitivity of the different solutions. Note that $k_r \cong 1 + \left[\mathrm{CV}\left(z_{j,r}\right)\right]^2$, where $\left[\mathrm{CV}\left(z_{jr}\right)\right]^2 = \sigma^2_{j,zr} \big/ \tilde{z}^2_{j,r}$. Therefore (4.11) holds if the $\left[\mathrm{CV}\left(z_{j,r}\right)\right]^2$ values are approximately constant.

Option 3.2. *Extremal case of Context 2, with uniformity of links in specific domains.* If the number or estimated number of clusters and of elementary units $N^B_{(q)}$ and $M^B_{(q)}$ of the domains $U^B_{(q)}$ $(q = 1, \ldots, Q)$ are available, then in the absence of information on the links $l_{j,ik}$, it might be reasonable to assume that these are homogeneous over the domains; that is, $l_{j,ik} \sim B\left(\lambda_{j,ik}\right)$, where $\lambda_{j,ik} = \gamma_j \big/ M^B_{(q)}$.

Furthermore, suppose that, in this context, the predictions $\tilde{y}_{i,r}$ and the sampling variances $\sigma^2_{i,r}$ could be assumed to be homogeneous within the domains $U^B_{(q)}$, i.e., $\tilde{y}_{i,r} = \tilde{y}_{r(q)}$ and $\sigma^2_{i,r} = \sigma^2_{r(q)}$ for $i \in U^B_{(q)}$. Then, the optimization problem may be dealt with as an extremal case of Context 2, with uniformity of links in specific domains.

Remark 4.3. Note that with this option, the predictions $\tilde{z}_{j,r}$ are equivalent to those expressed in (4.10). Indeed, it is reasonable to consider that, in the absence of information, the size in terms of elementary unit of the cluster U^B_i can be set as equal to its mean defined at the domain level: $M^B_i \cong \bar{M}^B_{(q)} = M^B_{(q)} \big/ N^B_{(q)}$ for $U^B_i \in U^B_{(q)}$. Then, the following approximations hold

$$\Lambda^B_{j,i} = \sum_{k \in U^B_i} \lambda_{j,ik} \cong \bar{M}^B_{(q)} \frac{\gamma_j}{M^B_{(q)}} = \frac{\gamma_j}{N^B_{(q)}}; \quad \Lambda^B_i = \sum_{j \in U^A_{(q)}} \Lambda^A_{j,i} \cong \frac{1}{N^B_{(q)}} \sum_{j \in U^A_{(q)}} \gamma_j .$$

Therefore, setting $\tilde{Y}^B_{r(q)} = \tilde{y}^B_{r(q)} N^B_{(q)}$ and postulating the independence of the process which generates the links $l_{j,ik}$ with the one that creates the variables of interest $y_{i,r}$, we can obtain

$$\tilde{z}_{j,r} \cong \sum_{i \in U^B_{(q)}} \frac{\Lambda^B_{j,i}}{\Lambda^B_i} \tilde{y}_{r(q)} = \sum_{i \in U^B_{(q)}} \frac{\gamma_j \big/ N^B_{(q)}}{\sum_{j \in U^A_{(q)}} \gamma_j \big/ N^B_{(q)}} \tilde{y}_{r(q)} = \frac{\gamma_j}{\sum_{j \in U^A_{(q)}} \gamma_l} \tilde{y}^B_{r(q)} N^B_{(q)} \quad \text{for} \quad j \in U^A_{(q)}.$$

Option 3.3. *Modeling the $z_{j,r}$ values.* Another alternative may be carried out by trying to model directly the z-values and the total number of links L^A_j with models of the type:

$$\begin{cases} z_{j,r} = \tilde{z}_{j,r} + u_{j,zr} = f_{zr}\left(\mathbf{x}_j; \boldsymbol{\varphi}_r\right) + u_{j,zr} \\ E_{M_{zr}}\left(u_{j,zr}\right) = 0, \; E_{M_{zr}}\left(u^2_{j,zr}\right) = \sigma^2_{j,zr}, \; \forall j \\ E_{M_{zr}}\left(u_{j,zr}, \; u_{j',zr}\right) = 0, \; \forall j \neq j' \end{cases} , \quad \begin{cases} L^A_j = \Lambda^A_j + u_{j,\Lambda} = f_\Lambda\left(\boldsymbol{\theta}_j; \boldsymbol{\varphi}_\Lambda\right) + u_{j,\Lambda} \\ E_{M_\Lambda}\left(u_{j,\Lambda}\right) = 0, \; E_{M_\Lambda}\left(u^2_{j,\Lambda}\right) = \sigma^2_{j,\Lambda}, \; \forall j \\ E_{M_r}\left(u_{j,\Lambda}, \; u_{j',\Lambda}\right) = 0, \; \forall j \neq j' \end{cases} \quad (4.13)$$

where $\mathbf{x}_j$ and $\boldsymbol{\theta}_j$ are vectors of auxiliary variables. The predictions Λ^A_j need to be positive. A useful model is the log-linear one (Xu and Lavallée, 2009): $\log\left(\Lambda^A_j\right) = \boldsymbol{\theta}'_j \boldsymbol{\varphi}_\Lambda$. The model on the right hand side of (4.13) allows the prediction of the total number of links Λ^A_j of the unit j, thus defining the expected survey cost attached to it. The optimization problem could be carried out using the variances of the predictions of the models (4.13).

Remark 4.4. Option 3.1 requires the minimum of information for the construction of the predictions $\tilde{z}_{j,r}$ and needs us to define of plausible values for the constants k_r. Option 3.2 involves the same information as Option 3.1 for the construction of the predictions $\tilde{z}_{j,r}$ (see Remark 4.3) but requires an estimate of the parameters $\sigma^2_{r(q)}$. These estimates can be obtained from either pilot or previous surveys conducted directly on the population U^B. Option 3.3 is the most complex and expensive, since it involves carrying out indirect pilot surveys on the population U^A for building plausible predictions of the parameters $\tilde{z}_{j,r}$, Λ^A_j, $\sigma^2_{j,zr}$ and $\sigma^2_{j,\Lambda}$.

Remark 4.5. A good strategy that should 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 $\mathbf{d}^*_j = \left(\mathbf{d}'_j, \mathbf{x}'_j \left/ \pi^A_j\right.\right)'$. For the calculation of the variances, the residuals $\eta_{j,v}$ are substituted by the modified residuals $\eta^*_{j,v} = y_{j,v} - \pi^A_j \left(\mathbf{d}^*_j\right)' \boldsymbol{\beta}^*_v$, where $\boldsymbol{\beta}^*_v = (\boldsymbol{\Delta}^*)^{-1} \sum_{j\in U^A} \pi^A_j \left(\frac{1}{\pi^A_j} - 1\right) \mathbf{d}^*_j y_{j,v}$ with $\boldsymbol{\Delta}^* = \sum_{j\in U^A} \mathbf{d}^*_j \left(\mathbf{d}^*_j\right)' \pi^A_j \left(1 - \pi^A_j\right)$. For the modified residuals $\eta^*_{j,r}$, similar expressions are used.

Remark 4.6. A proportional-to-population-size allocation may be a reasonable strategy for stratified sampling designs in which the total sample size m^A is fixed. In this case the stratum sample size, m^A_h, may be defined as $m^A_h = m^A \left(\sum_{j\in U^A_h} x_j \left/ \sum_{j\in U^A} x_j\right.\right)$, where x_j is the measure of the size.

5 Empirical results

The results herein illustrated are obtained using real data from Districts 7, 8, 9 of the Gaza Province, Mozambique. They summarize the empirical results from an evaluation study illustrated in FAO (2014). Other empirical results of the proposed strategies (FAO, 2015) have been conducted on the database of agricultural households from Burkina Faso's General Census of Agriculture and confirm the general results illustrated below.

In the analysis using Mozambique data, the population U^A refers to farms. The database used for the experimentation include environmental and economic variables and gathers the information from the 2007 census of large and medium farms and from a sample survey of small farms (for the same year). The overall number of records is about 36,890, of which 890 are large and medium farms.

The second population, U^B, is the 2007 household census. The database's records are the individuals involved in agricultural, fishing, or forestry activities. The database contains approximately 54,000 records and includes several socio-demographic environmental and economic variables. The databases of the two populations were merged, creating a Master Sampling Frame (MSF) with artificial links between individuals and farms. The merging procedure exploited the following variables: for individuals, the type of job and the district of residence; for farms, the sector, the district and the number of employed persons by type of job. Before merging, a cleaning step of U^B was carried out, discarding records that did not feature the job type variable (approximately 9,000 records). Subsequently, approximately 36,000 records of U^B declare to be a

farmholder without any employed persons. For these cases, a one-to-one farm-individual link was defined. The remaining individuals were linked with the 890 farms, according to the following hierarchical rules:

- Each farm was linked to a number of individuals equal to the number of workers, depending upon job type;
- Individuals and farms in the same district were linked;
- Individuals were linked with private/public governmental farms when the type of employment and the farm sector agree.

The links were generated randomly, according to the categories defined by the hierarchical rules. The exercise did not seek to predict the links that actually exist in the two populations, but rather to create a realistic dataset for the evaluation.

Although the datasets of the two populations include several variables, in this study we have decided to focus on two of these. For U^A, we consider the number of animals, while for U^B, we consider the number of trees. This is in order to better highlight the impact (in terms of both accuracy and sample size) that the different contexts, described in Section 4, have on the sample of the population U^B. Summary statistics on these variables are shown in Table 5.1.

Table 5.1
The variables used in the simulation with data from Mozambique

Population*	Number of records*	Variable	Mean value	%CV**
U^A: Farms	36,890***	Number of animals	11.1	681.6
U^B: Households	45,000	Number of trees	4.5	107.5

* Districts 7, 8, 9 of the Gaza Province, Mozambique.
** %CV = (population standard deviation/mean) × 100.
*** From these, 890 are large and medium farms.

For both populations, we have considered as domains of interest the districts (3 domains) and the province (1 domain). Therefore, in total we consider 8 target totals of interest (2 variables × 4 domains).

5.1 Optimal designs for the different contexts

In the following, we address four contexts:

Context 0. No control on the sample of the population U^B. The sample is planned, controlling only the accuracy of the estimates of the variables of the farms. Once the sample for U^A is selected, the units of U^B, linked to those selected for the sample of the U^A population are included in the sample via the indirect sampling mechanism. The expected percent CVs, %CV, of the estimates obtained from the indirect sample of households are then computed as $\%\mathrm{CV} = \left(\sqrt{\mathrm{AV}\left(\hat{Y}\right)}\Big/Y\right)\times 100$.

Context 1. Sampling frames exist for both populations. All links are known and an integrated sample design is used, finding an optimal solution considering both populations. Therefore, the multivariate allocation is

carried out, controlling the accuracy of estimates from both the direct sample of farms and the indirect sample of individuals.

Context 2. Sampling frames exist for both populations, but links are estimated probabilities and an integrated sample design is used.

Context 3. A frame exists only for the population U^A. An integrated sample design is studied considering Options 3.1 and 3.2, which represent the most feasible solutions in real contexts.

Contexts 1, 2 and 3 are those defined in Section 4. Context 0 is introduced because it represents a useful tool for the evaluation of the integrated strategy.

A stratified sampling mechanism is assumed for the first population U^A, where the strata U_h^A are defined as districts (7, 8 and 9) by size class (1, 2, 3-4, 5-9, 10-19, 20-49, 50-99, 100+) based on the number of farm workers, thus obtaining 21 strata. As regards models (4.1), we considered mean stratum models with $\tilde{y}_{j,v} = \tilde{y}_{v,(h)}$ and $\sigma^2_{j,v} = \sigma^2_{v,(h)}$ for $j \in U_h^A$. These specifications lead to a standard SSRSWOR design for the farms where the strata coincide with the planned domains (see Falorsi and Righi, 2015, Remark 4.2). For the evaluation we used the exact formula of the variance for a SSRSWOR instead of using the approximation of variance for a SSRSWOR given in Section 2; however the two expressions are substantially equivalent. For U^B, we also consider a mean model, defined at district level d, with $\tilde{y}_{i,r} = \tilde{y}_{r,(d)}$, and $\sigma^2_{i,r} = \sigma^2_{r,(d)}$ for $i \in U_d^B$.

The evaluation studies use software, developed in the R language, that implements the optimal sampling for the standard SSRSWOR designs as well as for more general sampling designs (e.g., balanced designs and incomplete stratification designs). It is available at http://www.istat.it/en/tools/methods-and-it-tools/design-tools/multiwaysampleallocation). Once installed, the software features a comprehensive user guide in English. Another software which considers only the SSRSWOR designs is MAUSS-R available at http://www.istat.it/it/strumenti/metodi-e-software/software/mauss-rdownload.

For each context, the variance constraints are expressed in terms of %CVs. The analyses presented in this section are focused on the contexts, and we use a symplified version of the cost functions. The cost c_j for observing the unit j in the population U^A with the linked units in the population U^B is fixed as equal to 1. More detailed analyses on costs are presented in Section 5.2.

Some further specifications for each context are herein illustrated (see Table 5.2).

Context 0. The variance constraints are fixed (only for the farm estimates: number of animals) at 6.5% at the province level and at 10% at the district level, resulting in a sample of 2,122 farms.

Context 1. The constraints for the farm and household estimates have been fixed so as to determine a sample roughly of 2,100 farms. In this way, the variance constraints are fixed for the farm estimates, animals at 10% at the province level and at 15% at the district level. Those for the household estimates are fixed at 2.5% at the province level and 5% at the district level. Note that this choice of constraints makes it possible to carry out the comparison between the two contexts with roughly the same sample size, even if in Context 1, the variance constraints on the estimates of the population U^A are larger than those fixed in Context 0.

Context 2. The CV constraints for the household and farm estimates are equal to those adopted in Context 1. The integrated observation is planned in the sample design phase by taking into account the uncertainty in the links. This has been carried out by considering a simplified model which assumes that, for each worker in a given farm, there is only one *strong* link (with value ψ) with an individual in the population of households and α weak links (with value τ) with other individuals in the same district, where ψ and τ are probabilities, where $\psi \gg \tau$. Let $l_{j\omega,ik}$ denote the link between the worker ω of the farm j and the individual k of the household i and suppose that these links follow a Bernoulli model M_l, where

$$E_{M_l}\left(l_{j\omega,ik}\right) = \lambda_{j\omega,ik} = \begin{cases} \psi \text{ for only one worker } j\omega \in U^A \text{ and one individual } ik \in U^B \\ \tau \text{ for only one worker } j\omega \in U^A \text{ and } \alpha \text{ individuals } ik \in U^B \end{cases}, \quad (5.1)$$

in which $\tau = \frac{1-\psi}{\alpha}$.

In the simulation we have considerered different combinations of values of the probabilities of strong links, ψ, of weak links, τ, and of the number of individuals, α, with a weak link. These combinations are illustrated in Table 5.3.

Context 3. The CV constrains for the households and farms estimates are equal to those adopted in Context 1. In Table 5.3, we derived the allocation considering the Option 3.2, proposed for Context 3. The results of Option 3.1 are presented at the end of this section.

Finally, note that for all the three contexts, the optimization problem has been set up in terms of π_j^A. With a SSRSWOR design, this may be seen as a problem of allocation for stratified sampling.

Table 5.2
Variance constraints in the different contexts

Contexts	Variance Constraints*			
	U^A: variable Animals		U^B: variable Trees	
	Province	District	Province	District
Context 0	6.5%	10%	No constraints	No constraints
Context 1	10%	15%	2.5%	5%
Context 2	10%	15%	2.5%	5%
Context 3	10%	15%	2.5%	5%

* Expressed in terms of %CV.

Table 5.3
Main results of the evaluation

Contexts		Sample size	Realized Coefficent of variations (%)						
			U^A: variable Animals				U^B: variable Trees		
			Province	District			Province	District	
				7	8	9		7	8	9
Context 0		2,122	6.5	10.0	10.0	10.0	1.5	6.8	12.7	1.4
Context 1		2,106	8.8	7.5	4.1	15.0	1.8	5.0	5.0	2.0
Context 2	$\psi = 0.90, \tau = 0.10, \alpha = 1$	2,146	8.8	7.2	4.1	15.0	2.2	5.0	5.0	2.4
	$\psi = 0.50, \tau = 0.10, \alpha = 5$	2,573	7.5	6.5	4.0	12.7	2.5	5.0	5.0	2.8
	$\psi = 0.30, \tau = 0.08, \alpha = 9$	2,767	7.0	6.4	4.0	11.9	2.5	5.0	5.0	2.8
	$\psi = 0.10, \tau = 0.09, \alpha = 9$	2,826	6.9	6.2	4.0	11.6	2.5	5.0	5.0	2.8
Context 3	Option 3.2	2,936	6.6	6.2	3.9	11.2	2.5	5.0	5.0	2.8

Looking at the main results of the evaluation, highlighted in Table 5.3, the following evidences emerge:

Context 0 vs Context 1. In the two contexts, the farm sample size is of about 2,100 farms.

- For Context 0, the expected %CVs of the farm estimates at the district level are exactly at the constraint level of 10%, defined for this context.
- In Context 1, we see that all the %CVs of the farm estimates at district level respect the constraints of 15% (defined for this context), being however considerably lower than 10% for the districts 7 and 8, showing that these districts are somewhat oversampled with respect to the target precisions. This is because in the second allocation, part of the farm sample is required to achieve the required indirect sample of households (FAO, 2014, studies this inefficiency issue in great detail).
- Considering now the precision of the estimates for the population U^B, we found that the expected sample sizes of households were approximately 5,300 records in both contexts. In Context 0, the %CVs are much higher than the desired level of 5%, being even larger than 12% in the District 8. With the sampling allocation resulting from Context 1, the desired precision of the estimates of population U^B are always respected, as well as those of population U^A, even if the constraints for these estimates have been defined larger than those adopted in Context 0.
- Thus, the integrated approach to the sampling allocation carried out in Context 1 enables control of the precision of the estimates for both populations of interest, however paying some loss in precision for the estimates for population U^A.

Context 1 vs Context 2. For the comparison between the Contexts 1 and 2, the analysis focuses upon the overall sample sizes, since the %CVs are under the constraint levels in both contexts.

- In the presence of strong links for Context 2 ($\psi = 0.90$, $\tau = 0.10$, $\alpha = 1$), there is only a small increase in the sample sizes (40 farms), while the CVs remain under the desired level of precision, altough being slightly increased for the household estimates.
- As the links become weaker, the sample sizes increase significantly. This is due to the achievement of the expected %CVs for the household estimates.
- Conversely in Context 2, the expected CVs for the farm estimates are lower than the targeted levels, suggesting that the the farms are somewhat oversampled with respect to the target levels of precision.

Context 3 vs other contexts. Having considered the Option 3.2 in Table 5.3, Context 3 may be considered as an extremal case of Context 2. Even in this case, the analysis focuses on the overall sample sizes, since all the %CVs are under the constraint levels:

- The maximization of the links uncertainty, represented by Option 3.2, causes an increase in the sample size of about 30%: from the sample size of 2,106 to that of 2,936.
- Examining Context 2, we note that we obtain results similar to those of Context 3 when the level, ψ, of the strong link is around 10%.

- Even in this case, the farms are somewhat oversampled with respect to the target levels of precision.

More detailed analysis of Context 3. Below, some more detailed analyzes are illustrated, aimed at better clarifying some aspects of the problem of sampling allocation for the integrated observation of two related populations. We explore Option 3.1 and the proportional allocation proposed in Remark 4.6 because of their practical importance. For the proportional allocation, we considered as measure of size (see Remark 4.6) the total number of employed people. The $\tilde{z}_{j,r}$ are obtained by expression (4.10). In this context, we have to define the k_r value. In order to identify a single k_r value, we exploited the data of Context 1 and first computed for each stratum the coefficent of variation of $z_{j,r}$, $\mathrm{CV}\left(z_{h,r}\right)$. Then, specific k_r values were computed at stratum level, as $k_{hr} = 1 + \left[\mathrm{CV}\left(z_{h,r}\right)\right]^2$ and finally the k_r value considered in this evaluation was obtained as a weighted mean of the k_{hr} values: $k_r = \sum_h k_{h,r} w_h$. We computed the weights w_h with two different alternatives, resulting in the two values: $k_r = 2.75$ and $k_r = 2.16$. With the first alternative, the w_h were defined proportional to the sum of the weights L_j^A at stratum level; while in the second alternative, the w_h were defined proportional to the quantity $\sqrt{\mathrm{CV}\left(z_{h,r}\right)}\,\overline{Y}_{r,h}^{B}\,N_h^A$, where $\overline{Y}_{r,h}^{B}$ and N_h^A are the mean value of variable y_r and the number of units in the stratum, respectively. For each alternative, we ran the problem (4.12), with the constraints defined in Table 5.2 for Context 1, obtaining an overall sample size, n^A, equal respectively to 1,639 and 1,517. The main results of the experiment are illustrated in Table 5.4, in which for both k_r values we show: (i) the expected %CVs, obtained as solution of problem (4.12) under the hypotesis that relation (4.11) holds; (ii) the *true* expected %CVs, that is, those obtained under Context 1 on the basis of the stratum sample sizes defined by the solution of the problem (4.12); and (iii) the true %CVs obtained, under Context 1, with the proportional allocation proposed in Remark 4.6.

Table 5.4
Expected and realized %CVs of the domain estimates of total number of trees with the sampling allocation obtained as solution of problem (4.12) and proportional allocation

Estimation Domains	$k_r = 2.75$, $n^A = 1{,}639$			$k_r = 2.16$, $n^A = 1{,}517$		
	Expected %CV, obtained as solution of problem (4.12), assuming that (4.11) holds	True expected %CV, under Context 1, with allocation defined by (4.12)	True expected %CV under Context 1, with proportional allocation	Expected %CV, obtained as solution of problem (4.12), assuming that (4.11) holds	True expected %CV, under Context 1, with allocation defined by (4.12)	True expected %CV under Context 1, with proportional allocation
Province	2.11	1.94	1.76	2.11	2.04	1.83
District 7	4.95	6.80	6.10	4.95	8.20	6.34
District 8	4.99	6.45	13.23	4.99	6.45	13.79
District 9	2.36	2.0	1.81	2.36	2.0	1.88

The main findings of this evaluation are the following:

- The strategy proposed by Option 3.1 seems to be effective, since it allows control of the sampling errors, avoiding the situation where these exceed by a large amount the desired accuracy for the different estimation domains.

- With the use of a unique k_r, the true expected %CVs (columns 3 and 7 of the Table 5.4) for some estimation domains are larger than the defined benchmarks and, in some others, the estimates are much more accurate than required.
- The choice of a larger value of the k_r parameter seems to be a safe choice, if the main objective of the sampling allocation is to avoid sampling errors in specific estimation domains that are too large.
- Even if it seems effective for the accuracy of the overall estimate at province level, the proportional allocation (columns 4 and 8 of the Table 5.4) does not allow control of extremal discrepancies from the expected accuracy in some estimation domains (see district 8).

5.2 Evaluation on costs

This evaluation considers Context 1 in which the sampling frames for both populations are available, and in which it is possible to build an integrated observation of the two populations. We focus on two observational strategies: the first considers two independent samples, one for farms and one for individuals. Therefore, a truly integrated analysis cannot be performed. The second observational strategy applies an integrated sampling design that selects a direct sample of farms and an indirect sample of the households of the workers of the sampled farms.

We adopted the variance constraints established for the Context 1 (see Table 5.5).

Table 5.5
Variance Constraints in the evaluation on costs

Variance Constraints *			
U^A: variable Animals		U^B: variable Trees	
Province	**District**	**Province**	**District**
10%	15%	2.5%	5%

* Expressed in terms of %CV.

For the direct sampling designs, we adopted a SSRSWOR design, where the population U^A was stratified by crossclassfying the districts and the size classes of the farms, and the population U^B was stratified by district. The cost for interviewing the farms varies ($C^A = 1, 2, 5$ and 10), which leads to performing four different evaluations. The cost C^B for interviewing an individual is set equal to 1.

For indirect sampling designs, we define the overall cost of interviewing the farm and the farms workers together by two different specifications of equation (3.2):

$$c_j = C^A + L_j^A C^B, \tag{5.2}$$

$$c_j = C^A + \sqrt{L_j^A}\, C^B. \tag{5.3}$$

The increase of the cost function (5.3) is lower than the increase of the cost function (5.2) when L_j^A increases.

We perform a precision-constrained optimal allocation for both independent sampling designs. The different C^A values (1, 2, 5 and 10) do not affect the farm sample size while the costs increase proportionally. Given the variance constraints in Table 5.5 with the independent strategy, the sample sizes of farms and individual are respectively 1,010 and 3,388. The total cost is then 4,398 when setting $C^A = 1$. In the integrated sample strategy, the costs do affect the allocation, essentially because if the farm interview costs increases, the number of sampled farms decreases and the allocation increases sample sizes of strata with the largest farms.

Table 5.6 below shows the sample sizes of farms and the expected sample sizes of individuals when cost model (5.2) is used to calculate the costs of individual interviews in the integrated allocation. We see that the farm sample is more than double the sample size, considering farms alone (1,101). The increase in size is due to precision constraints on the household estimates.

Table 5.6
Sample sizes for the integrated sample allocation, when the overall individual costs are given by (5.2)

Cost per farm interview (C^A)	1	2	5	10
Farms	2,388	2,289	2,190	2,137
Individuals	4,504	4,491	4,862	4,905

Table 5.7 below shows the allocation when equation (5.3) is used for the cost of individual interviews in the integrated allocation.

Table 5.7
Sample sizes for the integrated sample allocation, when the overall individual costs are given by (5.3)

Cost per farm interview (C^A)	1	2	5	10
Farms	2,135	2,121	2,111	2,108
Individuals	4,834	4,874	5,283	5,360

Tables (5.6) and (5.7) show that the integrated sample size of farms is roughly twice that of the independent allocation of farms. Thus the expected variance of the estimates will be much lower than the desidered variance constraints, suggesting that integrated sample allocation mainly depends on the variance constraints related to the individual parameters to be estimated.

Figures 5.1 and 5.2 show the cost for independent and integrated sampling. The integrated observational strategy is generally more expensive, except when the cost per farm interview is equal to 1 and the cost function given by (5.3). In this evaluation, the integrated nature of the sample is not needed as no cross tabulation of population U^A variables with population U^B variables are examined; then, the independent allocation will be more efficient in term of precision. Another cost function could however partially rebalance the two observational strategies in term of costs.

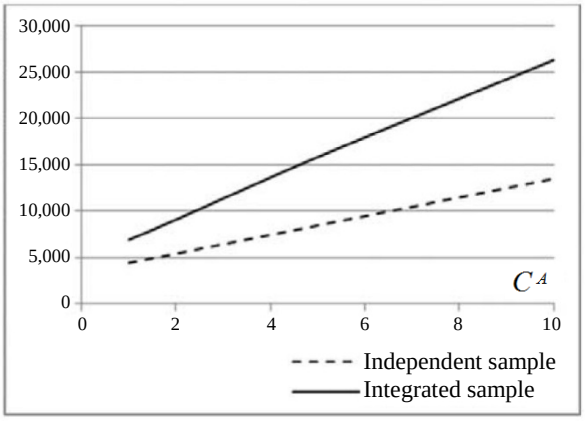

Figure 5.1 Overall costs integrated vs two independent allocations using (5.2).

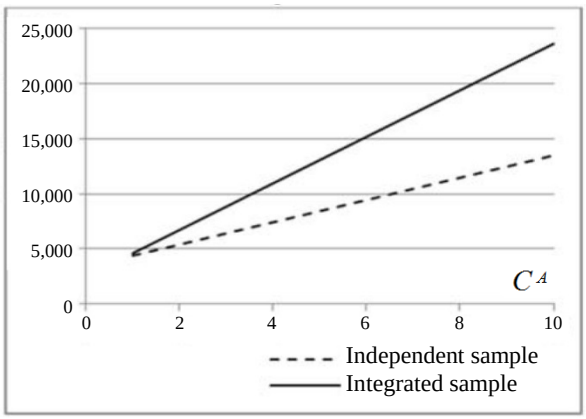

Figure 5.2 Overall costs – integrated vs two independent allocations using (5.3).

6 Conclusions

In this paper, we studied the problem of the definition of optimal sampling designs for survey strategies aiming at observing in an integrated way different statistical populations related to each other. This is particularly relevant in the agricultural sector where the integrated observation allows measurement of global phenomena that affect different statistical populations such as farms and households. The integrated observation is realized by directly sampling the first population and indirectly observing the second population, exploiting the links existing among the units of the two populations. We studied the problem considering three different contexts concerning information about the links. These range from two contexts in which the information is very rich, to the third context considering a case in which the information is very poor. The uncertainty on variables of the two populations, on links and on the z variables (built by the indirect sampling mechanism) is treated by introducing suitable superpopulation models for which expected values (of first and second order) are considered as known when launching the algorithm for the optimal sampling. Empirical studies were performed on real data of a developing country: Mozambique.

The main conclusions are summarized as follows.

Integrated vs independent observation. The integrated observation is essential to measure thoroughly global phenomena which impact on different populations. The main advantage is that it allows the cross tabulation of population U^A variables with population U^B variables. Furthermore, the integrated observation is necessary when the frame for the population U^B does not exist and an indirect sampling mechanism is needed. This is the case examined in Context 3. However, for Contexts 1 and 2, if only aggregates are examined independently from each other in the two populations, the independent allocation will be more efficient.

Cost issues. The loss in efficiency of the integrated observation can be reduced if, as assumed with cost function (5.3), the average cost of observing the elementary unit of U^B decreases when the size of the indirectly observed clusters increase. In this case, the performance of the integrated sample allocation and of the two independent allocations could be closer or similar as in the evaluation study. Nevertheless, it is complex to establish which relationship between C^A and C^B leads to two strategies with similar costs, since the allocations depend on not only on the cost of interview but also on the variability of the target parameters in the two populations and on the set of variance constraints.

Controlling the errors in the design phase. The integrated approach to allocation enables the CVs of the estimates for integrated populations to be controlled. If this is not done, the CVs of the indirectly observed population might be very high.

The impact on the uncertainty on the sample sizes. An increase in the model variances (on the variables or on the links) causes a significant increase in the sample sizes. This stresses the need of having good models for predicting the unknown variables or the links.

Appendix A

To obtain the model expectation $E_{M_v}\left(\eta_{j,v}^2\right)$, let $\boldsymbol{\eta}_v = \{\eta_{j,v}\}$ be the M^A vector of residuals, where

$$\boldsymbol{\eta}_v = \mathbf{Y}_v - \boldsymbol{\Pi}\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{Y}_v, \tag{A.1}$$

where $\mathbf{Y}_v = \{y_{j,v}\}$ denotes the M^A vector with the values of v^{th} variable of interest and $\boldsymbol{\Pi} = \text{diag}\{\pi_j^A\}$ indicates the diagonal matrix with the M^A inclusion probabilities. According to model (4.1), the vector $\mathbf{Y}_v$ may be expressed as $\mathbf{Y}_v = \tilde{\mathbf{Y}}_v + \mathbf{u}_v$, where $\tilde{\mathbf{Y}}_v = \{\tilde{y}_{i,v}\}$ and $\mathbf{u}_v = \{u_{i,v}\}$ denotes the M^A vectors of predictions and model residuals. Adopting the above matrix notation, the specific residuals $\eta_{j,v}$ can be expressed as $\eta_{j,v} = \left(\tilde{y}_{j,v} + u_{j,v}\right) - \pi_j\mathbf{d}_j'\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\left(\tilde{\mathbf{Y}}_v + \mathbf{u}_v\right)$. Therefore, the model expected values of the squared terms are given by:

$$\begin{aligned} E_{M_v}\left(\eta_{j,v}^2\right) = \tilde{y}_{j,v}^2 + \sigma_{j,v}^2 \quad & - 2\pi_j\mathbf{d}_j'\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\tilde{\mathbf{Y}}_v - 2\pi_j\mathbf{d}_j'\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{0}_{j\sigma_v} \\ & + \pi_j^2\tilde{\mathbf{Y}}_v'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{d}_j\mathbf{d}_j'\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\tilde{\mathbf{Y}}_v \\ & + \pi_j^2\boldsymbol{\sigma}_v'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{d}_j\mathbf{d}_j'\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\boldsymbol{\sigma}_v \end{aligned} \tag{A.2}$$

where $\boldsymbol{\sigma}_v = \{\sigma_{j,v}\}$ is the M^A column vector of model standard errors of the V variables and $\mathbf{0}_{j\sigma_v} = (0, \ldots, \sigma^2_{j,v}, \ldots, 0)'$ is a vector in which the j^{th} element is equal to $\sigma^2_{j,v}$ and all other elements are zeroes. Using the above matrix notation and according to Falorsi and Righi (2015), the anticipated variance can be approximated by the following expression:

$$E_{M_v}\left[V\left(\hat{\bar{Y}}_v^A \,\middle|\, \mathbf{m}^A\right)\right] = \left[M^A/(M^A - H)\right]\left[\tilde{\mathbf{Y}}_v'\boldsymbol{\Pi}^{-1}\tilde{\mathbf{Y}}_v + \boldsymbol{\sigma}_v'\boldsymbol{\Pi}^{-1}\boldsymbol{\sigma}_v - \tilde{\mathbf{Y}}_v'\tilde{\mathbf{Y}}_v + \boldsymbol{\sigma}_v'\boldsymbol{\sigma}_v + \text{AAV}_{3,v}\right].$$

Letting $\mathbf{a}_v = \mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I} - \boldsymbol{\Pi})\tilde{\mathbf{Y}}_v$, $\mathbf{b}_v = \boldsymbol{\sigma}_v'\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I} - \boldsymbol{\Pi})\,\boldsymbol{\sigma}_v$ and $\mathbf{c}_v = \boldsymbol{\sigma}_v'\text{diag}\left[\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I} - \boldsymbol{\Pi})'(\mathbf{I} - \boldsymbol{\Pi})\,\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{D}'\right]\boldsymbol{\sigma}_v$, we then have $\text{AAV}_{3,v} = \mathbf{a}_v'(\mathbf{I} - \boldsymbol{\Pi})\left(2\tilde{\mathbf{Y}}_v - \boldsymbol{\Pi}\mathbf{a}_v\right) + \mathbf{1}'(\mathbf{I} - \boldsymbol{\Pi})\left(2\mathbf{b}_v - \boldsymbol{\Pi}\mathbf{c}_v\right)$, where the scalars defined as (A.1.4), (A.1.7) and (A.1.8) in Falorsi and Righi (2015) are respectively the elements of the vectors $\mathbf{a}_v$, $\mathbf{b}_v$ and $\mathbf{c}_v$.

Appendix B

Adopting the matrix notation, the residuals $\eta_{j,r}$ can be expressed as

$$\eta_{j,r} = \mathbf{l}_j'\left(\tilde{\mathbf{Y}}_r + \mathbf{u}_r\right) - \pi_j\boldsymbol{\delta}_j'\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I} - \boldsymbol{\Pi})\,\mathbf{L}\left(\tilde{\mathbf{Y}}_r + \mathbf{u}_r\right), \tag{B.1}$$

where $\mathbf{L} = \{\tilde{L}^B_{j,i}\}$ is the $M^A \times N^B$ matrix of standardized links, and $\tilde{\mathbf{Y}}_r = \{y_{i,r}\}$ and $\mathbf{u}_r = \{u_{i,r}\}$ denote respectively the N^B vectors with the values of the predictions and of the residuals of the r^{th} variable of interest being $\mathbf{l}_j'$ is the j^{th} row of the matrix $\mathbf{L}$. Therefore, the model expected values of the squared terms is given by:

$$\begin{aligned} E_{M_r}\left(\eta^2_{j,r}\right) &= \tilde{\mathbf{Y}}_r'\mathbf{l}_j\mathbf{l}_j'\tilde{\mathbf{Y}}_r + \boldsymbol{\sigma}_r'\mathbf{l}_j\mathbf{l}_j'\boldsymbol{\sigma}_r \\ &\quad - 2\pi_j\tilde{\mathbf{Y}}_r'\mathbf{d}_j\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I} - \boldsymbol{\Pi})\,\mathbf{L}\tilde{\mathbf{Y}}_r - 2\pi_j\boldsymbol{\sigma}_r'\mathbf{d}_j\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I} - \boldsymbol{\Pi})\,\mathbf{L}\boldsymbol{\sigma}_r \\ &\quad + \pi_j^2\tilde{\mathbf{Y}}_r'\mathbf{L}'(\mathbf{I} - \boldsymbol{\Pi})\,\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{d}_j\mathbf{d}_j'\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I} - \boldsymbol{\Pi})\,\mathbf{L}\tilde{\mathbf{Y}}_r \\ &\quad + \pi_j^2\boldsymbol{\sigma}_r'\mathbf{L}'(\mathbf{I} - \boldsymbol{\Pi})\,\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{d}_j\mathbf{d}_j'\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I} - \boldsymbol{\Pi})\,\mathbf{L}\boldsymbol{\sigma}_r \end{aligned} \tag{B.2}$$

where $\boldsymbol{\sigma}_r = \{\sigma_{j,r}\}$ is the N^B column vector of model standard errors of the y_r variables. Following the above notation, we have:

$$E_{M_v}\left[V\left(\hat{\bar{Y}}_r^A \,\middle|\, \mathbf{m}^A\right)\right] = \left[M^A/(M^A - H)\right]\left[\tilde{\mathbf{Y}}_r'\mathbf{L}'\boldsymbol{\Pi}^{-1}\mathbf{L}\tilde{\mathbf{Y}}_r + \boldsymbol{\sigma}_r'\mathbf{L}'\boldsymbol{\Pi}^{-1}\mathbf{L}\boldsymbol{\sigma}_r - \tilde{\mathbf{Y}}_r'\mathbf{L}'\mathbf{L}\tilde{\mathbf{Y}}_r + \boldsymbol{\sigma}_r'\mathbf{L}'\mathbf{L}\boldsymbol{\sigma}_r + \text{AAV}_{3,r}\right].$$

Letting $\mathbf{a}_r = \mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I} - \boldsymbol{\Pi})\,\mathbf{L}\tilde{\mathbf{Y}}_r$, $\mathbf{b}_r = \boldsymbol{\sigma}_r'\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I} - \boldsymbol{\Pi})\,\boldsymbol{\sigma}_r$, $\mathbf{c}_r = \boldsymbol{\sigma}_r'\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I} - \boldsymbol{\Pi})'(\mathbf{I} - \boldsymbol{\Pi})\,\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{D}'\boldsymbol{\sigma}_r$, we have $\text{AAV}_{3,r} = \mathbf{a}_r'(\mathbf{I} - \boldsymbol{\Pi})\left(2\mathbf{L}\tilde{\mathbf{Y}}_r - \boldsymbol{\Pi}\mathbf{a}_r\right) + \mathbf{1}'(\mathbf{I} - \boldsymbol{\Pi})\left(2\mathbf{b}_r - \boldsymbol{\Pi}\mathbf{c}_r\right)$.

Finally, we note the following

$$\begin{aligned} E_{M_r}\left(z^2_{j,r}\right) &= \sum_{i=1}^{N^B}\left(\tilde{L}^B_{j,i}\right)^2\left(\tilde{y}^2_{i,r} + \sigma^2_{i,r}\right) + \sum_{i=1}^{N^B}\tilde{L}^B_{j,i}\,\tilde{y}_{i,r}\sum_{i'\neq i}\tilde{L}^B_{j,i'}\,\tilde{y}_{i',r} \\ &= \sum_{i=1}^{N^B}\left(\tilde{L}^B_{j,i}\right)^2\left(\tilde{y}^2_{i,r} + \sigma^2_{i,r}\right) + \sum_{i=1}^{N^B}\tilde{L}^B_{j,i}\,\tilde{y}_{i,r}\left(\tilde{z}_{j,r} - \tilde{L}^B_{j,i}\,\tilde{y}_{i,r}\right) \\ &= \sum_{i=1}^{N^B}\left(\tilde{L}^B_{j,i}\right)^2\left(\tilde{y}^2_{i,r} + \sigma^2_{i,r}\right) + \tilde{z}^2_{j,r} - \sum_{i=1}^{N^B}\left(\tilde{L}^B_{j,i}\right)^2\tilde{y}^2_{i,r} \\ &= \tilde{z}^2_{j,r} + \sigma^2_{j,zr}. \end{aligned}$$

Appendix C

Starting from (B.2), we have:

$$\begin{aligned} E_{M_l}E_{M_r}\left(\eta^2_{j,r}\right) &= \tilde{\mathbf{Y}}'_r E_{M_l}\left(\mathbf{l}_j\mathbf{l}'_j\right)\tilde{\mathbf{Y}}_r + \boldsymbol{\sigma}'_r E_{M_l}\left(\mathbf{l}_j\mathbf{l}'_j\right)\boldsymbol{\sigma}_r \\ &\quad - 2\pi_j \tilde{\mathbf{Y}}'_r E_{M_l}\left(\mathbf{l}_j\mathbf{d}'_j\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{L}\right)\tilde{\mathbf{Y}}_r \\ &\quad - 2\pi_j \boldsymbol{\sigma}'_r E_{M_l}\left(\mathbf{l}_j\mathbf{d}'_j\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{L}\right)\boldsymbol{\sigma}_r \\ &\quad + \pi_j^2 \tilde{\mathbf{Y}}'_r E_{M_l}\left(\mathbf{L}'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{d}_j\mathbf{d}'_j\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{L}\right)\tilde{\mathbf{Y}}_r \\ &\quad + \pi_j^2 \boldsymbol{\sigma}'_r E_{M_l}\left(\mathbf{L}'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{d}_j\mathbf{d}'_j\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{L}\right)\boldsymbol{\sigma}_r. \end{aligned} \tag{C.1}$$

The above expected value can be easily derived based on the following general result. Let $\mathbf{A} = \left\{a_{j,j'}\right\}$ be a generic $M^A \times M^A$ matrix. The generic element $g_{i,i'}$ in the position i, i' of the squared $N^B \times N^B$ matrix $\mathbf{L}'\mathbf{A}\mathbf{L} = \left\{g_{i,i'}\right\}$ is given by $g_{i,i'} = \sum_{j=1}^{M^A}\sum_{j'=1}^{M^A} \tilde{L}^B_{j,i}\,\tilde{L}^B_{j',i'}\,a_{j,j'}$.

We have $E_{M_l}\left(g_{i,i'}\right) = \sum_{j=1}^{M^A}\sum_{j'=1}^{M^A} \tilde{\Lambda}^B_{j,i}\,\tilde{\Lambda}^B_{j',i'}\,a_{j,j'} + \sum_{j=1}^{M^A}\sum_{j'=1}^{M^A} \mathrm{Cov}_{M_l}(\tilde{L}^B_{j,i}, \tilde{L}^B_{j',i'})\;a_{j,j'}$.

Taylor's series first order approximation of $\tilde{\Lambda}_{j,i}$ is given by $\tilde{L}^B_{j,i} \cong \frac{1}{\Lambda_i}\left(L^B_{j,i} - \tilde{\Lambda}^B_{j,i}\,L^B_i\right)$. Therefore, we have

$$\begin{aligned} \mathrm{Cov}_{M_l}(\tilde{L}^B_{j,i}, \tilde{L}^B_{j',i'}) &\cong \frac{1}{\Lambda^B_i}\frac{1}{\Lambda^B_{i'}}\mathrm{Cov}_{M_l}\left[\left(L^B_{j,i} - \tilde{\Lambda}^B_{j,i}L^B_i\right), \left(L^B_{j',i'} - \tilde{\Lambda}^B_{j',i'}L^B_{i'}\right)\right] \\ &= \begin{cases} \left[V_{M_l}(L^B_{j,i})(1-2\tilde{\Lambda}^B_{j,i}) + (\tilde{\Lambda}^B_{j,i})^2 V_{M_l}(L^B_i)\right]\Big/(\Lambda^B_i)^2 & \text{for } j=j' \text{ and } i=i' \\ \left[\tilde{\Lambda}^B_{j,i}\tilde{\Lambda}^B_{j,i'}V_{M_l}(L^B_i) - \tilde{\Lambda}^B_{j,i}V_{M_l}(L^B_{j,i}) - \tilde{\Lambda}^B_{j,i'}V_{M_l}(L^B_{j',i})\right]\Big/(\Lambda^B_i)^2 & \text{for } j\neq j' \text{ and } i=i' \\ 0 & \text{for } j=j' \text{ and } i\neq i' \\ 0 & \text{for } j\neq j' \text{ and } i\neq i' \end{cases} \end{aligned} \tag{C.2}$$

where

$$V_{M_l}\left(L^B_{j,i}\right) = \sum_{k=1}^{M^b_i}\lambda_{j,ik}\left(1-\lambda_{j,ik}\right),\; V_{M_l}(L^B_i) = \sum_{j=1}^{M^A} V_{M_l}\left(L^B_{j,i}\right). \tag{C.3}$$

Equation (C.2) is derived from the following result. For $i = i'$ and $j = j'$, we obtain

$$\begin{aligned} \mathrm{Cov}_{M_l}\left(\tilde{L}^B_{j,i}, \tilde{L}^B_{j,i}\right) &= V_{M_l}\left(\tilde{L}^B_{j,i}\right) \\ &\cong \frac{1}{\Lambda^2_i}\left[V_{M_l}\left(L^B_{j,i}\right) + \left(\tilde{\Lambda}^B_{j,i}\right)^2 V_{M_l}(L^B_i) - 2\tilde{\Lambda}^B_{j,i}\,\mathrm{Cov}_{M_l}\left(L^B_{j,i}, L^B_i\right)\right] \\ &= \frac{1}{\Lambda^2_i}\left[V_{M_l}\left(L^B_{j,i}\right)\left(1-2\tilde{\Lambda}^B_{j,i}\right) + \left(\tilde{\Lambda}^B_{j,i}\right)^2 V_{M_l}(L^B_i)\right] \end{aligned} \tag{C.4}$$

where $\mathrm{Cov}_{M_l}\left(L^B_{j,i}, L^B_i\right) = V_{M_l}\left(L^B_{j,i}\right)$. For $i = i'$ and $j \neq j'$, we obtain

$$\begin{aligned}\mathrm{Cov}_{M_l}\left(\tilde{L}^B_{j,i},\, \tilde{L}^B_{j',i}\right) &\cong \frac{1}{\Lambda_i^2}\mathrm{Cov}_{M_l}\left[\left(L^B_{j,i} - \tilde{\Lambda}^B_{j,i}\, L_i\right), \left(L^B_{j',i} - \tilde{\Lambda}^B_{j',i} L_i\right)\right] \\ &= \frac{1}{\Lambda_i^2} V_{M_l}\left[L^B_{j,i}L^B_{j',i} - \tilde{\Lambda}^B_{j',i}L^B_{j,i}L_i - \tilde{\Lambda}^B_{j,i}L^B_{j',i}L_i + \tilde{\Lambda}^B_{j,i}\,\tilde{\Lambda}^B_{j',i}\,L_i^2\right] \\ &= \frac{1}{\Lambda_i^2}\left[\tilde{\Lambda}^B_{j,i}\,\tilde{\Lambda}^B_{j',i} V_{M_l}\left(L^B_i\right) - \tilde{\Lambda}^B_{j,i}\, V_{M_l}\left(L^B_{j,i}\right) - \tilde{\Lambda}^B_{j',i}\, V_{M_l}\left(L^B_{j',i}\right)\right]. \qquad \text{(C.5)}\end{aligned}$$

Let $\mathbf{a} = \{a_j\}$ be a generic M^A vector. The generic element ${}^j g_{i,i'}$ in the position i, i' of the squared $N^B \times N^B$ matrix $\mathbf{l}_j\mathbf{a}'\mathbf{L} = \left\{{}^j g_{i,i'}\right\}$ is given by ${}^j g_{i,i'} = \sum_{j'=1}^{M^A} \tilde{L}^B_{j,i}\,\tilde{L}^B_{j',i'}\,a_{j'}$, where $E_{M_l}\left({}^j g_{i,i'}\right) = \sum_{j'=1}^{M^A} \tilde{\Lambda}^B_{j,i}\,\tilde{\Lambda}^B_{j',i'}\,a_{j'} + \sum_{j'=1}^{M^A} \mathrm{Cov}_{M_l}\left(\tilde{L}^B_{j,i},\, \tilde{L}^B_{j',i'}\right)\, a_{j'}$.

Finally, denote by $\left\{{}^{jj} g_{i,i'}\right\}$ the generic element in the i, i'^{th} position of the matrix $\mathbf{l}_j\mathbf{l}'_j$. Its generic expected value is given by $E_{M_l}\left({}^{jj} g_{i,i'}\right) = \tilde{\Lambda}^B_{j,i}\,\tilde{\Lambda}^B_{j',i'} + \mathrm{Cov}_{M_l}\left(\tilde{L}^B_{j,i},\, \tilde{L}^B_{j',i'}\right)$.

References

Bethel, J. (1989). Sample allocation in multivariate surveys. *Survey Methodology*, 15, 1, 47-57. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1989001/article/14578-eng.pdf.

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

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

Choudhry, G.H., Rao, J.N.K. and Hidiroglou, M.A. (2012). On sample allocation for efficient domain estimation. *Survey Methodology*, 38, 1, 23-29. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2012001/article/11682-eng.pdf.

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*. New York: John Wiley & Sons, Inc.

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

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, 1, 215-236. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2015001/article/14149-eng.pdf.

FAO (2014). *Technical Report on the Integrated Survey Framework*, Technical Report Series GO-02-2014. http://gsars.org/wp-content/uploads/2014/07/Technical_report_on-ISF-Final.pdf.

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

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.

Lavallée, P. (2002). Le sondage indirect, ou la méthode du partage des poids. Éditions de l'Université de Bruxelles (Belgique) et Éditions Ellipses (France), 215 pages.

Lavallée, P. (2007). *Indirect Sampling.* New York: Springer.

Lavallée, P., and Caron, P. (2001). Estimation using the generalised weight share method: The case of record linkage. *Survey Methodology*, 27, 2, 155-169. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2001002/article/6092-eng.pdf.

Lavallée, P., and Labelle-Blanchet, S. (2013). Indirect sampling applied to skewed populations. *Survey Methodology*, 39, 1, 183-215. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2013001/article/11829-eng.pdf.

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

Steel, D.G., and Clark, R.G. (2014). Potential gains from using unit level cost information in a model-assisted framework. *Survey Methodology*, 40, 2, 231-242. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2014002/article/14110-eng.pdf.

Wallgren, A., and Wallgren, B. (2014). *Register-Based Statistics: Administrative Data for Statistical Purposes*. New York: John Wiley & Sons, Inc. Chichester, UK. ISBN: ISBN 978-1-119-94213-9.

Winkler, W.E. (2001). *Multi-Way Survey Stratification and Sampling*. Research Report Series, Statistics #2001-01. Statistical Research Division U.S. Bureau of the Census Washington D.C. 20233.

Xu, X., and Lavallée, P. (2009). Treatments for link nonresponse in indirect sampling. *Survey Methodology*, 35, 2, 153-164. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2009002/article/11038-eng.pdf.