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

Survey Methodology

An extension of the weight share method when using a continuous sampling frame

by Guillaume Chauvet, Olivier Bouriaud and Philippe Brion

Release date: June 30, 2023

Statistics Canada Statistique Canada

Canadä

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 infostats@statcan.gc.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

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, June 2023
Vol. 49, No. 1, pp. 139-162
Statistics Canada, Catalogue No. 12-001-X

An extension of the weight share method when using a continuous sampling frame

Guillaume Chauvet, Olivier Bouriaud and Philippe Brion[1]

Abstract

The definition of statistical units is a recurring issue in the domain of sample surveys. Indeed, not all the populations surveyed have a readily available sampling frame. For some populations, the sampled units are distinct from the observation units and producing estimates on the population of interest raises complex questions, which can be addressed by using the weight share method (Deville and Lavallée, 2006). However, the two populations considered in this approach are discrete. In some fields of study, the sampled population is continuous: this is for example the case of forest inventories for which, frequently, the trees surveyed are those located on plots of which the centers are points randomly drawn in a given area. The production of statistical estimates from the sample of trees surveyed poses methodological difficulties, as do the associated variance calculations. The purpose of this paper is to generalize the weight share method to the continuous (sampled population) – discrete (surveyed population) case, from the extension proposed by Cordy (1993) of the Horvitz-Thompson estimator for drawing points carried out in a continuous universe.

Key Words: Continuous sampling design; Environmental statistics; Forest inventory; Synthetic variable; Variance estimation.

1. Introduction

The definition of statistical units is a recurring issue in the domain of sample surveys. Indeed, not all the populations surveyed have a readily available sampling frame. For these populations, the sampled units (for which a sampling frame is available from which to select units according to a given sampling design), are distinct from the observation units, which constitute the population of interest on which we are willing to infer.

This issue has been raised for a long time for studying populations that are difficult to reach, e.g., homeless people (see for example Ardilly and Le Blanc, 2001; De Vitiis, Falorsi and Inglese, 2014; Laporte, Vandentorren, Détrez, Douay, Le Strat, Le Méner, Chauvin and The Samenta Research Group, 2018), or nomad/non-localized populations (see for example Lohlé-Tart, Clairin, François and Gendreau, 1988; Clairin and Brion, 1996; Himelein, Eckman and Murray, 2014). It has also become recently more and more accurate for business statistics, with the use of a unit "enterprise" not necessarily equivalent to the unit available in the business registers (Lorenc, Smith and Bavdaž, 2018). In this case, producing estimates on the population of interest raises complex questions, linked to the fact that the weights of the observation units need to be based on the design weights of the units selected in the sampling frame.

To deal with this issue, Deville and Lavallée (2006) proposed the so-called weight share method. It is based on a principle of duality between the sampled population and the observed population, where a variable of interest defined on the observed population may be written as a synthetic variable defined on

1. Guillaume Chauvet, Ensai (Irmar), Campus de Ker Lann, Bruz, France. E-mail: guillaume.chauvet@ensai.fr; Olivier Bouriaud, Université Stefan cel Mare de Suceava, 13 rue de l'Université, 720229 Suceava, Roumania and IGN, Laboratoire d'Inventaire Forestier, 14 rue Girardet, 54000 Nancy, France. E-mail: obouriaud.lif@gmail.com; Philippe Brion, Irmar, France. E-Mail: philippe.brion55@gmail.com.

the sampling frame (see also Lavallée, 2009). Because it creates a link between the observation units and the sampling units, this method enables the properties of the sampling design to be used to define unbiased estimators of totals for the observed populations, and to derive variance formulas. In particular, the sampling weights of the sampling units are used to assign estimation weights to the observation units. This paper deals with the extension of this method to the case when the sampled population is a continuous frame. For the sampled population and the observed population, we will use the notations U^A and U^B in case of discrete populations, and $\mathcal{U}^A$ and $\mathcal{U}^B$ in case of continuous populations.

We are particularly interested in applications encountered in forest inventories, in which it is common practice to use a sample of points selected in a continuum and then fixed-shape supports defined from these points to perform the survey on a discrete population of trees. The approach which consists of transporting a variable from the discrete population to the continuous population is not new, and has been considered by Stevens and Urquhart (2000), Gregoire and Valentine (2007) and Mandallaz (2007), for example. While these previous works were quite similar in their overall approach of the indirect sampling, the link between the units from the population sampled and the units of the target population are only implicit.

The work by Stevens and Urquhart (2000) is very similar to ours. They studied the situation when a finite population of interest is linked to a continuous territory, and they considered a way to transfer a variable of interest onto the continuous sampling frame. This is similar to the synthetic variable that we present in equation (3.15). They derived a so-called "aggregation-unbiasedness" requirement, in order to obtain unbiased total estimators. They also proposed a Horvitz-Thompson variance estimator, making use of the theory by Cordy (1993). Despite the importance of this paper, it did not have a significant impact in the literature. It is in particular telling that the article is not quoted in textbooks like Gregoire and Valentine (2007) and Mandallaz (2007). Therefore, we feel the need for a simple presentation of the approach, where we clearly state what are the estimation weights, the resulting estimators of totals for the finite population, and the associated Horvitz-Thompson variance estimators. The weight share method is a very useful and simple tool for this, as illustrated in the applications considered in Section 3.

In natural populations such as forest trees, the units are distributed spatially over a territory. Estimating totals of any given attribute of this population requires undergoing spatial sampling. To this end, Gregoire and Valentine (2007, Chapter 10) introduce the so-called Monte Carlo integration approach, and call the synthetic variable the "attribute density". Several examples are given for devices used in the practice of forest inventory (e.g., point relascope sampling, line intersect sampling). However, the link with Cordy's set-up is not pointed out, and variance estimation is restricted to the case when the points are selected by independent uniform sampling. Mandallaz (2007, Section 4.2) also develops a related approach, where the link between the observed, finite population and the sampled, continuous population is performed by means of the so-called "local density". The approach is first presented for plot sampling, and then extended to more complex situations like cluster (trakt) sampling, which is popular for forest inventories (Lawrence, McRoberts, Tomppo, Gschwantner and Gabler, 2010). However, the method is quite ad-hoc, since the local density needs to be computed differently in each situation. On the other hand, the weight share method enables one to produce general formulas for both point estimators and variance estimators. In particular, we consider in Section 3.4 the situation of spatial cluster sampling for forest inventories, for which the weight share method provides a general solution for estimation and variance estimation under an arbitrary continuous sampling design.

In what follows, we first recall in Section 2 the basic principles of the weight share method in the case of two discrete populations U^A and U^B. In Section 3, we extend the method to cover the case when a continuous population $\mathcal{U}^A$ is sampled, and we want to infer on a discrete population U^B. The results of two simulation studies are presented in Section 4. We conclude in Section 5.

2. Sampling in a discrete population

In this section, we first define in Section 2.1 our notations when sampling in a discrete population U^A. We then recall in Section 2.2 how the weight share method may be used to produce estimates in another discrete population U^B linked to U^A. A simple example is presented in Section 2.3 for illustration.

2.1 Notations

We are interested in a discrete population U^A, for which the units in the population are enumerable and a sampling frame may therefore be available. For example, this may be a population of households or individuals in social surveys, or a register in business surveys. The size of the population U^A is denoted as N^A. Suppose that we are interested in a variable of interest y^A taking the value y_i^A for unit $i \in U^A$, and that we wish to estimate the population total

$$\tau_y^A = \sum_{i \in U^A} y_i^A. \tag{2.1}$$

A random sample S^A is selected in U^A by means of a sampling design $p^A(\cdot)$, and we let s^A denote a possible realization of S^A. The Horvitz-Thompson (HT) estimator is

$$\hat{\tau}_y^A = \sum_{i \in S^A} d_i^A y_i^A, \tag{2.2}$$

where $d_i^A = 1/\pi_i^A$ is the design weight of unit i, and π_i^A the probability for unit i of inclusion in the sample. This estimator is design-unbiased for τ_y^A, provided that all the π_i^A's are > 0.

The variance of $\hat{\tau}_y^A$ is

$$V_p\left(\hat{\tau}_y^A\right) = \sum_{i,j \in U^A} \frac{y_i^A}{\pi_i^A} \frac{y_j^A}{\pi_j^A} \left(\pi_{ij}^A - \pi_i^A \pi_j^A\right), \tag{2.3}$$

where π_{ij}^A is the probability that the units i and j are jointly selected in S^A. If all the π_{ij}^A's are positive, this variance is unbiasedly estimated by

$$\hat{V}^A\left(\hat{\tau}_y^A\right) = \sum_{i,j \in S^A} \frac{y_i^A}{\pi_i^A} \frac{y_j^A}{\pi_j^A} \left(\frac{\pi_{ij}^A - \pi_i^A \pi_j^A}{\pi_{ij}^A}\right). \tag{2.4}$$

2.2 Weight share method

Suppose that we are interested in another population U^B, with a variable of interest y^B taking the value y_k^B for unit $k \in U^B$. We wish to estimate the population total

$$\tau_y^B = \sum_{k \in U^B} y_k^B. \tag{2.5}$$

We suppose that no sampling frame is available for U^B, but that this population is linked to the population U^A. The link between the units in U^A and U^B is represented by the indicator variables

$$L^{AB}(i,k) = \begin{cases} 1 & \text{if units } i \in U^A \text{ and } k \in U^B \text{ are linked,} \\ 0 & \text{otherwise.} \end{cases} \tag{2.6}$$

The set of *ancestors* for some unit $k \in U^B$ is $\text{Anc}_k = \left\{i \in U^A;\ L^{AB}(i,k) = 1\right\}$. The set of *descendants* for some unit $i \in U^A$ is $\text{Des}_i = \left\{k \in U^B;\ L^{AB}(i,k) = 1\right\}$. For any unit $k \in U^B$,

$$N_{+k}^{AB} = \sum_{i \in U^A} L^{AB}(i,k) \tag{2.7}$$

is the total number of ancestors. It is required that any unit $k \in U^B$ be linked to at least one unit in U^A; that is, we suppose that $N_{+k}^{AB} > 0$ for any unit $k \in U^B$.

A sample S^B is obtained in U^B by surveying all the descendants of the units i selected in S^A. More formally, we have

$$S^B = \bigcup_{i \in S^A} \text{Des}_i. \tag{2.8}$$

To obtain an estimator of τ_y^B, the weight share method (Deville and Lavallée, 2006) makes use of a principle of duality between populations U^A and U^B, based on the link function given in (2.6). The total τ_y^B may be written as

$$\tau_y^B = \sum_{i \in U^A} y_i^A \quad \text{with} \quad y_i^A = \sum_{k \in U^B} \frac{L^{AB}(i,k)\, y_k^B}{N_{+k}^{AB}}, \tag{2.9}$$

see Deville and Lavallée (2006, Result 2). Equation (2.9) represents the fact that the variable y_k^B may be distributed over the units in U^A to obtain a synthetic variable y_i^A. This is done by sharing each value y_k^B equally among the ancestors in Anc_k.

From equation (2.9), the total τ_y^B can be unbiasedly estimated by performing HT-estimation on the sample S^A, with the synthetic variable y_i^A. This HT-estimator may be rewritten as

$$\hat{\tau}_y^B = \sum_{i \in S^A} d_i^A y_i^A = \sum_{k \in S^B} w_k^B y_k^B, \tag{2.10}$$

with

$$w_k^B = \frac{1}{N_{+k}^{AB}} \sum_{i \in S^A} L^{AB}(i, k) \, d_i^A,$$

see Deville and Lavallée (2006, Result 3). Each unit $k \in S^B$ is given the sum of the weights of the sampled units $i \in S^A$ which are linked to k, divided by the number of links N_{+k}^{AB}. The weights d_i^A of the units $i \in S^A$ are therefore shared among the units $k \in S^B$, hence the name of the method. It is important to note that the weights w_k^B can only be computed if the number of ancestors N_{+k}^{AB} is known for any unit $k \in S^B$. Therefore, this information needs to be collected during the survey. In some situations, it may be difficult or even impossible to state whether or not a unit in U^A is related to another unit in U^B. This is referred to as link nonresponse by Xu and Lavallée (2009), who propose treatment methods to handle this problem.

From equation (2.10), the weight share method enables one to attribute to each unit $k \in S^B$ a weight w_k^B, which is usable for any variable of interest y_k^B and such that the estimator $\hat{\tau}_y^B$ is unbiased. This is a very strong property. In contrast, the HT-estimator for the sample S^B cannot be computed. The inclusion probability of unit k in the sample S^B is

$$\Pr\left(k \in S^B\right) = \sum_{\substack{s^A \subset U^A \\ s^A \cap \text{Anc}_k \neq \varnothing}} p^A\left(s^A\right). \tag{2.11}$$

Computing these inclusion probabilities would require a full specification of both the sampling design $p^A(\cdot)$ and of the links between both populations, which is usually impossible.

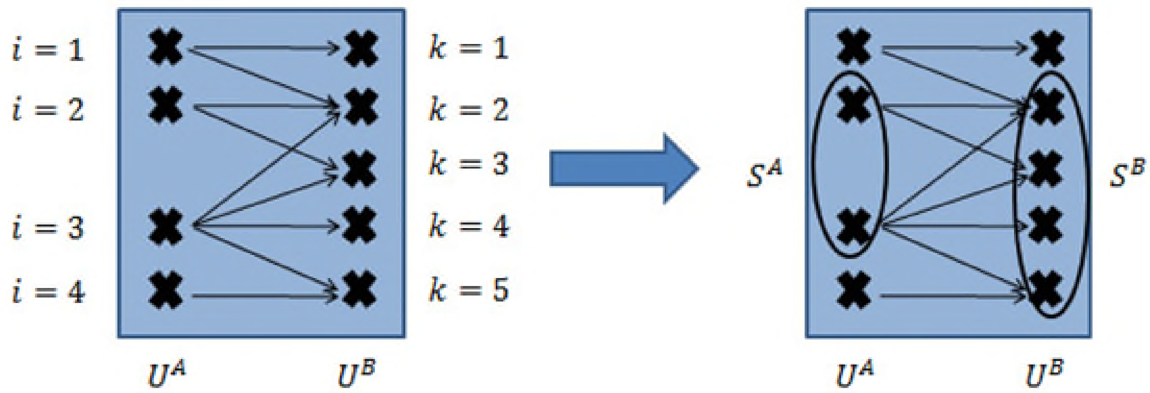
Since $\hat{\tau}_y^B$ may be written as a HT-estimator on the sample S^A, the variance of $\hat{\tau}_y^B$ is given by equation (2.3), with y_i^A the synthetic variable given in (2.9), and a variance estimator is given by (2.4). Note that the variable y_i^A can be exactly computed for any unit $i \in S^A$, since it is assumed that all the units in U^B linked to the units in S^A are surveyed, see equation (2.8).

2.3 A simple example

For illustration, we present a toy example in Figure 2.1. The population U^A contains $N^A = 4$ units, and the population U^B contains $N^B = 5$ units. The links between units are represented by the arrows. For example, unit $i = 3$ in U^A has four descendants, namely units $k = 2, 3, 4$ and 5. Therefore, we have $L^{AB}(3,1) = 0$ and $L^{AB}(3,2) = L^{AB}(3,3) = L^{AB}(3,4) = L^{AB}(3,5) = 1$. Unit $k = 4$ has a single ancestor $i = 3$, whereas unit $k = 5$ has two ancestors $i = 3$ and $i = 4$. Suppose that the sampling design leads to the selection of the subset $s^A = \{2, 3\}$. Then all the descendants of the units in s^A are selected, resulting in the observation of the subset

$$s^B = \{2, 3, 4, 5\}. \tag{2.12}$$

Figure 2.1 A simple example of links between two discrete populations U^A and U^B.



Now, suppose that the sampled values are

$$y_2^B = 1, \quad y_3^B = 3, \quad y_4^B = 3, \quad y_5^B = 5. \tag{2.13}$$

We first compute the synthetic variable y_i^A for the units $i \in s^A$, making use of equation (2.9). We obtain

$$\begin{aligned} y_2^A &= \frac{y_2^B}{3} + \frac{y_3^B}{2} \simeq 1.83, \\ y_3^A &= \frac{y_2^B}{3} + \frac{y_3^B}{2} + \frac{y_4^B}{1} + \frac{y_5^B}{2} \simeq 7.33. \end{aligned}$$

Now, we compute the weights w_k^B for the units $k \in s^B$ by means of the weight share method, making use of equation (2.10). Suppose that the sample S^A is selected in U^A by simple random sampling without replacement, leading to $\pi_i^A = 0.5$ and $d_i^A = 2$ for any $i \in U^A$. We obtain

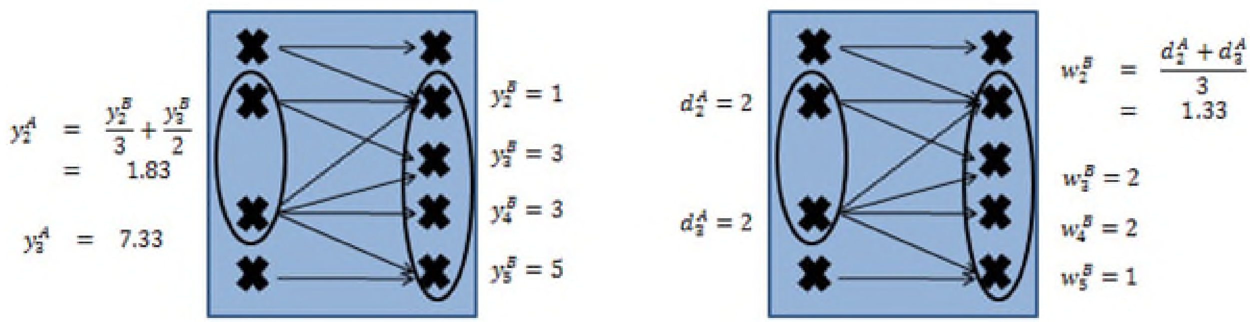
$$\begin{aligned} w_2^B &= \frac{d_2^A + d_3^A}{3} \simeq 1.33, \\ w_3^B &= \frac{d_2^A + d_3^A}{2} = 2, \\ w_4^B &= \frac{d_3^A}{1} = 2, \\ w_5^B &= \frac{d_3^A}{2} = 1. \end{aligned}$$

The estimate for the total τ_y^B is therefore

$$\hat{\tau}_y^B\left(s^B\right) = \sum_{k \in s^B} w_k^B y_k^B \simeq 18.33.$$

The results are summarized in Figure 2.2.

Figure 2.2 Computation of the synthetic variable y_i^A and of the weights d_k^B obtained by the weight share method on a simple example.



3. Sampling in a continuous population

In this section, we first define in Section 3.1 our notations when sampling in a continuous universe $\mathcal{U}^A$, following Cordy (1993). We explain in Section 3.2 how the weight share method may be extended to produce estimates in some discrete population U^B linked to $\mathcal{U}^A$. We consider applications to sampling designs used in forest inventories, for which some analog approaches have been proposed in the literature. The case of direct plot sampling is first considered in Section 3.3. The application to cluster sampling is considered in Section 3.4.

3.1 Notations

We first review the set-up for sampling and estimation in a continuous universe $\mathcal{U}^A$ introduced by Cordy (1993). This framework was strongly motivated by environmental applications, and led to an extension of the HT-estimator existing for a discrete population, see Section 2.1.

Suppose that we are interested in a continuous population or universe, i.e., as defined by Gregoire and Valentine (2007, page 93) "that does not naturally divide into smaller discrete units". This may be a landscape or a lake, for example. We suppose that the universe $\mathcal{U}^A$ is included in $\mathbf{R}^q$ with $q \geq 1$. We are interested in some Lebesgue integrable function $y^A : \mathcal{U}^A \rightarrow \mathbf{R}$, and we wish to estimate the total (integral)

$$\tau_y^A = \int_{\mathcal{U}^A} y^A(x)\, dx \tag{3.1}$$

of this function over $\mathcal{U}^A$.

A random sample $S^A = \left\{S_1^A, \ldots, S_{n^A}^A\right\}$ of n^A locations is selected from $\mathcal{U}^A$, and we let $s^A = \left\{s_1^A, \ldots, s_{n^A}^A\right\}$ denote a possible realization of S^A. We assume the existence of the joint probability density function (PDF)

$$f\left(s_1^A, \ldots, s_{n^A}^A\right) \tag{3.2}$$

of the sample locations, along with the existence of the marginal PDF and of the joint PDF

$$f_i\left(s_i^A\right) \quad \text{and} \quad f_{ij}\left(s_i^A, s_j^A\right) \tag{3.3}$$

of s_i^A and s_j^A for $i \neq j$. For example, if the sample S^A is obtained by n^A independent selections of some point, performed uniformly in $\mathcal{U}^A$, we have

$$f_i\left(s_i^A\right) = \frac{1\left(s_i^A \in \mathcal{U}^A\right)}{M^A} \tag{3.4}$$

and $f\left(s_1^A, \ldots, s_{n^A}^A\right) = \prod_{i=1}^{n^A} f_i\left(s_i^A\right)$, with $M^A = \int_{\mathcal{U}^A} dx$ the global measure of the universe, and with $1(\cdot)$ an indicator function.

Suppose that the PDF is absolutely continuous with respect to the Lebesgue measure. As noted by Cordy (1993) and Stevens (1997), there are sampling designs used in practice such that this assumption does not hold true, such as systematic sampling, for example. For any point $x \in \mathcal{U}^A$, the inclusion density function is defined by

$$\pi^A(x) = \sum_{i=1}^{n^A} f_i(x). \tag{3.5}$$

This may be seen as a local measure of the number of sampled points by unit of measure. We have in particular $\int_{\mathcal{U}^A} \pi^A(x)\, dx = n^A$, which is a usual property for sampling designs of fixed size n^A. Similarly, the joint inclusion density function is defined by

$$\pi^A\left(x, x'\right) = \sum_{i=1}^{n^A} \sum_{\substack{j=1 \\ j \neq i}}^{n^A} f_{ij}\left(x, x'\right) \tag{3.6}$$

for $x, x' \in \mathcal{U}^A$.

The HT-estimator of τ_y^A is

$$\hat{\tau}_y^A = \sum_{s \in S^A} d^A(s)\, y^A(s), \tag{3.7}$$

with $d^A(x) = 1/\pi^A(x)$ the design weight of some point x. This estimator is unbiased for τ_y, provided that $\pi^A(x) > 0$ almost everywhere, see Cordy (1993, Theorem 1).

If the function $y^A(\cdot)$ is bounded and $\int_{\mathcal{U}^A} \left\{1/\pi^A(x)\right\} dx < \infty$, then the variance of $\hat{\tau}_y^A$ is given by the Horvitz-Thompson formula

$$\begin{aligned} V\left(\hat{\tau}_y^A\right) &= \int_{x \in \mathcal{U}^A} \frac{\left\{y^A(x)\right\}^2}{\pi^A(x)} dx \\ &+ \int_{x \in \mathcal{U}^A} \int_{x' \in \mathcal{U}^A} \left\{\pi^A\left(x, x'\right) - \pi^A(x)\, \pi^A\left(x'\right)\right\} \frac{y^A(x)}{\pi^A(x)} \frac{y^A\left(x'\right)}{\pi^A\left(x'\right)} dx dx', \end{aligned} \tag{3.8}$$

or equivalently by the Sen-Yates-Grundy formula

$$V\left(\hat{\tau}_y^A\right) = \frac{1}{2}\int_{x\in\mathcal{U}^A}\int_{x'\in\mathcal{U}^A}\left\{\pi^A(x)\,\pi^A(x') - \pi^A\left(x, x'\right)\right\}\left\{\frac{y^A(x)}{\pi^A(x)} - \frac{y^A(x')}{\pi^A(x')}\right\}^2 dxdx', \tag{3.9}$$

see Cordy (1993, Section 2). The corresponding variance estimators are respectively

$$\begin{aligned}\hat{V}_{\mathrm{HT}}\left(\hat{\tau}_y^A\right) = & \sum_{s\in S^A}\left\{\frac{y^A(s)}{\pi^A(s)}\right\}^2 \\ & + \sum_{s\in S^A}\sum_{\substack{s'\in S^A \\ s'\neq s}}\left\{\frac{\pi^A\left(s, s'\right) - \pi^A(s)\,\pi^A(s')}{\pi^A\left(s, s'\right)}\right\}\frac{y^A(s)}{\pi^A(s)}\frac{y^A(s')}{\pi^A(s')}\end{aligned} \tag{3.10}$$

and

$$\hat{V}_{\mathrm{YG}}\left(\hat{\tau}_y^A\right) = \frac{1}{2}\sum_{s\in S^A}\sum_{\substack{s'\in S^A \\ s'\neq s}}\left\{\frac{\pi^A(s)\,\pi^A(s') - \pi^A\left(s, s'\right)}{\pi^A\left(s, s'\right)}\right\}\left\{\frac{y^A(s)}{\pi^A(s)} - \frac{y^A(s')}{\pi^A(s')}\right\}^2. \tag{3.11}$$

If in addition both $\pi^A(x) > 0$ and $\pi^A\left(x, x'\right) > 0$ almost everywhere on $\mathcal{U}^A$, then these two variance estimators are unbiased, see Cordy (1993, Theorem 2). Note that the condition on $\pi^A\left(x, x'\right)$ may not be true, for example when using systematic sampling designs.

3.2 Weight share method: the continuous-discrete case

Suppose that we are still interested in the population U^B and in the estimation of the total τ_y^B given in equation (2.5). The units in U^B are not directly sampled, but a continuous universe $\mathcal{U}^A$ linked to U^B is sampled instead. The links between the units inside the populations $\mathcal{U}^A$ and U^B are represented by the indicator function

$$L^{AB}\left(x, k\right) = \begin{cases} 1 & \text{if } x\in\mathcal{U}^A \text{ and } k\in U^B \text{ are linked,} \\ 0 & \text{otherwise.}\end{cases} \tag{3.12}$$

We keep the same terminology as in Section 2.2, and we note Anc_k for the *ancestor subset* of some $k\in U^B$, and $\mathrm{Des}(x)$ for the *descendant subset* of some $x\in\mathcal{U}^A$. For any $k\in U^B$,

$$M_{+k}^{AB} = \int_{x\in\mathcal{U}^A} L^{AB}\left(x, k\right) dx \tag{3.13}$$

is the measure of the ancestor subset of unit k. As in the discrete case, we suppose that $M_{+k}^{AB} > 0$ for any $k\in U^B$.

A sample S^B is obtained in U^B by surveying all the descendants of the points selected in S^A. Formally, we have therefore

$$S^B = \bigcup_{s\in S^A}\mathrm{Des}(s). \tag{3.14}$$

To obtain an estimator of $\tau_y^B,$ we establish a duality principle between populations $\mathcal{U}^A$ and $U^B.$ This is summarized in Proposition 1. The duality principle is similar to that obtained with a discrete population: each value y_k^B is equally shared among the points in the ancestor subset of $k.$ The synthetic function $y^A(x)$ may therefore be interpreted as a local measure of density of the variable y^B per unit area. This approach has already been considered in the domain of forest inventory by Mandallaz (2007, Section 4.2) and Gregoire and Valentine (2007, Chapter 10), for example, as discussed in the introduction.

Proposition 1. *The total* τ_y^B *may be written as*

$$\tau_y^B = \int_{x \in \mathcal{U}^A} y^A(x)\, dx \tag{3.15}$$

with

$$y^A(x) = \sum_{k \in U^B} \frac{L^{AB}(x, k)\, y_k^B}{M_{+k}^{AB}}.$$

Proof. *We have*

$$\begin{aligned} \tau_y^B &= \sum_{k \in U^B} y_k^B = \sum_{k \in U^B} y_k^B \times \frac{1}{M_{+k}^{AB}} \int_{x \in \mathcal{U}^A} L^{AB}(x, k)\, dx \\ &= \int_{x \in \mathcal{U}^A} \sum_{k \in U^B} \frac{L^{AB}(x, k)\, y_k^B}{M_{+k}^{AB}}\, dx = \int_{x \in \mathcal{U}^A} y^A(x)\, dx. \end{aligned}$$

Proposition 1 makes it possible to rewrite τ_y^B as an integral over the universe $\mathcal{U}^A,$ and therefore to make use of the extended HT-estimator given in (3.7). In turn, this estimator may be written as a weighted sum over the sample $S^B.$ This is summarized in Proposition 2.

Proposition 2. *The total* τ_y^B *may be unbiasedly estimated by*

$$\hat{\tau}_y^B = \sum_{k \in S^B} w_k^B y_k^B = \sum_{s \in S^A} d^A(s)\, y^A(s) \tag{3.16}$$

with

$$w_k^B = \frac{1}{M_{+k}^{AB}} \sum_{s \in S^A} L^{AB}(s, k)\, d^A(s).$$

Proof. *We can rewrite*

$$\begin{aligned} \hat{\tau}_y^B &= \sum_{s \in S^A} \sum_{k \in S^B} \frac{L^{AB}(s, k)\, y_k^B d^A(s)}{M_{+k}^{AB}} \\ &= \sum_{s \in S^A} \sum_{k \in U^B} \frac{L^{AB}(s, k)\, y_k^B d^A(s)}{M_{+k}^{AB}} \end{aligned}$$

where the last equality follows from the fact that if $s \in S^A$, *all the units* k *in its descendant subset* $\mathrm{Des}(x)$ *are selected in* S^B. *It follows that*

$$\hat{\tau}_y^B = \sum_{s \in S^A} d^A(s) \sum_{k \in U^B} \frac{L^{AB}(s,k)\, y_k^B}{M_{+k}^{AB}} = \sum_{s \in S^A} d^A(s)\, y^A(s),$$

which is simply the HT-estimator of the integral $\int_{x \in \mathcal{U}^A} y^A(x)\, dx$.

The weight share method thus brings a solution to the estimation of τ_y^B, by using a weighted estimator computed on the sample S^B where the weights w_k^B are given in equation (3.16). Each unit $k \in S^B$ is given the sum of the weights of the sampled points $s \in S^A$ which are linked to k, divided by M_{+k}^{AB}, the measure of the ancestor subset of unit k. The principle is therefore the same as with the usual weight share method applied to discrete populations. It is important to note that, for any unit $k \in S^B$, we need to know the measure M_{+k}^{AB} of its ancestor subset.

Since $\hat{\tau}_y^B$ may be written as a HT-estimator on the sample S^A, the variance is obtained from equation (3.8) or equation (3.9), by using the synthetic variable $y^A(x)$ given in equation (3.15). A variance estimator can be obtained by applying equation (3.10) or equation (3.11). Therefore, variance estimation is straightforward for the weight share estimator.

3.3 Application to plot sampling for forest inventories

We first consider an application of the weight share method to the case where a forest inventory is performed by direct plot sampling. We are interested in a population U^B of trees located on a territory $\mathcal{U}^A$. A sample of points S^A is first selected in $\mathcal{U}^A$ by using a continuous sampling design. For each point $s \in S^A$, the circle $C_r(s)$ centered on s with some predetermined radius r is drawn. All the trees $k \in U^B$ such that their center x_k is inside these circles are selected in the sample S^B and surveyed.

The link function is therefore

$$L^{AB}(x,k) = \begin{cases} 1 & \text{if } x_k \in C_r(x), \\ 0 & \text{otherwise,} \end{cases} = \begin{cases} 1 & \text{if } x \in C_r(x_k), \\ 0 & \text{otherwise.} \end{cases} \tag{3.17}$$

For any tree $k \in U^B$, the quantity M_{+k}^{AB} is

$$M_{+k}^{AB} = \int_{x \in \mathcal{U}^A} L^{AB}(x,k)\, dx = \int_{x \in \mathcal{U}^A} 1\{x \in C_r(x_k)\}\, dx, \tag{3.18}$$

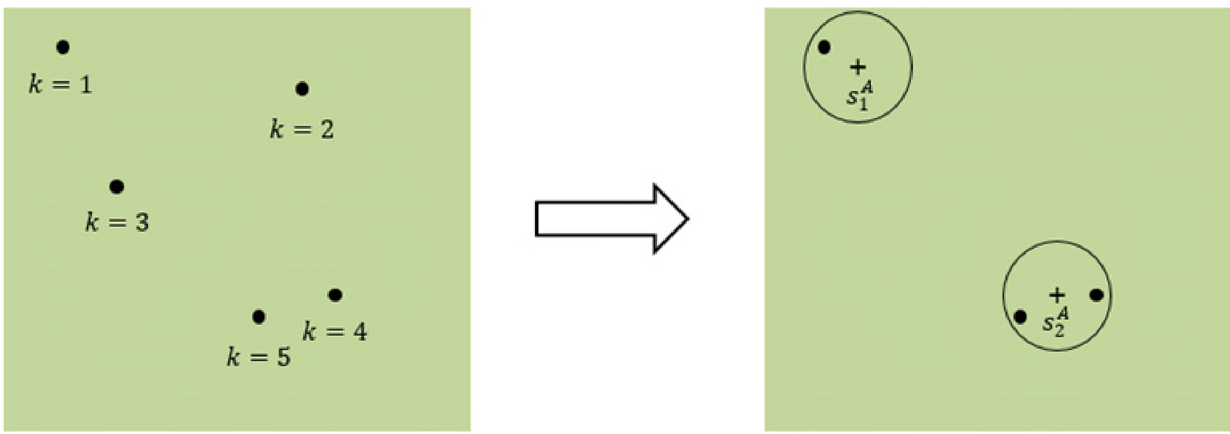
which is the area of the intersection between $\mathcal{U}^A$ and the circle centered on x_k. For any point $x \in \mathcal{U}^A$, the synthetic variable is

$$y^A(x) = \sum_{k \in U^B} \frac{L^{AB}(x,k)\, y_k^B}{M_{+k}^{AB}} = \sum_{x_k \in C_r(x)} \frac{y_k^B}{M_{+k}^{AB}}. \tag{3.19}$$

The solution obtained in this case is equivalent to the aggregation function in Stevens and Urquhart (2000, Section 4.2), to the attribute density in Gregoire and Valentine (2007, Section 10.2) or to the local density in Mandallaz (2007, equation 4.5).

For illustration, we present a toy example in Figure 3.1. We are interested in a rectangular territory $\mathcal{U}^A$ with a global measure $M^A = 7 \times 8 = 56\, m^2$. Inside this area, we have a population U^B of $N^B = 5$ trees. A sample of $n^A = 2$ points is selected by independent drawings with the marginal PDF given in (3.4), which leads to the observation of, say, $s^A = \{s_1^A, s_2^A\}$.

Figure 3.1 A simple example of links between a continuous population $\mathcal{U}^A$ and a discrete population U^B.



For each point $s \in s^A$, the circle $C_r(s)$ centered on s with radius $r = 1$ is drawn, and all the trees k such that their center x_k is inside these circles are surveyed. In the example presented in Figure 3.1, we obtain $s^B = \{1, 4, 5\}$. The values of the variable of interest for the units in s^B are, say:

$$y_1^B = 1,\ y_4^B = 4,\ y_5^B = 3.$$

We compute the quantities M_{+k}^{AB} for the trees $k \in s^B$, applying equation (3.18). For the trees $k = 4$ and 5 the circle $C_r(x_k)$ is included in $\mathcal{U}^A$, resulting in $M_{+k}^{AB} \simeq 3.14$. For the tree $k = 1$, we have $M_{+1}^{AB} \simeq 3.00$. We compute the synthetic variable $y^A(x)$ for the points in s^A, making use of equation (3.19). We obtain

$$y^A(s_1) = \frac{y_1^B}{3.00} \simeq 0.33,$$

$$y^A(s_2) = \frac{y_4^B}{3.14} + \frac{y_5^B}{3.14} \simeq 2.23.$$

Finally, we compute the weights w_k^B for the trees $k \in s^B$ by means of the weight share method, making use of equation (3.16). We obtain

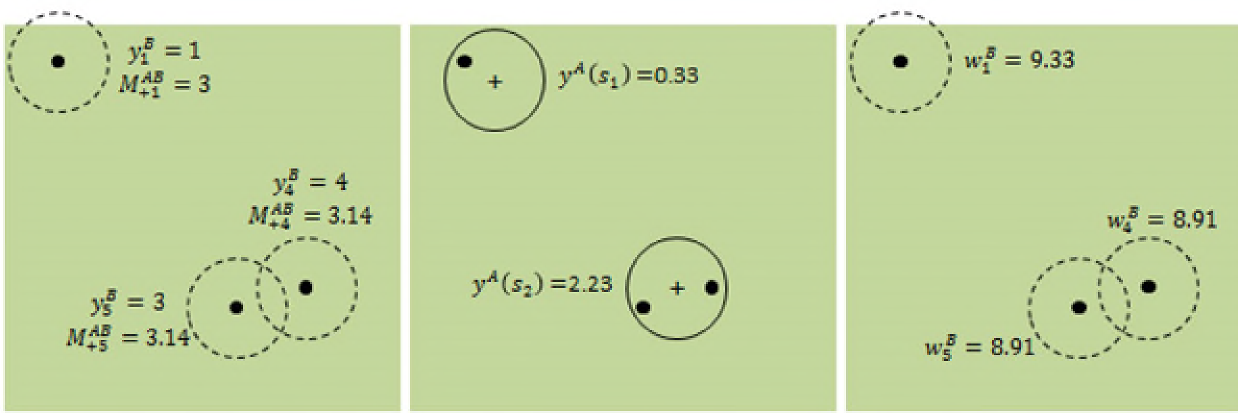
$$\begin{aligned} w_1^B &= \frac{d^A(s_1)}{M_{+1}^{AB}} = \frac{28}{3.00} \simeq 9.33, \\ w_4^B &= \frac{d^A(s_2)}{M_{+4}^{AB}} = \frac{28}{3.14} \simeq 8.91, \\ w_5^B &= \frac{d^A(s_2)}{M_{+5}^{AB}} = \frac{28}{3.14} \simeq 8.91. \end{aligned}$$

The estimate for the total τ_y^B is therefore

$$\hat{\tau}_y^B(s^B) = \sum_{k \in s^B} w_k^B y_k^B \simeq 71.72.$$

The results are summarized in Figure 3.2.

Figure 3.2 Computation of the synthetic variable $y^A(x)$ and of the weights w_k^B obtained by the weight share method on a simple example.

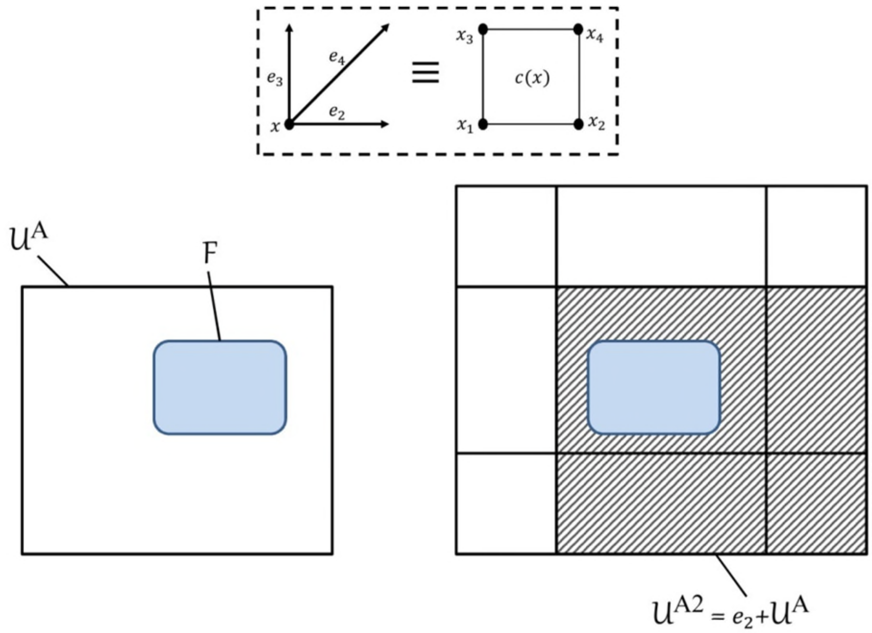


3.4 Application to spatial cluster sampling for forest inventories

When the accessibility to the field is difficult, it is common practice in forest inventories to use clusters of plots (Köhl and Magnussen, 2015). These clusters have a fixed geometric form determined before the survey. For instance, plots may be positioned at the corners of a square of 50m size (see Mandallaz, 2007, Section 4.3).

Suppose that we are again interested in a population U^B of trees located on a forest $\mathcal{F}$. Let $\mathcal{U}^A$ denote a set such that $\mathcal{F} \subset \mathcal{U}^A$, and let $e_1, \ldots, e_L$ denote a set of L vectors in $\mathbf{R}^2$. For any point $x \in \mathcal{U}^A$, the cluster $c(x)$ is defined as the set of points $\{x_l \equiv x + e_l;\ l = 1, \ldots, L\}$. Following the notation by Mandallaz (2007), we take e_1 as the null vector, and $x_1 = x$ is seen as the origin of the cluster. Let us denote by $\mathcal{U}^{Al} = e_l + \mathcal{U}^A$ for $l = 1, \ldots, L$, and $\mathcal{V}^C = \bigcup_{l=1}^L \mathcal{U}^{Al}$ their union. It is also supposed that the set $\mathcal{U}^A$ is large enough to ensure that $\mathcal{F} \subset \mathcal{U}^{Al}$ for any $l = 1, \ldots, L$ (see Mandallaz, 2007, Section 4.3). For illustration, an example in the case $L = 4$ is presented in Figure 3.3.

Figure 3.3 An example of cluster sampling in continuous populations.



Lecture note. In the upper panel, an example of cluster of size $L = 4$ originated in the point x is presented. In the left panel, two continuous populations such that $\mathcal{F} \subset \mathcal{U}^A$ are given. In the right panel, the sets $\mathcal{U}_{Al}, l = 1, \ldots, 4$ associated to $\mathcal{U}^A$ are presented, where the hatched set stands for $\mathcal{U}_{A2}$.

Cluster sampling is performed by first selecting a sample S^A of n^A points in $\mathcal{U}^A$, according to a continuous sampling design. For each point $x \in S^A$, we obtain the associated cluster $c(x) = \{x_1, \ldots, x_L\}$ originated in $x_1 = x$, and the associated circles $C_r(x_l)$ are drawn with some predetermined radius r. All the trees $k \in U^B$ such that their center x_k is inside one of the circles $C_r(x_l), l = 1, \ldots, L$ for some $x \in S^A$ are surveyed. An example is presented in Figure 3.4.

For any point $x \in \mathcal{U}^A$, the synthetic variable $y^A(x)$ is obtained in two steps, using $\mathcal{V}^C$ as a pivotal population. We first define a link function between $\mathcal{V}^C$ and U^B as

$$L^{CB}(z, k) = \begin{cases} 1 & \text{if } x_k \in C_r(z), \\ 0 & \text{otherwise,} \end{cases} \tag{3.20}$$

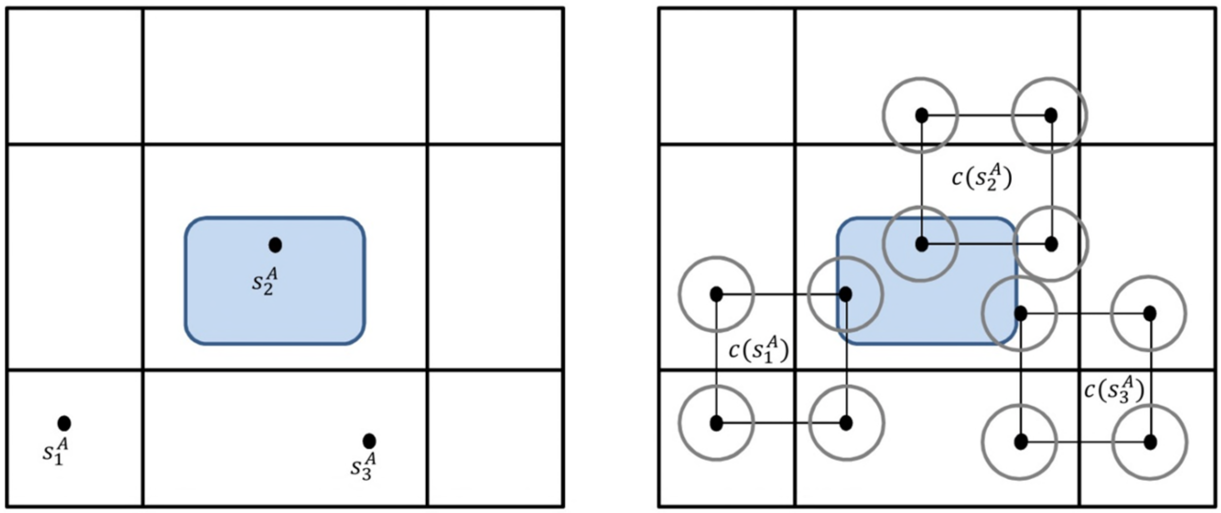
with x_k the center of the tree k and z a point in the union set $\mathcal{V}^C$. This is similar to the link function defined in (3.17), and following the same lines of reasoning we obtain the intermediary synthetic variable

$$y^C(z) = \sum_{x_k \in C_r(z)} \frac{y_k^B}{M_{+k}^{CB}} \quad \text{for any} \quad z \in \mathcal{V}^C, \tag{3.21}$$

where $M_{+k}^{CB} = \int_{z \in \mathcal{V}^C} 1\{z \in C_r(x_k)\}\, dz$ is the area of the intersection between $\mathcal{V}^C$ and the circle centered on x_k. Then, we define a link function between $\mathcal{U}^A$ and $\mathcal{V}^C$ as

$$L^{AC}(x, z) = \begin{cases} 1 & \text{if } z \in c(x) = \{x_1, \ldots, x_L\}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.22}$$

Figure 3.4 An example of cluster sampling where $n^A = 3$ points are initially selected from $\mathcal{U}^A$.



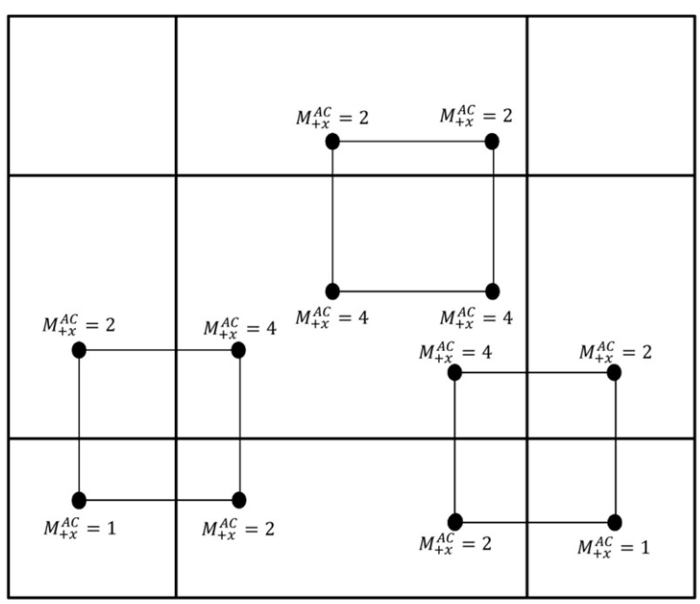
In other words, x and z are linked if z is one of the vertices of the cluster originated in x. The quantity

$$M_{+z}^{AC} = \int_{x \in \mathcal{U}^A} L^{AC}(x, z)\, dx \tag{3.23}$$

is the number of clusters having z as a vertex. We give, in Figure 3.5, the values obtained for the vertices of the clusters in the example initiated in Figure 3.4. We finally obtain the synthetic variable

$$y^A(x) = \sum_{z \in c(x)} \frac{y^C(z)}{M_{+z}^{AC}} = \sum_{z \in c(x)} \frac{1}{M_{+z}^{AC}} \sum_{x_k \in C_r(z)} \frac{y_k^B}{M_{+k}^{CB}} \quad \text{for any} \quad x \in \mathcal{U}^A. \tag{3.24}$$

Figure 3.5 Values of the quantities M_{+x}^{AC} for the vertices in a cluster sample.



As proved in Proposition 2, τ_y^B may be unbiasedly estimated by $\hat{\tau}_y^B$, and an unbiased variance estimator is directly obtained by applying equation (3.10) or equation (3.11). Therefore, the weight share method provides a general solution for both estimation and variance estimation for cluster sampling, under an arbitrary sampling design performed in $\mathcal{U}^A$. On the other hand, the solution described in Mandallaz (2007, equation 4.17) is suitable only when S^A is obtained by independent uniform selections from $\mathcal{U}^A$.

4. Simulation study

In this simulation study, we consider estimation and variance estimation for spatial cluster sampling. We wish to compare the weight share method with the solution described in Mandallaz (2007, equation 4.17), in the situation when the sample S^A is obtained by independent uniform selections from $\mathcal{U}^A$. This is the purpose of the first simulation study described in Section 4.1. We also wish to evaluate the weight share method in the situation when the sample S^A of cluster origins is not selected by independent uniform sampling, and when Mandallaz's technique may therefore not be applied. This is the purpose of the second simulation study described in 4.2.

The continuous population $\mathcal{U}^A$ that we consider is a square of length 1,000 meters. Inside $\mathcal{U}^A$, the forest is located on a square territory $\mathcal{F}$ of length 600 meters, but the location of the forest and its area are seen as unknown prior to the survey. A population of $N^B = 30{,}942$ black pines is generated in the forest $\mathcal{F}$. The main characteristics of the population of pines in terms of volume (y_{1k}^B, cube meters) and breast-height diameter (y_{2k}^B, centimeters) are summarized in Table 4.1.

Table 4.1
Mean, standard deviation, minimum and maximum for the breast-height diameter and volume for the population of black pines

	Mean	Standard deviation	Minimum	Maximum
Breast-height diameter (centimeters)	17.28	6.38	3.57	41.78
Volume (cube meters)	0.17	0.16	0.00	0.99

We are interested in estimating the following parameters: the total number of trees $\tau_{y0}^B = \sum_{k \in U^B} y_{0k}^B$ with $y_{0k}^B = 1$, the total volume of wood $\tau_{y1}^B = \sum_{k \in U^B} y_{1k}^B$, the mean volume of trees $\mu_{y1}^B = \tau_{y1}^B / N^B$, the mean breast-height diameter $\mu_{y2}^B = \tau_{y2}^B / N^B$, the average volume of wood per square meter of forest $\bar{Y}_1^B = \tau_{y1}^B / M^{\mathcal{F}}$, where $M^{\mathcal{F}}$ stands for the area of the forest $\mathcal{F}$.

4.1 Comparison between the weight share method and Mandallaz's estimator for cluster sampling

We first compare the performance of the proposed estimators to those proposed by Mandallaz (2007, Section 4.3), in case of cluster sampling (see Section 3.4) where the sample S^A of cluster origins is selected by independent uniform sampling from the territory $\mathcal{U}^A$. We select a sample S^A of size $n^A = 100, 200$ or 400. A sample of trees is then selected and surveyed by using the cluster sampling technique described in Section 3.4, where we use square clusters of size $L = 4$ and length 60 meters, and plots with a radius r of 25 meters.

For a given sample, the estimators under the weight share method are obtained as follows. The estimators of the totals τ_{y0}^{B} and τ_{y1}^{B} are obtained from equation (3.16) as

$$\hat{\tau}_{y0}^{B} = \frac{M^{A}}{n^{A}} \sum_{x \in S^{A}} y_{0}^{A}(x), \tag{4.1}$$

$$\hat{\tau}_{y1}^{B} = \frac{M^{A}}{n^{A}} \sum_{x \in S^{A}} y_{1}^{A}(x),$$

where y_{0}^{A} and y_{1}^{A} are obtained by plugging into equation (3.24) the variables y_{0k}^{B} and y_{1k}^{B}, respectively, and where M^{A} is the area of $\mathcal{U}^{A}$. The estimators of the population means μ_{y1}^{B} and μ_{y2}^{B} are

$$\hat{\mu}_{y1}^{B} = \frac{\hat{\tau}_{y1}^{B}}{\hat{\tau}_{y0}^{B}} \quad \text{and} \quad \hat{\mu}_{y2}^{B} = \frac{\hat{\tau}_{y2}^{B}}{\hat{\tau}_{y0}^{B}},$$

respectively, obtained by using the plug-in principle, and where $\hat{\tau}_{y2}^{B}$ is obtained as described in equation (4.1). The estimator of $\bar{Y}_{1}^{B}$ is

$$\hat{\bar{Y}}_{1}^{B} = \frac{\hat{\tau}_{y1}^{B}}{\hat{M}^{\mathcal{F}}},$$

where

$$\hat{M}^{\mathcal{F}} = \frac{M^{A}}{n^{A}} \sum_{x \in S^{A}} y_{3}^{A}(x) \quad \text{with} \quad y_{3}^{A}(x) = \sum_{z \in c(x)} \frac{1\{z \in \mathcal{F}\}}{M_{+z}^{AC}} \tag{4.2}$$

is an unbiased estimator of $M^{\mathcal{F}}$. The estimators proposed by Mandallaz are obtained as follows. The estimator of $\bar{Y}_{1}^{B}$ is

$$\hat{\bar{Y}}_{1,\text{mand}}^{B} = \frac{\sum_{x \in S^{A}} \sum_{z \in c(x)} 1\{z \in \mathcal{F}\} \sum_{k \in C_{r}(z)} \frac{y_{1k}^{B}}{M_{+k}^{\mathcal{F}}}}{\sum_{x \in S^{A}} \sum_{z \in c(x)} 1\{z \in \mathcal{F}\}}$$

with $M_{+k}^{\mathcal{F}}$ the area of the intersection between $\mathcal{F}$ and the circle centered on x_{k}, see equation (4.17) in Mandallaz (2007). The estimators of the totals τ_{y0}^{B} and τ_{y1}^{B} are

$$\hat{\tau}_{y0,\text{mand}}^{B} = \frac{1}{L} \frac{M^{A}}{n^{A}} \sum_{x \in S^{A}} \sum_{z \in c(x)} 1\{z \in \mathcal{F}\} \sum_{k \in C_{r}(z)} \frac{y_{0k}^{B}}{M_{+k}^{\mathcal{F}}},$$

$$\hat{\tau}_{y1,\text{mand}}^{B} = \frac{1}{L} \frac{M^{A}}{n^{A}} \sum_{x \in S^{A}} \sum_{z \in c(x)} 1\{z \in \mathcal{F}\} \sum_{k \in C_{r}(z)} \frac{y_{1k}^{B}}{M_{+k}^{\mathcal{F}}},$$

see equation (4.16) in Mandallaz (2007). The estimators of the population means μ_{y1}^{B} and μ_{y2}^{B} are

$$\hat{\mu}_{y1,\text{mand}}^{B} = \frac{\hat{\tau}_{y1,\text{mand}}^{B}}{\hat{\tau}_{y0,\text{mand}}^{B}} \quad \text{and} \quad \hat{\mu}_{y2,\text{mand}}^{B} = \frac{\hat{\tau}_{y2,\text{mand}}^{B}}{\hat{\tau}_{y0,\text{mand}}^{B}},$$

respectively, obtained by using the plug-in principle.

The sampling and estimation steps are repeated $D = 10{,}000$ times. For an estimator $\hat{\theta}$ of a parameter θ, we compute the Monte Carlo Percent Relative Bias

$$\text{RB}\{\hat{\theta}\} = 100 \times \frac{D^{-1}\sum_{d=1}^{D}\hat{\theta}_d - \theta}{\theta}, \tag{4.3}$$

and the Monte Carlo Mean Square Error

$$\text{MSE}\{\hat{\theta}\} = \frac{1}{D}\sum_{d=1}^{D}\left(\hat{\theta}_d - \theta\right)^2. \tag{4.4}$$

The results are given in Table 4.2. For any of the five parameters, both estimation methods lead to virtually unbiased estimators, and the mean square errors are very close. The estimator of Mandallaz performs slightly better for $\bar{Y}_1^B$, while the weight share method performs slightly better for the other parameters.

Table 4.2
Percent relative bias and mean square error for five parameters estimated by means of the weight share method or by estimators proposed by Mandallaz

			τ_{y0}^{B}	τ_{y1}^{B}	μ_{y1}^{B}	μ_{y2}^{B}	$\bar{Y}_1^{B}$
$n^A = 100$	Weight share	RB (%)	-0.01	-0.01	0.00	0.00	-0.15
		MSE	$2.04\ 10^{7}$	$6.24\ 10^{5}$	$1.22\ 10^{-6}$	$1.81\ 10^{-3}$	$8.76\ 10^{-7}$
	Mandallaz	RB (%)	-0.02	-0.02	0.00	0.00	-0.17
		MSE	$2.06\ 10^{7}$	$6.29\ 10^{5}$	$1.25\ 10^{-6}$	$1.85\ 10^{-3}$	$8.65\ 10^{-7}$
$n^A = 200$	Weight share	RB (%)	-0.01	-0.01	0.00	0.00	-0.15
		MSE	$9.91\ 10^{6}$	$3.03\ 10^{5}$	$6.07\ 10^{-7}$	$9.09\ 10^{-4}$	$4.15\ 10^{-7}$
	Mandallaz	RB (%)	-0.01	-0.01	0.00	0.00	-0.15
		MSE	$9.99\ 10^{6}$	$3.05\ 10^{5}$	$6.20\ 10^{-7}$	$9.27\ 10^{-4}$	$4.08\ 10^{-7}$
$n^A = 400$	Weight share	RB (%)	-0.02	-0.02	0.00	0.00	-0.03
		MSE	$5.03\ 10^{6}$	$1.54\ 10^{5}$	$2.99\ 10^{-7}$	$4.44\ 10^{-4}$	$2.15\ 10^{-7}$
	Mandallaz	RB (%)	-0.02	-0.03	0.00	0.00	-0.04
		MSE	$5.07\ 10^{6}$	$1.55\ 10^{5}$	$3.05\ 10^{-7}$	$4.52\ 10^{-4}$	$2.12\ 10^{-7}$

Note: Relative bias (RB); Mean square error (MSE).

For the estimators obtained under the weight share method, we also considered variance estimation. We did not perform variance estimation for Mandallaz's estimators, since in Mandallaz (2007, Section 4.3), variance estimators are only proposed for spatial means like $\bar{Y}_1^B$, and not for population totals or population means. Under independent uniform sampling from $\mathcal{U}^A$, applying equation (3.10) leads to the unbiased variance estimator

$$\hat{V}_{\mathrm{HT}}\left(\hat{\tau}_{y0}^{B}\right) = \frac{\left(M^{A}\right)^{2}}{n^{A}} \times \frac{1}{n^{A}-1} \sum_{x \in S^{A}} \left\{ y_{0}^{A}(x) - \frac{1}{n^{A}} \sum_{s \in S^{A}} y_{0}^{A}(s) \right\}^{2}, \tag{4.5}$$

for $\hat{\tau}_{y0}^{B}$, and a similar expression for $\hat{\tau}_{y1}^{B}$. An approximately unbiased variance estimator for $\hat{\mu}_{y1}^{B}$ is obtained by replacing in (4.5) the variable $y_0^A(x)$ by the linearized variable

$$u_{1}^{A}(x) = \frac{1}{\hat{\tau}_{y0}^{B}} \left\{ y_{1}^{A}(x) - \hat{\mu}_{y1}^{B}\, y_{0}^{A}(x) \right\}, \tag{4.6}$$

and can similarly be obtained for $\hat{\mu}_{y2}^{B}$. An approximately unbiased variance estimator for $\hat{\bar{Y}}_{1}^{B}$ is obtained by replacing in (4.5) the variable $y_0^A(x)$ by the linearized variable

$$u_{2}^{A}(x) = \frac{1}{\hat{\tau}_{y3}^{B}} \left\{ y_{1}^{A}(x) - \hat{\bar{Y}}_{1}^{B} y_{3}^{A}(x) \right\}. \tag{4.7}$$

The sampling and estimation steps are repeated $E = 1{,}000$ times. To measure the bias of a variance estimator $\hat{V}_{\mathrm{HT}}(\hat{\theta})$, we use the Monte Carlo Percent Relative Bias

$$\mathrm{RB}\left\{\hat{V}_{\mathrm{HT}}(\hat{\theta})\right\} = 100 \times \frac{E^{-1}\sum_{e=1}^{E} \hat{V}_{\mathrm{HT}}\left(\hat{\theta}_{e}\right) - \mathrm{MSE}(\hat{\theta})}{\mathrm{MSE}(\hat{\theta})}, \tag{4.8}$$

where $\mathrm{MSE}(\hat{\theta})$ is obtained independently from the first run of $D = 10{,}000$ simulations, see equation (4.4). The results are presented in Table 4.3. All the variance estimators are approximately unbiased.

Table 4.3
Percent relative bias of a variance estimator for five parameters estimated by means of the weight share method

n^A	$\hat{V}\left(\hat{\tau}_{y0}^{B}\right)$	$\hat{V}\left(\hat{\tau}_{y1}^{B}\right)$	$\hat{V}\left(\hat{\mu}_{y1}^{B}\right)$	$\hat{V}\left(\hat{\mu}_{y2}^{B}\right)$	$\hat{V}\left(\hat{\bar{Y}}_{1}^{B}\right)$
100	-1.21	-1.17	-0.48	-0.04	-0.17
200	1.75	1.83	-0.59	-0.76	3.35
400	0.49	0.62	0.40	1.09	-0.29

4.2 Evaluation of the weight share method for cluster sampling with non-uniform sampling of the clusters

We consider a second case of cluster sampling when the sample S^A is not selected by independent uniform sampling from $\mathcal{U}^A$, so that the estimators proposed by Mandallaz (2007) can not be used. The population $\mathcal{U}^A$ is first partitioned into a sub-population $\mathcal{U}_1^A$ of length 300 meters and height 1,000 meters (west part), and a sub-population $\mathcal{U}_2^A$ of length 700 meters and height 1,000 meters (east part). The area of $\mathcal{U}_1^A$ and $\mathcal{U}_2^A$ are denoted as M_1^A and M_2^A, respectively. A sample of size $n_1 = 75, 150$ or 300 points is first selected globally from $\mathcal{U}^A$, and a second sample of size $n_2 = 25, 50$ or 100 is then selected from $\mathcal{U}_2^A$. For example, this case may arise if it is of specific interest to perform estimation on the sub-population $\mathcal{U}_2^A$, and if an extension sample may be funded to ensure that the global sample selected from

$\mathcal{U}_2^A$ is sufficiently large. The union of these two samples is denoted as S^A. We also let n_1^A denote the (random) size of $S_1^A = S^A \cap \mathcal{U}_1^A$, and n_2^A denote the (random) size of $S_2^A = S^A \cap \mathcal{U}_2^A$.

Conditionally on n_1^A and n_2^A, unbiased estimators for the totals τ_{y0}^B and τ_{y1}^B are

$$\hat{\tau}_{y0,\text{cond}}^B = \frac{M_1^A}{n_1^A} \sum_{x \in S_1^A} y_0^A(x) + \frac{M_2^A}{n_2^A} \sum_{x \in S_2^A} y_0^A(x),$$

$$\hat{\tau}_{y1,\text{cond}}^B = \frac{M_1^A}{n_1^A} \sum_{x \in S_1^A} y_1^A(x) + \frac{M_2^A}{n_2^A} \sum_{x \in S_2^A} y_1^A(x),$$

where y_0^A and y_1^A are obtained by plugging into equation (3.24) the variables y_{0k}^B and y_{1k}^B, respectively. The estimators of the population means μ_{y1}^B and μ_{y2}^B are

$$\hat{\mu}_{y1,\text{cond}}^B = \frac{\hat{\tau}_{y1,\text{cond}}^B}{\hat{\tau}_{y0,\text{cond}}^B} \quad \text{and} \quad \hat{\mu}_{y2,\text{cond}}^B = \frac{\hat{\tau}_{y2,\text{cond}}^B}{\hat{\tau}_{y0,\text{cond}}^B},$$

respectively, obtained by using the plug-in principle. The estimator of $\bar{Y}_1^B$ is

$$\hat{\bar{Y}}_{1,\text{cond}}^B = \frac{\hat{\tau}_{y1,\text{cond}}^B}{\hat{M}_{\text{cond}}^{\mathcal{F}}},$$

where

$$\hat{M}_{\text{cond}}^{\mathcal{F}} = \frac{M_1^A}{n_1^A} \sum_{x \in S_1^A} y_3^A(x) + \frac{M_2^A}{n_2^A} \sum_{x \in S_2^A} y_3^A(x)$$

is an unbiased estimator of $M^{\mathcal{F}}$, where $y_3^A(x)$ is defined in equation (4.2).

The sampling and estimation steps are repeated $D = 10{,}000$ times. For an estimator $\hat{\theta}$ of a parameter θ, we compute the Monte Carlo Percent Relative Bias given in (4.3) and the Monte Carlo Mean Square Error given in (4.4). The results are given in Table 4.4. For any of the five parameters, the estimators are virtually unbiased. The mean square error decreases as the sample size increases, as could be expected.

Table 4.4
Percent relative bias and Mean square error for five parameters estimated by means of the weight share method

		τ_{y0}^B	τ_{y1}^B	μ_{y1}^B	μ_{y2}^B	$\bar{Y}_1^B$
$n^A = 100$	RB (%)	0.03	0.03	0.00	0.00	-0.02
	MSE	$1.85\ 10^7$	$5.67\ 10^5$	$1.19\ 10^{-6}$	$1.78\ 10^{-3}$	$8.5\ 10^{-7}$
$n^A = 200$	RB (%)	-0.02	-0.02	0.01	0.00	-0.06
	MSE	$9.40\ 10^6$	$2.87\ 10^5$	$5.83\ 10^{-7}$	$8.70\ 10^{-4}$	$4.34\ 10^{-7}$
$n^A = 400$	RB (%)	0.09	0.09	0.00	0.00	0.05
	MSE	$4.67\ 10^6$	$1.43\ 10^5$	$2.90\ 10^{-7}$	$4.35\ 10^{-4}$	$2.18\ 10^{-7}$

Note: Relative bias (RB); Mean square error (MSE).

We now consider variance estimation. Conditionally on n_1^A and n_2^A, an unbiased variance estimator for $\hat{\tau}_{y0,\text{cond}}^B$ is

$$\hat{V}_2\left(\hat{\tau}_{y0}^B\right) = \frac{\left(M_1^A\right)^2}{n_1^A} \times \frac{1}{n_1^A - 1} \sum_{x \in S_1^A} \left\{ y_0^A(x) - \frac{1}{n_1^A} \sum_{s \in S_1^A} y_0^A(s) \right\}^2 + \frac{\left(M_2^A\right)^2}{n_2^A} \times \frac{1}{n_2^A - 1} \sum_{x \in S_2^A} \left\{ y_0^A(x) - \frac{1}{n_2^A} \sum_{s \in S_2^A} y_0^A(s) \right\}^2, \tag{4.9}$$

and can be expressed similarly for $\hat{\tau}_{y1,\text{cond}}^B$. An approximately unbiased variance estimator for $\hat{\mu}_{y1,\text{cond}}^B$ is obtained by replacing in (4.9) the variable $y_0^A(x)$ by the linearized variable given in equation (4.6), and can be obtained similarly for $\hat{\mu}_{y2}^B$. An approximately unbiased variance estimator for $\hat{\bar{Y}}_{1,\text{cond}}^B$ is obtained by replacing in (4.9) the variable $y_0^A(x)$ by the linearized variable given in equation (4.7).

The sampling and estimation steps are repeated $E = 1{,}000$ times. To measure the bias of a variance estimator $\hat{V}_2(\hat{\theta})$, we use the Monte Carlo Percent Relative Bias defined in (4.3). The results are presented in Table 4.5. All the variance estimators are approximately unbiased.

Table 4.5
Percent relative bias of a variance estimator for five parameters estimated by means of the weight share method

n^A	$\hat{V}_2\left(\hat{\tau}_{y0}^B\right)$	$\hat{V}_2\left(\hat{\tau}_{y1}^B\right)$	$\hat{V}_2\left(\hat{\mu}_{y1}^B\right)$	$\hat{V}_2\left(\hat{\mu}_{y2}^B\right)$	$\hat{V}_2\left(\hat{\bar{Y}}_1^B\right)$
100	3.18	3.16	-3.02	-2.96	-0.28
200	0.97	0.92	-1.59	-1.01	-1.88
400	1.12	1.03	0.05	-0.36	-1.46

5. Discussion

There are several reasons in practice why a population has no tractable sampling frame. When the population of interest may be linked to a discrete population for which a sampling frame is available, the usual weight share method (Deville and Lavallée, 2006) makes it possible to obtain probability samples from the population of interest, as well as unbiased estimators and variance estimators for this population. We showed that this approach may be generalized in a natural way when the population of interest is linked to a continuous population, by using a synthetic function on this continuous population which may be interpreted as a local measure of density. In case of spatial cluster sampling with independent uniform selection of the cluster origins, our simulation results show that the weight share method and Mandallaz's estimator perform similarly. Mandallaz's estimator can not be applied when the cluster origins are not selected by independent uniform sampling. In this case, our simulation results confirm that the weight share method leads to unbiased estimators.

In our view, making use of the weight share method has several advantages. Firstly, it enables to easily handle the so-called edge corrections, i.e., the fact that some units have an ancestor subset which intersects with the area of interest. Using the area of the ancestor subset M_{+k}^{AB} in the estimation weights (see equation (3.16)) leads to exactly unbiased estimators, while alternative edge corrections may be somewhat cumbersome, see Gregoire and Valentine (2007, Section 10.7) or Roesch, Green and Scott (1993), for example. Also, our approach enables us to go back to the population which was indeed sampled. This is necessary to make use of the Horvitz-Thompson estimator, and to compute an unbiased variance estimator. This is possible in full generality, i.e., under an arbitrary continuous sampling design, by using the theory developed by Cordy (1993).

This method is obviously not limited to forest inventory. One example is the survey on crop practices conducted by the French statistical office of the Ministry of Agriculture, until 2006. The sample for this survey consisted of parcels, selected from the points of the survey Ter-Uti (Chapelle-Barry, 2008). The Ter-Uti survey is dedicated to the production of statistics on the land use, and the points where the land use was associated to crop practices are the basis from which to select the parcels. Calculating weights for the parcels was done by considering that each parcel had a probability of being drawn proportional to its surface. In this way, the method produced the same weights than the weight share method, except for edge effects. However, using the weight share method would lead to have a general methodological framework enabling the derivation of a variance estimator. Other topics, linked to environmental issues, could also take advantage from the application of this method. This extension of the weight share method should also be considered for estimators that are not directly Horvitz-Thompson estimators, such as those relative to two-phase sampling, which are often used for environmental issues.

Acknowledgements

The authors would like to thank Pierre Lavallée and Minna Pulkkinen for useful discussions. The authors would also like to thank the referee, the associate editor and the assistant editor for helpful comments on an earlier version of the manuscript.

References

Ardilly, P., and Le Blanc, D. (2001). Sampling and weighting a survey of homeless persons: A French example. *Survey Methodology*, 27, 1, 109-118. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2001001/article/5859-eng.pdf.

Chapelle-Barry, C. (2008). Enquête sur les pratiques culturales en 2006. *Agreste Chiffres et Données*.

Clairin, R., and Brion, P. (1996). *Manuel de Sondages : Applications aux Pays en Développement*. Documents et manuels du CEPED, 3.

Cordy, C.B. (1993). An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe. *Statistics & Probability Letters*, 18(5), 353-362.

De Vitiis, C., Falorsi, S. and Inglese, F. (2014). Implementing the first ISTAT survey of homeless population by indirect sampling and weight sharing method, Springer. *Contributions to Sampling Statistics*, 119-138.

Deville, J.-C., and Lavallée, P. (2006). Indirect sampling: The foundations of the generalized weight share method. *Survey Methodology*, 32, 2, 165-176. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006002/article/9551-eng.pdf.

Gregoire, T.G., and Valentine, H.T. (2007). Sampling Strategies for Natural Resources and the Environment. CRC Press.

Himelein, K., Eckman, S. and Murray, S. (2014). Sampling nomads: A new technique for remote, hard-to-reach, and mobile populations. *Journal of Official Statistics*, 30(2), 191-213.

Köhl, M., and Magnussen, S. (2015). Sampling in forest inventories, Springer. *Tropical Forestry Handbook*, 1-50.

Laporte, A., Vandentorren, S., Détrez, M.-A., Douay, C., Le Strat, Y., Le Méner, E., Chauvin, P. and The Samenta Research Group (2018). Prevalence of mental disorders and addictions among homeless people in the greater Paris area, France. *International Journal of Environmental Research and Public Health*, 15(2), 241.

Lavallée, P. (2009). *Indirect Sampling*. Springer Science & Business Media.

Lawrence, M., McRoberts, R.E., Tomppo, E., Gschwantner, T. and Gabler, K. (2010). Comparisons of national forest inventories, Springer. *National Forest Inventories*, 19-32.

Lohlé-Tart, L., Clairin, R., François, M. and Gendreau, F. (1988). De l'Homme au Chiffre. Réflexions sur l'Observation Démographique en Afrique.

Lorenc, B., Smith, P.A. and Bavdaž, M. (2018). *The Unit Problem and Other Current Topics in Business Survey Methodology*. Cambridge Scholars Publishing.

Mandallaz, D. (2007). Sampling Techniques for Forest Inventories. CRC Press.

Roesch, F.A., Green, E.J. and Scott, C.T. (1993). An alternative view of forest sampling. *Survey Methodology*, 19, 2, 199-204. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1993002/article/14453-eng.pdf.

Stevens, D. (1997). Variable density grid-based sampling designs for continuous spatial populations. *Environmetrics: The Official Journal of the International Environmetrics Society*, 8(3), 167-195.

Stevens, D.L., and Urquhart, N.S. (2000). Response designs and support regions in sampling continuous domains. *Environmetrics: The Official Journal of the International Environmetrics Society*, 11(1), 13-41.

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.