



HAL
open science

Asymptotic properties of pivotal sampling with application to spatial sampling

Guillaume Chauvet, Ronan Le Gleut

► **To cite this version:**

Guillaume Chauvet, Ronan Le Gleut. Asymptotic properties of pivotal sampling with application to spatial sampling. 2018. hal-01853832v1

HAL Id: hal-01853832

<https://hal.science/hal-01853832v1>

Preprint submitted on 5 Aug 2018 (v1), last revised 23 Aug 2019 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Asymptotic properties of pivotal sampling with application to spatial sampling

August 3rd, 2018

Guillaume Chauvet* and Ronan Le Gleut†

Abstract

Unequal probability sampling without replacement is commonly used for sample selection. To produce estimators with associated confidence intervals, some basic statistical properties like consistency and asymptotic normality of the Horvitz-Thompson estimator are desirable. These properties have been mainly studied for large entropy sampling designs. On the other hand, spatial sampling designs rather make use of sampling algorithms which take into account the order of units in the population, like systematic sampling or pivotal sampling. So far, the statistical properties of such procedures have not been investigated. In this work, we study the asymptotic properties of the pivotal sampling design. Under mild assumptions, we prove that the Horvitz-Thompson estimator is asymptotically normally distributed and that a conservative variance estimator can always be computed. We also introduce a general spatial sampling design which is spatially balanced, which possesses good statistical properties and which is computationally very efficient, even for large databases.

Keywords: asymptotic normality, conservative variance estimator, martingale algorithm, spatial balance.

*ENSAI/IRMAR, Campus de Ker Lann, 35170 Bruz, France

†INSEE, 88 avenue Verdier, 92100 Montrouge, France

1 Introduction

Unequal probability sampling without replacement is commonly used for sample selection, since the accuracy of estimators may be significantly improved if the inclusion probabilities are chosen roughly proportional to the variable of interest. Numerous sampling algorithms have been proposed in the literature, see Tillé (2006) for a recent review. To produce consistent estimators with associated confidence intervals, some statistical properties are desirable for a sampling algorithm: namely, that (a) the Horvitz-Thompson (HT) estimator is weakly consistent for the true total; (b) the HT estimator satisfies a central-limit theorem; (c) an asymptotically valid confidence interval may be computed for the HT-estimator.

As underlined by Chauvet (2017), the property (a) of consistency for the HT estimator follows under a simple moment condition if the sampling algorithm under study is more efficient than with-replacement unequal probability sampling, a.k.a multinomial sampling (Tillé, 2006, Section 5.4). This property holds for the Sampford design and for the conditional Poisson sampling design (Gabler, 1981, 1984; Qualité, 2008). A general discussion about the consistency of the HT-estimator for large entropy sampling designs is given in Berger (2011). The property (b) has been proved by Hájek (1964) for conditional Poisson sampling, Rosén (1972) for successive sampling, and Ohlsson (1986) for the Rao-Hartley-Cochran (1962) procedure. Sufficient conditions for large entropy sampling designs are given in Berger (1998). Concerning property (c), the normality-based confidence interval obtained by replacing the unknown variance by an estimator is asymptotically valid if the variance estimator is weakly consistent. Such property holds for the so-called Horvitz-Thompson variance estimator under assumptions on first through fourth-order inclusion probabilities (see Breidt and Opsomer, 2000, assumptions H6 and H7). These assumptions may be shown to hold for conditional Poisson sampling (Boistard et al., 2012). In summary, the properties (a)-(c) have been mainly studied for large entropy sampling algorithms, where the units are closed to being independently selected conditionally on a fixed sample size.

In the context of spatial sampling, the variables of interest often present a positive spatial correlation, in the sense that neighbouring units tend to resemble each other. It is intuitively relevant in this case to select samples which are well spread over space, in order to optimize the information col-

lected. Such samples are called spatially balanced (e.g., Stevens and Olsen, 2004). Large entropy sampling designs are therefore not suitable, since they do not account for the distribution of units in space. On the other hand, sampling algorithms which take into account the order of units in the population have been extensively used in spatial sampling. Systematic sampling on a grid is commonly used, with the drawback that properties (a)-(c) usually do not hold unless we are willing to make some strong model assumptions. Systematic sampling is generalized to multiple dimensions by Stevens and Olsen (2004) who introduce the Generalized Random Tessellation Stratified (GRTS) sampling design. The method consists in defining an order on the spatial units in the population, and then applying systematic sampling after a partial randomization of the units. Though GRTS involves more randomization than direct systematic sampling, properties (a)-(c) are not guaranteed to hold either.

The pivotal method (Deville and Tillé, 2004; Tillé, 2006; Chauvet, 2012) is a very simple sequential sampling algorithm, which avoids selecting neighbouring units and therefore enables selecting spatially balanced samples. A vast literature has recently considered such applications for pivotal sampling, see for example Grafström et al. (2012); Grafström and Ringvall (2013); Grafström et al. (2014); Grafström and Tillé (2013); Dickson et al. (2014); Benedetti et al. (2015); Dickson and Tillé (2016); Fattorini et al. (2015); Vallée et al. (2015). However, these authors have mainly considered the properties of spatial sampling procedures in terms of spatial spread and in terms of accuracy of the HT-estimator. The statistical properties (a)-(c) have not been considered yet for spatial sampling designs, and this is the purpose of the present paper.

Pivotal sampling does not belong to the class of large entropy sampling designs. In applications for spatial sampling, some intricate dependencies are introduced to avoid the selection of neighbouring units. Therefore, the general tools for large entropy sampling designs considered by Berger (1998, 2011) are not applicable to prove the required statistical properties. Chauvet (2017) proved that pivotal sampling is more efficient than multinomial sampling, so that property (a) holds. We tackle properties (b)-(c) in this paper. We make use of a version of the martingale central-limit theorem to prove the asymptotic normality of the HT-estimator. It is not possible to produce a consistent variance estimator, since many second-order inclusion probabilities are equal to zero. However, we propose some very simple variance estimators which are proved to be conservative for the true variance.

This enables to produce very simple conservative confidence intervals. We also introduce a general spatial sampling design which is spatially balanced, which possesses good statistical properties and which is computationally very efficient, even for large databases. This is therefore a good alternative to the GRTS sampling design.

The paper is organized as follows. In Section 2, the notation is defined and our assumptions are introduced and discussed. In Section 3, a recursive algorithm for ordered pivotal sampling is first presented in Section 3.1, and a martingale characterization of the HT-estimator is given in Section 3.2, along with some useful properties. They are used in Section 3.3 to obtain a design-based martingale central-limit theorem for the HT-estimator. Variance estimation is then discussed in Section 3.4, and some very simple conservative variance estimators are proposed. In Section 4, we consider an application to spatial sampling. We first give some reminders on the GRTS sampling design in Section 4.1, and we propose in Section 4.2 a modification that we call the Pivotal Tessellation Method (PTM). By a comparison by simulations with alternative sampling designs in Section 5, we demonstrate that the proposed method is competitive in terms of spatial balance. We also study the properties of the proposed variance estimators. The proofs of Propositions and Theorems are given in Appendix. The proofs of additional Lemmas are given in the Supplementary Material.

2 Notation and assumptions

We consider a finite population U of size N . In order to study the asymptotic properties of the sampling designs and estimators that we treat below, we consider the asymptotic framework of Isaki and Fuller (1982). We assume that the population belongs to a nested sequence $\{U_\nu\}$ of finite populations with increasing sizes N_ν , and all limiting processes will be taken as $\nu \rightarrow \infty$. Though all quantities under consideration depend on ν , this subscript is omitted in what follows for simplicity of notation.

We assume that the units in U are ordered, prior to sampling, according to some permutation τ , random or not. We reason conditionally on τ , and therefore we do not need particular assumptions on this permutation. An interesting case is when the permutation is deterministic, obtained by ordering the units in U according to some auxiliary variable known for any unit in

the population. This leads to so-called ordered pivotal sampling (Chauvet, 2012). We consider in Section 4.2 an application to spatial sampling through a modification of the GRTS sampling design.

We note $U = \{1, \dots, N\}$ the units in the population after re-arrangement with respect to τ . We denote $\pi_U = (\pi_1, \dots, \pi_N)^\top$ a vector of probabilities, with $0 < \pi_k \leq 1$ for any unit k in U and $n = \sum_{k \in U} \pi_k$ the sample size. We note

$$\pi_M = \max_{k \in U} \pi_k. \quad (1)$$

We are interested in estimating the total $t_y = \sum_{k \in U} y_k$ for some variable of interest taking the value y_k for unit $k \in U$. We note $y_U = (y_1, \dots, y_N)^\top$ the vector of the population values. A random sample S is selected with inclusion probabilities π_U , and the total t_y is unbiasedly estimated by the Horvitz-Thompson (HT) estimator

$$\hat{t}_{y\pi} = \sum_{k \in S} \check{y}_k, \quad (2)$$

with $\check{y}_k = y_k/\pi_k$. We note $E(\cdot)$ and $V(\cdot)$ for the expectation and the variance of some estimator, and $E_{\{\mathcal{F}\}}(\cdot)$ and $V_{\{\mathcal{F}\}}(\cdot)$ for the expectation and the variance of some estimator conditionally on some σ -algebra \mathcal{F} .

We define the cumulative inclusion probabilities for unit k as $C_k = \sum_{l=1}^k \pi_l$, with $C_0 = 0$. The unit k is said to be cross-border if the cumulated inclusion probabilities exceed an integer for this specific unit. That is, the cross-border unit k_i is such that $C_{k_i-1} < i$ and $C_{k_i} \geq i$ for some positive integer $i = 1, \dots, n-1$. The inclusion probability for the cross-border unit k_i may be split as $\pi_{k_i} = a_i + b_i$ with

$$a_i = i - C_{k_i-1} \quad \text{and} \quad b_i = C_{k_i} - i. \quad (3)$$

We also note

$$c_i = \frac{a_i b_i}{(1 - a_i)(1 - b_i)}. \quad (4)$$

A microstratum U_i , $i = 1, \dots, n$, is a set of units that are between two cross-border units. We have $U_i = \{k \in U; k_{i-1} \leq k \leq k_i\}$, with $k_0 = 0$ and $k_n = N + 1$. We take by convention $a_0 = b_0 = 0$ and $a_n = b_n = 0$. For any

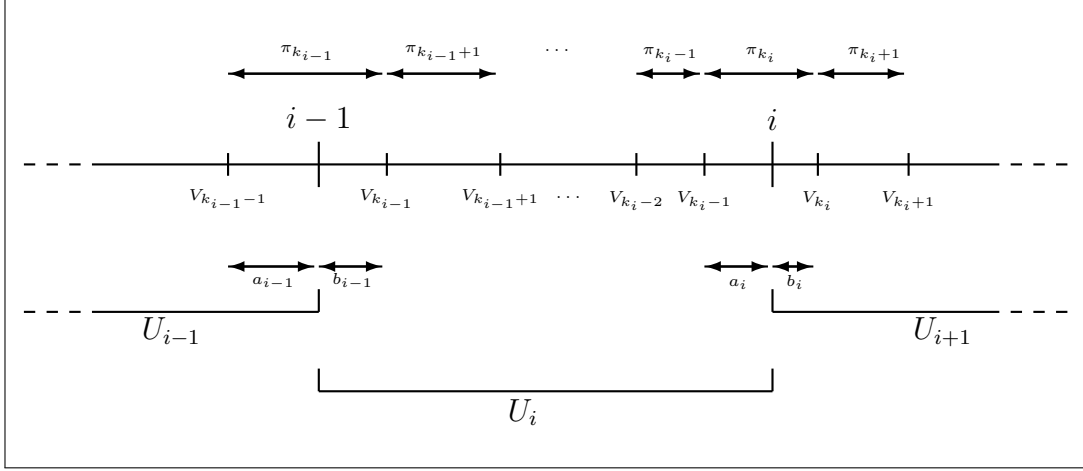


Figure 1: Probabilities and cross-border units in the microstratum U_i

unit $k \in U_i$, we note

$$\alpha_{ik} = \begin{cases} b_{i-1} & \text{if } k = k_{i-1}, \\ \pi_k & \text{if } k_{i-1} < k < k_i, \\ a_i & \text{if } k = k_i, \end{cases} \quad (5)$$

and $\alpha_i = (\alpha_{ik})_{k \in U_i}$. We have in particular $\sum_{k \in U_i} \alpha_{ik} = 1$.

The microstrata are overlapping, since a cross-border unit usually belongs to two adjacent microstrata: k_i belongs both to the microstratum U_i with an associated probability a_i , and to the microstratum U_{i+1} with an associated probability b_i . To fix ideas, useful quantities for population U are presented in Figure 1.

We make the following assumptions:

H1: There exists some constant $f_1 < 1$ such that for any $k \in U$:

$$\pi_k \leq f_1. \quad (6)$$

H2: There exists some constant C_1 such that:

$$\sum_{k \in U} \pi_k \left(\check{y}_k - \frac{t_y}{n} \right)^4 \leq C_1 \frac{N^4}{n^3}. \quad (7)$$

H3: There exists some constant $C_2 > 0$ such that:

$$\sum_{i=1}^n \sum_{k \in U_i} \alpha_{ik} \left(\check{y}_k - \sum_{l \in U_i} \alpha_{il} \check{y}_l \right)^2 \geq C_2 \frac{N^2}{n}. \quad (8)$$

It assumed in (H1) that the first-order inclusion probabilities are bounded away from 1. This is not a severe restriction in practice, since some unit with an inclusion probability equal to 1 is automatically surveyed, and is thus not involved in the selection process. The condition (H2) holds in particular if the variable y has a finite centered moment of order 4, and if there exists some constants $D_1, D_2 > 0$ such that for any $k \in U$, $D_1 \leq Nn^{-1}\pi_k \leq D_2$. Assumption (H3) requires that the dispersion within the micro-strata does not vanish.

3 Pivotal sampling and statistical properties

3.1 The pivotal method

We suppose that the random sample S is selected by means of pivotal sampling with inclusion probabilities π_U . Pivotal sampling (Deville and Tillé, 1998) is based on duels between units, and may be summarized as follows. At the first step, the two first units in the population fight, and the loser is definitely discarded from the sample while the winner gets their cumulated probabilities $\pi_1 + \pi_2$. The remaining unit then faces unit 3 in a similar principle. The fights go on until all non cross-border units in a micro-stratum have been considered. The successive duels result in discarding all the units inside the micro-stratum, except one which gets the cumulated inclusion probabilities.

When this surviving unit (denoted as S_1) faces the first cross-border unit k_1 , the cumulated inclusion probabilities exceed 1. In this case, one of the two units (denoted as F_1) is selected in the sample, while the other unit (denoted as L_1) goes on with the residual probability $C_{k_1} - 1 = b_1$. The unit L_1 then faces the next unit $k_1 + 1$, and the duels go on. The algorithm stops at step $N - 1$, when the two last units fight. A recursive description of pivotal sampling is presented in Algorithm 1.

Algorithm 1 Pivotal sampling in the population U

- We initialize with $L_0 = k_0$.
- At any step $i = 1, \dots, n$:
 - The unit L_{i-1} jumps to microstratum U_i with the residual probability b_{i-1} .
 - One unit, denoted as S_i , is selected among $\{L_{i-1}, k_{i-1} + 1, \dots, k_i - 1\}$ with probabilities proportional to $(b_{i-1}, \pi_{k_{i-1}+1}, \dots, \pi_{k_i-1})$.
 - The unit S_i faces k_i . One of these two units, denoted as F_i , is selected while the other one, denoted as L_i , jumps to microstratum U_{i+1} with the residual probability b_i . We have

$$(F_i, L_i) = \begin{cases} (S_i, k_i) & \text{with probability } \frac{1-a_i-b_i}{1-b_i}, \\ (k_i, S_i) & \text{with probability } \frac{a_i}{1-b_i}. \end{cases} \quad (9)$$

- The final sample is $\{F_1, \dots, F_n\}$.
-

Pivotal sampling is a fixed-size sampling design which matches exactly the set π_U of prescribed inclusion probabilities (Deville and Tillé, 1998). By construction, the selection of neighbouring units is avoided, since two non cross-border units inside a same microstratum U_i may not be selected together in the sample. This method is therefore of interest in situations where contiguous units are similar with respect to the variables of interest. It is often the case in spatial sampling. In the particular case when the cumulated inclusion probabilities sum to integers, so that $C_{k_i} = i$ for any $i = 1, \dots, n - 1$, pivotal sampling is strictly equivalent to a one-per-stratum stratified sampling design.

3.2 Some preliminary results

We derive some properties which are needed to establish the asymptotic normality of the HT-estimator, by following the approach in Ohlsson (1986). Let $\sigma(X)$ denote the σ -field generated by some random X . We introduce the following σ -fields:

$$\begin{aligned} \mathcal{F}_0 &= \sigma(\tau), \\ \mathcal{F}_i &= \sigma(\tau, S_1, F_1, L_1, \dots, S_i, F_i, L_i) \text{ for } i = 1, \dots, n. \end{aligned} \quad (10)$$

Conditioning on the σ -field \mathcal{F}_i amounts to conditioning on all the random events, up to those in the microstratum U_i . We first state in Proposition 1 that the HT estimator is a sum of martingale increments. The proof of this proposition is given in Appendix A.

Proposition 1. *We can write*

$$\hat{t}_{y\pi} - t_y = \sum_{i=1}^n \xi_i \quad \text{where} \quad \xi_i = \check{y}_{F_i} + b_i \check{y}_{L_i} - \left\{ \sum_{k \in U'_i} \alpha_{ik} \check{y}_k + b_i \check{y}_{k_i} \right\}, \quad (11)$$

and where $U'_i = \{L_{i-1}, k_{i-1} + 1, \dots, k_i - 1, k_i\}$. Also, $\{\xi_i; i = 1, \dots, n\}$ is a martingale difference sequence with respect to the filtration $\{\mathcal{F}_i; i = 0, \dots, n\}$.

We note

$$\eta_i = \frac{\xi_i}{\sqrt{V_{\{\mathcal{F}_0\}}(\hat{t}_{y\pi})}}. \quad (12)$$

We prove the martingale central-limit theorem for the HT-estimator by applying Proposition 4.1 in Ohlsson (1986), for which we need the following conditions:

$$\sum_{i=1}^n E_{\{\mathcal{F}_0\}} |\eta_i|^{2+\delta} \rightarrow 0 \text{ for some } \delta > 0, \quad (13)$$

$$\sum_{i=1}^n E_{\{\mathcal{F}_{i-1}\}} (\eta_i^2) \rightarrow_{Pr} 1, \quad (14)$$

and where \rightarrow_{Pr} stands for the convergence in probability. Under Assumptions (H1) and (H2), these conditions will follow from Propositions 2 and 3. The proof of Proposition 2 is given in Appendix B. The proof of Proposition 3 is lengthy, and is therefore given in the Supplementary Material. We also need to prove that the variance of the HT-estimator is not vanishing, in the sense that its order of magnitude is $O(N^2 n^{-1})$. This will follow from Proposition 4, whose proof is given in Appendix C.

Proposition 2. *We have*

$$E_{\{\mathcal{F}_0\}} \left(\sum_{i=1}^n \xi_i^4 \right) \leq 16 \left\{ 2 + \frac{1}{1 - \pi_M} \right\} \left\{ \sum_{l \in U} \pi_l \left(\check{y}_l - \frac{t_y}{n} \right)^4 \right\}. \quad (15)$$

Proposition 3. *We have*

$$V_{\{\mathcal{F}_0\}} \left\{ \sum_{i=1}^n V_{\{\mathcal{F}_{i-1}\}}(\xi_i) \right\} \leq 8 \left\{ 3 + \frac{2}{1 - \pi_M} \right\} \left\{ 2 + \frac{1}{1 - \pi_M} \right\} \sum_{k \in U} \pi_k \left(\check{y}_k - \frac{t_y}{n} \right)^4. \quad (16)$$

Proposition 4. *We have*

$$V_{\{\mathcal{F}_0\}}(\hat{t}_{y\pi}) \geq \{1 - \pi_M\}^2 \left\{ \sum_{i=1}^n \sum_{k \in U_i} \alpha_{ik} \left(\check{y}_k - \sum_{l \in U_i} \alpha_{il} \check{y}_l \right)^2 \right\}. \quad (17)$$

3.3 A martingale central-limit theorem

We can now formulate our main result. The proof is given in Appendix D.

Theorem 1. *Suppose that the sample S is selected by means of pivotal sampling, and that assumptions (H1)-(H3) hold. Then*

$$\frac{\hat{t}_{y\pi} - t_y}{\sqrt{V_{\{\mathcal{F}_0\}}(\hat{t}_{y\pi})}} \xrightarrow[\mathcal{L}]{} \mathcal{N}(0, 1), \quad (18)$$

where $\xrightarrow[\mathcal{L}]{}$ stands for the convergence in distribution.

From Theorem 1, an approximate two-sided $100(1 - 2\alpha)\%$ confidence interval for t_y is thus given by

$$\left[\hat{t}_{y\pi} \pm u_{1-\alpha} \sqrt{V_{\{\mathcal{F}_0\}}(\hat{t}_{y\pi})} \right] \quad (19)$$

with $u_{1-\alpha}$ the quantile of order $1 - \alpha$ of the standard normal distribution. In practice, the variance of the HT-estimator in (19) needs to be replaced with a variance estimator, which is discussed in Section 3.4.

3.4 Variance estimation

Two customary choices for variance estimation for a fixed-size sampling design are the Sen-Yates-Grundy (SYG) variance estimator

$$v_{SYG}(\hat{t}_{y\pi}) = \frac{1}{2} \sum_{k \neq l \in S} \frac{\pi_k \pi_l - \pi_{kl}}{\pi_{kl}} (\check{y}_k - \check{y}_l)^2, \quad (20)$$

or the Horvitz-Thompson (HT) variance estimator

$$v_{HT}(\hat{t}_{y\pi}) = \sum_{k,l \in S} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}, \quad (21)$$

with π_{kl} the probability that units k and l are selected jointly in the sample. These two variance estimators are unbiased if and only if all the second-order inclusion probabilities are strictly positive. Otherwise, $v_{SYG}(\hat{t}_{y\pi})$ is biased downwards, while $v_{HT}(\hat{t}_{y\pi})$ is biased upwards for a variable of interest with positive values.

In case of pivotal sampling, many second-order inclusion probabilities are equal to 0, since in particular two non cross-border units inside a same microstratum U_i may not be selected together in the sample. Both the SYG variance estimator and the HT variance estimator may therefore be severely biased. Another drawback is that the second-order inclusion probabilities are required. Computing these probabilities is possible for ordered pivotal sampling (see Chauvet, 2012), but this requires the knowledge of the original ranking τ and of the first-order inclusion probabilities for all the units in U . Consequently, this computation may be impossible for a data user with limited knowledge of the sampling frame.

We propose an alternative variance estimator which does not make use of the second-order inclusion probabilities. This variance estimator is

$$v_{DIFF}(\hat{t}_{y\pi}) = \sum_{i=1}^{\lfloor n/2 \rfloor} (1 + \delta_i) (\check{y}_{F_{2i}} - \check{y}_{F_{2i-1}})^2 + (\check{y}_{F_n} - \check{y}_{F_{n-1}})^2 1(n \text{ is odd}),$$

$$\text{where } \delta_i = \frac{b_{2i-1}c_{2i-1} + c_{2i}}{1 - c_{2i}}, \quad (22)$$

where the quantities b_i and c_i are given in equations (3) and (4), where $\lfloor \cdot \rfloor$ stands for the integer part, and where $1(\cdot)$ stands for the indicator function. Roughly speaking, the i^{th} term of the sum in equation (22) accounts for the variance in the microstrata U_{2i-1} and U_{2i} , and $(\check{y}_{F_n} - \check{y}_{F_{n-1}})^2$ is a correction term to account for the variance in the last microstratum U_n if n is odd.

Theorem 2. *We have*

$$E_{\{\mathcal{F}_0\}}\{v_{DIFF}(\hat{t}_{y\pi})\} \geq V_{\{\mathcal{F}_0\}}(\hat{t}_{y\pi}). \quad (23)$$

The proof of Theorem 2 is given in Appendix E. The proposed variance estimator enables to compute conservative confidence intervals. The multinomial variance estimator

$$v_{MULT}(\hat{t}_{y\pi}) = \frac{n}{n-1} \sum_{k \in S} \left(\frac{y_k}{\pi_k} - \frac{\hat{t}_{y\pi}}{n} \right)^2 \quad (24)$$

is another possible conservative variance estimator, see Chauvet (2017). The proposed variance estimator $v_{DIFF}(\hat{t}_{y\pi})$ better accounts for the features of the sampling design, and we therefore expect it to be less conservative. This will be evaluated in Section 5 through a simulation study. From a close look at the proof of Theorem 2, the bias of $v_{DIFF}(\hat{t}_{y\pi})$ will be small if the means in consecutive microstrata are close, in the sense that for any $i = 1, \dots, \lfloor n/2 \rfloor$:

$$\sum_{k \in U_{2i-1}} \alpha_{2i-1,k} \check{y}_k \simeq \sum_{k \in U_{2i}} \alpha_{2i,k} \check{y}_k. \quad (25)$$

In many situations, the proposed variance estimator may be further simplified by omitting the factors δ_i . It can be shown that

$$\delta_i \leq \frac{\pi_M^2(1 + \pi_M)}{2(2 - \pi_M)}, \quad (26)$$

where π_M is the maximum inclusion probability, see equation (1). The proof of equation (26) is given in the Supplementary Material. Consequently, with moderately large inclusion probabilities no greater than 0.35, the factors δ_i will be no greater than 0.05. In such case, they may be safely ignored which leads to the simplified variance estimator

$$v_{DIFF2}(\hat{t}_{y\pi}) = \sum_{i=1}^{\lfloor n/2 \rfloor} (\check{y}_{F_{2i}} - \check{y}_{F_{2i-1}})^2 + (\check{y}_{F_n} - \check{y}_{F_{n-1}})^2 \mathbf{1}(n \text{ is odd}). \quad (27)$$

This variance estimator only requires the knowledge of 1) the values of the variable of interest for the selected units, along with their inclusion probabilities, and 2) their rank of selection in the sampling process. It can be therefore easily computed by a data user.

4 Application to spatial sampling

The degree of spatial balance of a sampling design is very important in order to limit a lack of efficiency due to positive spatial auto-correlation between units. The interest in spatial sampling has increased in the last decade, see for example Stevens and Olsen (2004); Grafström et al. (2012); Grafström and Tillé (2013); Dickson and Tillé (2016). It has led to applications in various domains, including the drawing of primary sampling units in the context of household surveys (see Favre-Martinoz and Merly-Alpa, 2017).

In Section 4.1, we give a brief overview of the Generalized Random Tessellation Stratified (GRTS) sampling design. Though this sampling method is popular in practice, some basic statistical results like the consistency of the HT estimator and the asymptotic normality are difficult to prove. Therefore, we introduce a modification in Section 4.2 that we call the Pivotal Tessellation Method (PTM). By substituting pivotal sampling to systematic sampling at the selection process, this new sampling algorithm makes sure that the required statistical properties hold true. Also, it enables to significantly reduce the computational time needed to select a sample, see Section 6 for an illustration.

4.1 The GRTS sampling design

The GRTS design (Stevens and Olsen, 2004) is one of the most popular spatial sampling methods. It is suitable to select a sample in several situations, including that of a finite discrete population (e.g., trees within a forest), of a linear continuous population (e.g., rivers), or of an areal continuous population (e.g., forests or lakes). In this Section, we describe the GRTS design for a discrete two-dimensional population. The main idea is to use some function that maps a two-dimensional space into one dimension, while preserving some proximity relationships between units. A sample is then selected in the one-dimensional space through systematic sampling.

To apply the GRTS design, the two-dimensional space under study is first mapped to the unit square $[0; 1] \times [0; 1]$. This unit square is then mapped to a one-dimensional interval, by using a function $f(\cdot)$ which preserves two-dimensional proximity relationships. For this purpose, Stevens and Olsen (2004) propose to use quadrant-recursive functions (Mark, 1990) which en-

sure that, when recursively decomposing a rectangular region into sub-quadrants, the image of any sub-quadrant is an interval. In this case, the function $f(\cdot)$ can be seen as the limit of successive intensifications of a grid covering the unit square, where the square is divided into four sub-squares, each of which being subsequently divided into four sub-squares, and so on. In Section 4.2, we propose a simple way to obtain such tessellation of the unit square in cells, by using the decomposition of a number in Bit code.

The quadrant-recursive function $f(\cdot)$ maps each cell to a so-called address, which is a decimal number on the one-dimensional interval resulting from the order in which the divisions are carried out. This mapping preserves the proximity relationships between sampling units, in the sense that consecutive cells in the two-dimensional space have consecutive addresses on the unit line. Prior to sampling, the cells may then be randomized within each quadrant to gain entropy in the selection process; a so-called hierarchical randomization is obtained if the permutations are independent from one sub-quadrant to another (Stevens and Olsen, 2004).

Finally, a sample of cells is selected through systematic sampling of addresses on the line. Stevens and Olsen (2004) proved that the GRTS sampling design matches the required first-order inclusion probabilities, and leads to a spatially balanced sample. However, the statistical properties (a)-(c) are fairly difficult to prove when using a systematic sampling design, even when the units are randomized.

4.2 The Pivotal Tessellation Method

We propose a modification of the GRTS method where pivotal sampling is used in replacement of systematic sampling. Like for the GRTS design, we use some quadrant-recursive function to map the two-dimensional space into one dimension, and a sample is then selected on a one-dimensional line by means of pivotal sampling. This leads to the selection of a spatially balanced sample, while matching the required first-order inclusion probabilities. Also, from the results in Section 3, the HT-estimator is consistent and asymptotically normally distributed, and a conservative variance estimator may be easily produced.

We now present an efficient way to obtain a tessellation of the space under

study, by using the decomposition of a number in Bit code which is readily obtained in \mathbf{R} . The two-dimensional space is mapped by Euclidean transformations to the square $[0, 2^{31} - 1] \times [0, 2^{31} - 1]$: only 31 out of the 32 positions in the decomposition are useful, since the first position is always 0 for positive numbers. These 31 positions are successively considered, to obtain an intensification of the grid by subdividing each square previously obtained in four sub-squares.

This division is obtained as follows: if some point in the square has coordinates with on i^{th} position $(x_i, y_i) \in \{0, 1\}^2$, then the corresponding position in the address is $y_i + 2x_i \in \{0, 1, 2, 3\}$. For example, if we have on i^{th} position $(x_i, y_i) = (1, 0)$, then the corresponding position in the address is 2. This leads to an address in 31 positions. A cell may contain several sampling units, but the proposed method leads to a very fine tessellation with $4^{31} \approx 4.6 \cdot 10^{18}$ addresses, making this case fairly unlikely. The proposed tessellation may be easily generalized to spaces of dimension $d \geq 3$, which can be of interest in a factorial space, for example (Le Gleut, 2017).

The address in 31 positions that we obtain define a mapping between the two-dimensional space and a line which preserves the proximity relationships between sampling units. A sample is obtained by applying the pivotal method. Here again, the cells obtained in the tessellation may be randomized prior to sampling to gain entropy. However, the use of pivotal sampling guarantees that the HT-estimator is consistent and asymptotically normally distributed, even without this randomization.

4.3 An illustration of the proposed method

To fix ideas, we apply the proposed method on a small two-dimensional population. The whole process is given in Figure 2. The mapping of the space on a square and the ranking of the units in the population is described in the top part. The mapping on a one-dimensional line and the sample selection by means of the pivotal method is described in the medium part. The mapping back to the selected points in the original space is described in the bottom part.

The population under study contains $N = 16$ units (first scheme from the left, top part of Figure 2), where we wish to select a spatially balanced sam-

ple of size $n = 4$ with equal inclusion probabilities $\pi_k = 1/4$. This population is mapped into a square (second scheme). In view of the small size of the population, the fine tessellation on the square $[0, 2^{31} - 1] \times [0, 2^{31} - 1]$ is not required: to simplify the presentation, the population is mapped into the square $[0, 2^2 - 1] \times [0, 2^2 - 1]$. The coordinates of each point are then considered, and the two first positions of their decomposition in Bit code are used to obtain addresses ranging from 00 to 33 (third scheme). This defines a path between the units in the square (fourth scheme).

The two-dimensional space is then mapped on a one-dimensional line of length $n = 4$, where each point k is represented by a segment of length π_k , and where the addresses of the units define their ranking on the line (medium part of Figure 2). A sample of n points, represented as gray points, is then selected by means of pivotal sampling. In this particular case, pivotal sampling amounts to a one-per stratum stratified sampling design where each stratum is made of four consecutive points. In this example, the addresses 01, 12, 21 and 32 are selected. For comparison, a similar sample selection by means of the GRTS sampling design is described in the Supplementary Material.

5 Simulation study

In this Section, we present some simulation results. We first use an artificial population introduced in Example 5 of Grafström et al. (2012). We then use the Meuse dataset available in the **R** *package* `gstat`, and considered in Grafström and Tillé (2013). We compare the performances of the proposed Pivotal Tessellation Method (PTM) with alternative spatial sampling designs. The simulation set-up is described in Section 5.1. The results of the simulation study for the two populations are given in Sections 5.2 and 5.3, respectively.

5.1 Simulation set-up

The sampling designs that we consider as competitors are the Generalized Random Tessellation Stratified sampling design (GRTS); the two versions of the Local Pivotal Method (LPM1 and LPM2; see Grafström et al., 2012); the pivotal method through Traveling Salesman Problem order (TSP; see Dickson and Tillé, 2016); and the Conditional Poisson Sampling design (CPS; see

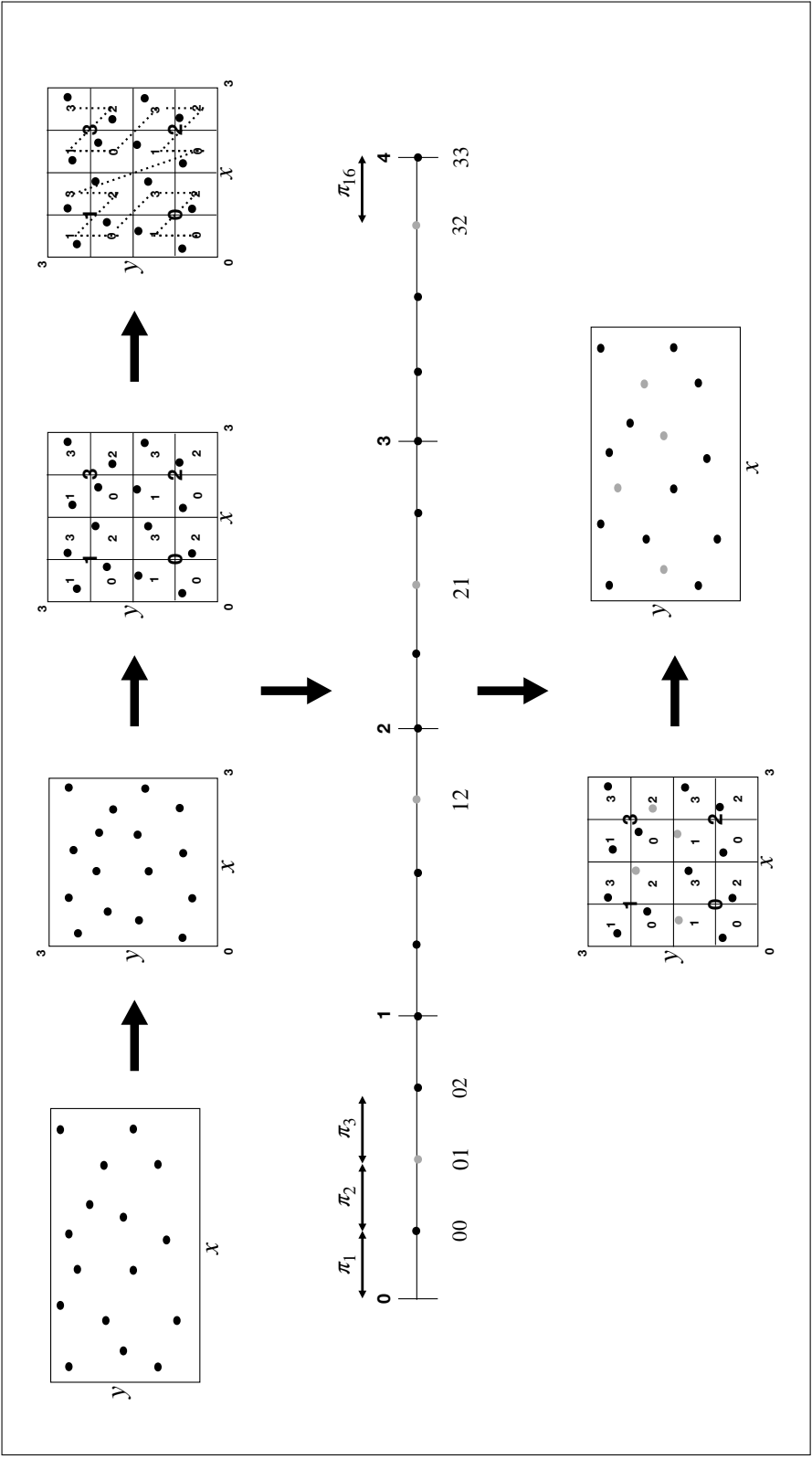


Figure 2: Selection process on a two-dimensional population by means of the Pivotal Tesselation Method

Hájek, 1964). In order to implement the pivotal method and the CPS design, we use the **R** *package* `Sampling`. To solve the traveling salesman problem, we use the algorithm "2-Opt" of the **R** *package* `TSP`. The GRTS design is implemented through the **R** *package* `spsurvey`, and the LPM1 and LPM2 are implemented through the **R** *package* `BalancedSampling`.

In both simulation studies, we are interested in the spatial balance of the sampling designs, using the approach of Voronoi polygons suggested by Stevens and Olsen (2004). For a given sample s , the Voronoi polygon for some sampled unit k includes all units in the population which are closer to k than to any other sampled unit. The quantity

$$\Delta(s) = \frac{1}{n} \sum_{k \in s} (\delta_k - 1)^2 \quad (28)$$

is used as a measure of spatial balance, with δ_k the sum of the inclusion probabilities of all units in the polygon associated to k . If the sample is spatially balanced, it is expected that all the δ_k 's are close to 1, and that $\Delta(s)$ is small. From both populations, we select $B = 10,000$ samples by means of the PTM, GRTS, LPM1, LPM2, TSP and CPS. In each sample s_b , $b = 1, \dots, 10,000$, we compute the quantity $\Delta(s_b)$. As a measure of spatial balance of the sampling design, we compute their Monte Carlo Mean

$$E_{MC}(\Delta) = \frac{1}{B} \sum_{b=1}^B \Delta(s_b). \quad (29)$$

In both simulation studies, we are also interested in the variance of the HT-estimator which is evaluated by the Monte Carlo Variance

$$V_{MC}(\hat{t}_{y\pi}) = \frac{1}{B} \sum_{b=1}^B \left\{ \hat{t}_{y\pi}(s_b) - \frac{1}{B} \sum_{c=1}^B \hat{t}_{y\pi}(s_c) \right\}^2, \quad (30)$$

with $\hat{t}_{y\pi}(s_b)$ the value of the HT-estimator evaluated on the sample s_b , $b = 1, \dots, 10,000$.

Finally, we are interested in comparing variance estimators for the proposed PTM. To measure the performances of some variance estimator $v(\hat{t}_{y\pi})$, we

compute its Monte Carlo Mean

$$E_{MC}\{v(\hat{t}_{y\pi})\} = \frac{1}{B} \sum_{b=1}^B v\{\hat{t}_{y\pi}(s_b)\} \quad (31)$$

where $v\{\hat{t}_{y\pi}(s_b)\}$ denotes the variance estimator in the b -th sample. We also compute the percent relative stability

$$RS_{MC}\{v(\hat{t}_{y\pi})\} = 100 \times \frac{\left[B^{-1} \sum_{b=1}^B \{v(\hat{t}_{y\pi}(s_b)) - E_{MC}(v(\hat{t}_{y\pi}))\}^2 \right]^{1/2}}{V_{MC}(\hat{t}_{y\pi})} \quad (32)$$

We compare the performance of the proposed simplified variance estimator v_{DIFF2} given in (27) with the Sen-Yates-Grundy (SYG) variance estimator given in equation (20), and the Horvitz-Thompson (HT) variance estimator given in equation (21). We also consider the Hájek-Rósen (HR) variance estimator

$$v_{HR}(\hat{t}_{y\pi}) = \frac{n}{n-1} \sum_{k \in S} (1 - \pi_k) \left(\frac{y_k}{\pi_k} - \hat{R} \right)^2 \quad \text{with } \hat{R} = \frac{\sum_{k \in S} \frac{y_k}{\pi_k} (1 - \pi_k)}{\sum_{k \in S} (1 - \pi_k)}. \quad (33)$$

Finally, we consider stratified multinomial variance estimators

$$v_{MULT_h}(\hat{t}_{y\pi}) = \sum_{i=1}^p \frac{h}{h-1} \sum_{k \in S_i} \left(\frac{y_k}{\pi_k} - \frac{1}{h} \sum_{l \in S_i} \frac{y_l}{\pi_l} \right)^2. \quad (34)$$

This is the variance estimator we would use if the population was stratified into $p = n/h$ strata, with selection of a sample S_i of h units by multinomial sampling inside the i^{th} stratum. We compute $v_{MULT_h}(\hat{t}_{y\pi})$ for several values of h . The multinomial variance estimator in (24) is a particular case obtained with $h = n$.

5.2 Results of the first simulation study

The first population that we consider is introduced in Example 5 of Grafström et al. (2012). It is obtained by dividing the unit square according to a regular 20×20 grid, resulting in a population of $N = 400$ units. For any unit k , the variable of interest y_k is the area within the cell under the function $f(x_1, x_2) = 3(x_1 + x_2) + \sin\{6(x_1 + x_2)\}$.

Table 1: Monte Carlo Mean of the spatial balance and Monte Carlo Variance of the Horvitz-Thompson estimator for Population 1

	PTM	GRTS	LPM1	LPM2	TSP	CPS
	$E_{MC}(\Delta)$					
$n = 16$	0.07	0.12	0.08	0.09	0.11	0.33
$n = 32$	0.08	0.11	0.07	0.07	0.10	0.30
$n = 48$	0.09	0.11	0.07	0.07	0.10	0.29
	$V_{MC}(\hat{t}_{y\pi}) (\times 100)$					
$n = 16$	1.53	2.49	1.94	1.96	2.65	12.48
$n = 32$	0.39	0.89	0.54	0.57	0.65	6.18
$n = 48$	0.16	0.34	0.26	0.27	0.28	3.91

For comparability, we use the same simulation set-up than in Grafström et al. (2012), selecting samples of size $n = 16, 32$ or 48 with equal probabilities. Note that in case of sampling with equal probabilities, Conditional Poisson Sampling amounts to simple random sampling. Stevens and Olsen (2004) underlined that their method should perform better in terms of spatial balance for sample sizes which are multiples of 4. Therefore, this simulation set-up is expected to be favorable for GRTS, and presumably for PTM.

The simulation results for the spatial balance and the variance of the HT-estimator are given in Table 1. As expected, all sampling designs that use spatial auxiliary information produce much more balanced samples than the conditional Poisson sampling design which does not. PTM, LPM1 and LPM2 are the best methods in terms of spatial balance, with LPM1 performing slightly better. The proposed PTM performs best in terms of variance.

The simulation results for the possible variance estimators for PTM are given in Table 2. All variance estimators are biased, since the PTM leads to several second order inclusion probabilities that are equal to zero. The HT variance estimator is heavily positively biased, and the SYG variance estimator is heavily negatively biased. For $n = 16$, the SYG variance estimator is equal to zero because in this case, the sampling design amounts to stratified simple random sampling of size 1 inside each stratum. Therefore, we have $\pi_{kl} = 0$ for two units k and l in the same stratum, and $\pi_{kl} = \pi_k \pi_l$ otherwise.

Table 2: Monte Carlo Mean and Relative Stability of the variance estimators for Population 1

		v_{HT}	v_{SYG}	v_{HR}	v_{DIFF2}	v_{MULT2}	v_{MULT4}	v_{MULT}
	$V_{MC}(\hat{t}_{y\pi}) (\times 100)$	$E_{MC}\{v(\hat{t}_{y\pi})\} (\times 100)$						
$n = 16$	1.53	66.57	0.00	13.26	5.46	5.40	6.63	13.82
$n = 32$	0.39	29.25	0.04	6.19	1.16	1.15	2.19	6.73
$n = 48$	0.16	17.12	0.03	3.91	0.36	0.35	0.56	4.45
		$RS_{MC}\{v(\hat{t}_{y\pi})\}$						
$n = 16$		306	0	93	149	150	90	97
$n = 32$		745	18	94	104	107	111	101
$n = 48$		1499	17	95	71	71	77	108

The five other variance estimators are all positively biased. Among them, the proposed estimator v_{DIFF2} and the estimator v_{MULT2} perform similarly and present the best results with the smallest bias. Their relative stability is larger than that of v_{HR} for small sample sizes, but is smaller for $n = 48$. Overall, v_{HR} is slightly better than v_{MULT} , but the estimators v_{MULT2} and v_{MULT4} are less biased with comparable or better stability.

5.3 Results of the second simulation study

The second population that we consider is the "Meuse" data set available in the **R** package `gstat`. It gives locations and top soil heavy metal concentrations (ppm) collected in a flood plain of the river Meuse, sampled from an area of approximately $15 \text{ m} \times 15 \text{ m}$. The variables that we consider are the topographical map coordinates (`x` and `y`), the topsoil concentration in cadmium (`cadmium`), copper (`copper`), lead (`lead`) and zinc (`zinc`), the relative elevation (`elev`) and the percentage of organic matter (`om`).

As explained by Grafström and Tillé (2013), this data set exhibits an important spatial correlation. The computation of Moran's I leads to the same conclusion. The sampling design that we use consists in selecting 50 among the $N = 164$ locations in the data set, with probabilities proportional to the copper concentration. In view of the high correlations between the concentrations in heavy metals (see Figure 3), the variance for the estimation of the

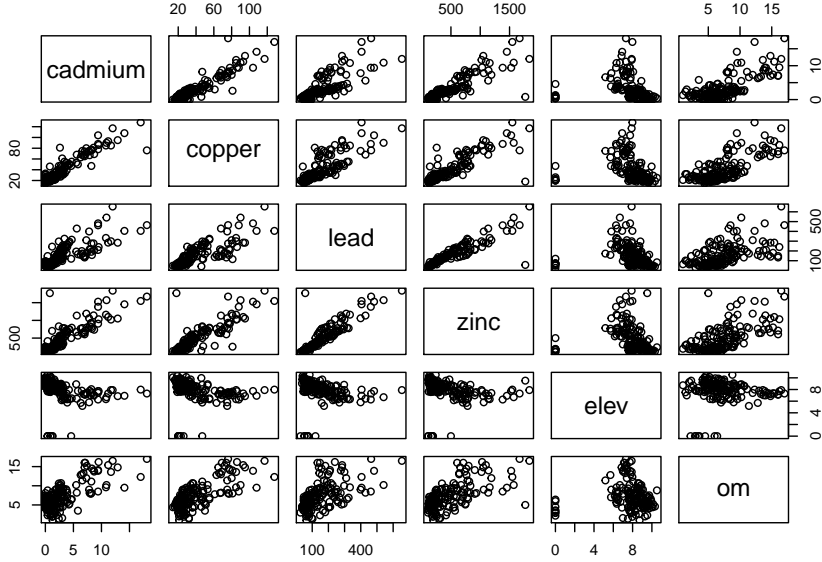


Figure 3: Scatter plot matrix of the variables in the Meuse data set

total of these variables is expected to be small.

The simulation results for the spatial balance and the variance of the HT-estimator are given in Table 3. LPM1 and LPM2 produce the most balanced samples, followed by GRTS and TSP, while PTM performs slightly worse. Here again, all spatial sampling designs produce much more balanced samples than CP. In terms of variance, the five spatial sampling designs show comparable results, with LPM1 and LPM2 performing slightly better.

The simulation results for the possible variance estimators for PTM are given in Table 4. The HT-variance estimator is heavily positively biased, and the SYG-variance estimator is negatively biased, as expected. The six other variance estimators are all positively biased, and among them v_{HR} presents the best results with the smallest bias and the smallest relative stability. Among the five other estimators, the proposed estimator v_{DIFF2} and the estimator v_{MULT2} perform similarly and present the smallest bias, but are slightly more unstable. We observe that the variance estimators are particularly unstable for *zinc*, the value of the fourth central moment for this variable being particularly huge in the data set.

Table 3: Monte Carlo Mean of the spatial balance and Monte Carlo Variance of the Horvitz-Thompson estimator for Population 2

		PTM	GRTS	LPM1	LPM2	TSP	CPS
		$E_{MC}(\Delta)$					
		0.18	0.16	0.13	0.14	0.16	0.29
		$V_{MC}(\hat{t}_{y\pi})$					
zinc	($\times 10^{-7}$)	2.06	2.03	1.98	2.00	2.02	2.14
lead	($\times 10^{-5}$)	5.79	5.77	5.33	5.42	5.63	9.38
cadmium	($\times 10^{-2}$)	5.14	5.83	5.93	5.77	5.88	7.92
elev	($\times 10^{-3}$)	6.90	6.03	5.27	5.43	5.37	9.24
om	($\times 10^{-3}$)	2.24	2.18	1.97	1.99	2.06	2.71

6 Computational time

In this Section, we compared the computational time needed to select a sample by means of spatial sampling designs. We worked on a remote server Intel(R) Xeon(R) CPU E5-2640 v4 @ 2.40GHz with 80 Go RAM. We used various population sizes, with equal probabilities of $\pi_k = 1/10$. The sampling designs considered are the proposed PTM, GRTS, LPM1, LPM2 and TSP. The coordinates x and y of the units in the population were generated independently from a standard normal distribution.

The results are presented in Table 5, with NC (not computed) indicating the cases where the computational resources were not sufficient to allow the selection of a sample. We first note that TSP is not feasible for the medium or the large population, while GRTS is not feasible for the large population. The computational time for PTM, LPM1 and LPM2 is considerably lower than that for GRTS, even in the small population. The computational time for PTM is approximately the same than for LPM1 and LPM2 for the small population, but is considerably reduced for the medium and the large populations.

Table 4: Monte Carlo Mean and Relative Stability of the variance estimators for Population 2

	v_{HT}	v_{SYG}	v_{HR}	v_{DIFF2}	v_{MULT2}	v_{MULT5}	v_{MULT10}	v_{MULT}
	$V_{MC}(\hat{t}_{y\pi})$							
zinc	4.42	0.81	2.21	2.42	2.44	2.66	2.69	2.90
lead	33.68	3.85	9.25	10.63	10.59	12.41	12.58	15.76
cadmium	10.74	3.94	7.88	10.59	10.40	14.53	13.78	14.53
elev	23.17	4.28	9.19	12.21	11.85	14.32	15.22	14.69
om	10.98	1.32	2.67	3.49	3.45	3.71	3.81	4.19
	$E_{MC}\{v(\hat{t}_{y\pi})\}$							
zinc	259	297	183	202	209	207	206	213
lead	481	70	23	46	47	39	37	34
cadmium	205	48	28	54	51	68	56	51
elev	131	56	25	48	45	42	38	35
om	281	38	19	42	43	33	29	26
	$RS\{v(\hat{t}_{y\pi})\}$							

Table 5: Computational time to select a sample for 5 spatial sampling designs and 3 populations

Population size	PTM	GRTS	LPM1	LPM2	TSP
$N = 10,000$	0.3s	36.5s	1.4s	0.6s	18h30m
$N = 100,000$	2s	1h30m	1m8s	43s	NC
$N = 1,000,000$	21s	NC	4h8m	1h26m	NC

7 Conclusion

The pivotal method is widely used in spatial sampling designs since it avoids selecting neighbouring units. It does not belong to the class of large entropy sampling designs, and basic statistical properties such as the asymptotic normality of the Horvitz-Thompson estimator do not follow from general results obtained for this type of sampling designs (Berger, 1998, 2011). In this paper, we proved the asymptotic normality of the HT-estimator under mild assumptions. We also proposed a very simple variance estimator which does not require second-order inclusion probabilities. This variance estimator is very simple to compute for a data user, and enables computing conservative confidence intervals.

Among the spatial sampling designs proposed in the literature, the Generalized Random Tessellation Stratified (GRTS) sampling design is widely used but its statistical properties have not been investigated. We proposed a modification of the GRTS sampling design, by replacing the systematic sampling step with a pivotal sampling step. The proposed Pivotal Tessellation Method (PTM) enjoys very good statistical properties, namely consistency and asymptotic normality of the HT-estimator and availability of a very simple conservative variance estimator. Also, our simulation results indicate that PTM is very competitive both in terms of spatial balance and of accuracy of estimators. We also proposed a very efficient way to obtain a tessellation of the space under study with the proposed method. Among the spatial sampling designs that we consider, PTM appears as the less computationally intensive and enables selecting spatially balanced samples, even for very large populations.

The statistical properties established in Section 3 hold true for any population to which ordered pivotal sampling is applied, after a ranking of the units with respect to some criterion. This is in particular true for the pivotal method through Traveling Salesman Problem (TSP), see Dickson and Tillé (2016). However, our results do not hold for the Local Pivotal Method (LPM, see Grafström et al., 2012), since the ranking of the units is not fixed in advance, but varies during the sampling procedure. The study of similar statistical properties for the LPM is a very challenging problem for the future.

8 Supplementary Material

The four supplemental files are contained in a single archive.

README: description of the supplemental files. (txt file)

Functions: Functions for the tessellation of the space and the mapping in one dimension. (R file)

Example: An example of sample selection by means of the Pivotal Tessellation Method (calls Functions). (R file)

SuppMaterial: Additional proofs of results. (pdf file)

Acknowledgments

The authors would like to thank Martin Chevalier (Insee) for his help in programming an efficient implementation of the proposed Pivotal Tessellation Method in **R**.

Appendix

A Proof of Proposition 1

Using the identity

$$\sum_{k \in U'_i} \alpha_{ik} \check{y}_k = b_{i-1} \check{y}_{L_{i-1}} + \sum_{k_{i-1} < k < k_i} \alpha_{ik} \check{y}_k + a_i \check{y}_{k_i},$$

we can write

$$\sum_{i=1}^n \xi_i = \sum_{i=1}^n (\check{y}_{F_i} + b_i \check{y}_{L_i} - b_{i-1} \check{y}_{L_{i-1}}) + \sum_{i=1}^n \left\{ \sum_{k_{i-1} < k < k_i} \alpha_{ik} \check{y}_k + (a_i + b_i) \check{y}_{k_i} \right\}. \quad (35)$$

We first consider the first term in the right-hand side of (35), which we can write as

$$\begin{aligned} \sum_{i=1}^n (\check{y}_{F_i} + b_i \check{y}_{L_i} - b_{i-1} \check{y}_{L_{i-1}}) &= \sum_{i=1}^n \check{y}_{F_i} + \sum_{i=1}^n (b_i \check{y}_{L_i} - b_{i-1} \check{y}_{L_{i-1}}) \\ &= \hat{t}_{y\pi} + (b_n \check{y}_{L_n} - b_0 \check{y}_{L_0}) \\ &= \hat{t}_{y\pi}. \end{aligned} \quad (36)$$

We now consider the second term in the right-hand side of (35), which we can write as

$$\begin{aligned} \sum_{i=1}^n \left\{ \sum_{k_{i-1} < k < k_i} \alpha_{ik} \check{y}_k + (a_i + b_i) \check{y}_{k_i} \right\} &= \sum_{i=1}^n \sum_{k_{i-1} < k \leq k_i} y_k \\ &= t_y. \end{aligned} \quad (37)$$

By plugging (36) and (37) into (35), we obtain (11). Also, from equation (9), we obtain

$$E_{\{L_{i-1}, S_i\}}(\check{y}_{F_i} + b_i \check{y}_{L_i}) = (1 - a_i) \check{y}_{S_i} + (a_i + b_i) \check{y}_{k_i} \quad (38)$$

and from the definition of ξ_i and S_i , we obtain

$$\begin{aligned} E_{\{\mathcal{F}_{i-1}\}}(\xi_i) &= E_{\{L_{i-1}\}}(\xi_i) = E_{\{L_{i-1}\}}E_{\{L_{i-1}, S_i\}}(\xi_i) \\ &= E_{\{L_{i-1}\}} \left\{ (1 - a_i)\check{y}_{S_i} - \sum_{k \in U'_i \setminus \{k_i\}} \alpha_{ik}\check{y}_k \right\} = 0. \end{aligned}$$

B Proof of Proposition 2

B.1 Preliminary Lemmas

The proofs are given in the Supplementary Material.

Lemma 1. *We can write*

$$\xi_i = b_i(\check{y}_{S_i} - \check{y}_{F_i}) + \sum_{k \in U'_i} \alpha_{ik}(\check{y}_{F_i} - \check{y}_k) \quad (39)$$

$$= (1 - b_i) \sum_{k \in U'_i} \alpha_{ik}(\check{y}_{F_i} - \check{y}_k) + b_i \sum_{k \in U'_i} \alpha_{ik}(\check{y}_{S_i} - \check{y}_k). \quad (40)$$

Lemma 2. *For any convex function ϕ , we have:*

$$\phi(\xi_i) \leq (1 - b_i) \sum_{k \in U'_i} \alpha_{ik}\phi(\check{y}_{F_i} - \check{y}_k) + b_i \sum_{k \in U'_i} \alpha_{ik}\phi(\check{y}_{S_i} - \check{y}_k), \quad (41)$$

$$E_{\{L_{i-1}\}}\phi(\xi_{vi}) \leq \sum_{k, l \in U'_i} \alpha_{ik}\alpha_{il}\phi(\check{y}_l - \check{y}_k). \quad (42)$$

Lemma 3. *For any $i \leq j = 1, \dots, n$, we note $c_i = \frac{a_i b_i}{(1-a_i)(1-b_i)}$, and*

$$c(i, j) = \begin{cases} 1 & \text{if } j = i, \\ \prod_{l=i}^{j-1} c_l & \text{if } j > i. \end{cases} \quad (43)$$

We have

$$\sum_{j=i}^n c(i, j) \leq 1 + \frac{1}{1 - \pi_M} \quad \text{and} \quad \sum_{i=1}^j c(i, j) \leq 1 + \frac{1}{1 - \pi_M}. \quad (44)$$

Lemma 4. *We note z for some variable of interest, and $z_i^+ = \max(z_i, 0)$.*

Then

$$E_{\{\mathcal{F}_0\}} \left(\sum_{j=1}^{n-1} b_j z_{L_j} \right) \leq \left\{ 1 + \frac{1}{1 - \pi_M} \right\} \sum_{l \in U} \pi_l z_l^+. \quad (45)$$

B.2 Proof of Proposition 2

We can write

$$\sum_{i=1}^n E_{\{\mathcal{F}_0\}}(\xi_i^4) = E_{\{\mathcal{F}_0\}} \left[\sum_{i=1}^n E_{\{\mathcal{F}_0, L_{i-1}\}}(\xi_i^4) \right]. \quad (46)$$

Applying equation (42) in Lemma 2 with $\phi(x) = x^4$, we obtain

$$\begin{aligned} \sum_{i=1}^n E_{\{\mathcal{F}_0\}}(\xi_i^4) &\leq E_{\{\mathcal{F}_0\}} \left[\sum_{i=1}^n \sum_{k, l \in U'_i} \alpha_{ik} \alpha_{il} (\check{y}_l - \check{y}_k)^4 \right] \\ &= E_{\{\mathcal{F}_0\}} \left[\sum_{i=1}^n \sum_{k, l \in U'_i} \alpha_{ik} \alpha_{il} \left\{ \left(\check{y}_l - \frac{t_y}{n} \right) - \left(\check{y}_k - \frac{t_y}{n} \right) \right\}^4 \right] \\ &\leq 16 E_{\{\mathcal{F}_0\}} \left[\sum_{i=1}^n \sum_{k, l \in U'_i} \alpha_{ik} \alpha_{il} \left(\check{y}_k - \frac{t_y}{n} \right)^4 \right] \\ &= 16 E_{\{\mathcal{F}_0\}} \left[\sum_{i=1}^n \sum_{k \in U'_i} \alpha_{ik} \left(\check{y}_k - \frac{t_y}{n} \right)^4 \right]. \end{aligned} \quad (47)$$

We note $z_k = \left(\check{y}_k - \frac{t_y}{n} \right)^4$. From equation (47), we obtain

$$\begin{aligned} \sum_{i=1}^n E_{\{\mathcal{F}_0\}}(\xi_i^4) &\leq 16 E_{\{\mathcal{F}_0\}} \left[\sum_{i=1}^n \left\{ b_{i-1} z_{L_{i-1}} + \sum_{k \in U_i} \alpha_{ik} z_k \right\} \right] \\ &\leq 16 \left\{ E_{\{\mathcal{F}_0\}} \left(\sum_{i=1}^n b_{i-1} z_{L_{i-1}} \right) + \sum_{l \in U} \pi_l z_l \right\}. \end{aligned} \quad (48)$$

By applying Lemma 4, we obtain

$$E_{\{\mathcal{F}_0\}} \left(\sum_{i=1}^n \xi_i^4 \right) \leq 16 \left\{ 2 + \frac{1}{1 - \pi_M} \right\} \left\{ \sum_{l \in U} \pi_l \left(\check{y}_l - \frac{t_y}{n} \right)^4 \right\}.$$

C Proof of Proposition 4

C.1 Preliminary Lemma

The proof is given in the Supplementary Material.

Lemma 5. *We have*

$$\begin{aligned} V_{\{L_{i-1}\}}(\xi_i) &= \sum_{k < l \in U'_i \setminus \{k_i\}} \alpha_{ik} \alpha_{il} (\check{y}_k - \check{y}_l)^2 \\ &+ \frac{1 - a_i - b_i}{1 - a_i} \sum_{k \in U'_i \setminus \{k_i\}} \alpha_{ik} a_i (\check{y}_k - \check{y}_{k_i})^2. \end{aligned} \quad (49)$$

C.2 Proof of Proposition 4

Since the ξ_i 's define a martingale difference sequence with respect to the filtration $\{\mathcal{F}_i; i = 0, \dots, n\}$, we have $E_{\{\mathcal{F}_{i-1}\}}(\xi_i) = 0$, which leads to

$$V_{\{\mathcal{F}_0\}}(\hat{t}_{y\pi}) = \sum_{i=1}^n V_{\{\mathcal{F}_0\}}(\xi_i) = \sum_{i=1}^n E_{\{\mathcal{F}_0\}} V_{\{\mathcal{F}_{i-1}\}}(\xi_i). \quad (50)$$

From Lemma 5, we obtain

$$\begin{aligned} V_{\{\mathcal{F}_{i-1}\}}(\xi_i) &\geq (1 - a_i - b_i) \left\{ \sum_{k < l \in U'_i \setminus \{k_i\}} \alpha_{ik} \alpha_{il} (\check{y}_k - \check{y}_l)^2 \right. \\ &\quad \left. + \sum_{k \in U'_i \setminus \{k_i\}} \alpha_{ik} a_i (\check{y}_k - \check{y}_{k_i})^2 \right\} \\ &= (1 - a_i - b_i) \sum_{k, l \in U'_i} \alpha_{ik} \alpha_{il} (\check{y}_k - \check{y}_l)^2 \\ &= (1 - a_i - b_i) \sum_{k \in U'_i} \alpha_{ik} \left(\check{y}_k - \sum_{l \in U'_i} \alpha_{il} \check{y}_l \right)^2. \end{aligned} \quad (51)$$

Since $L_{i-1} = k_{i-1}$ with probability $\frac{1-a_{i-1}-b_{i-1}}{1-b_{i-1}}$, we obtain from (51)

$$\begin{aligned} E_{\{\mathcal{F}_0\}} V_{\{\mathcal{F}_{i-1}\}}(\xi_i) &\geq \frac{(1-a_{i-1}-b_{i-1})(1-a_i-b_i)}{1-b_{i-1}} \sum_{k \in U_i} \alpha_{ik} \left(\check{y}_k - \sum_{l \in U_i} \alpha_{il} \check{y}_l \right)^2 \\ &\geq \left\{ 1 - \max_{k \in U} \pi_k \right\}^2 \left\{ \sum_{k \in U_i} \alpha_{ik} (\check{y}_k - \sum_{l \in U_i} \alpha_{il} \check{y}_l)^2 \right\}. \end{aligned} \quad (52)$$

From (50) and (52), we obtain (17).

D Proof of Theorem 1

From Proposition 4.1 in Ohlsson (1986), it is sufficient to prove that the sequence $\{\eta_i; i = 1, \dots, n\}$ fulfills equations (13) and (14). We first note that from Propositions 2-4 and from Assumptions (H1)-(H3), we obtain

$$E_{\{\mathcal{F}_0\}} \left(\sum_{i=1}^n \xi_i^4 \right) = O\left(\frac{N^4}{n^3}\right), \quad (53)$$

$$V_{\{\mathcal{F}_0\}} \left\{ \sum_{i=1}^n V_{\{\mathcal{F}_{i-1}\}}(\xi_i) \right\} = O\left(\frac{N^4}{n^3}\right), \quad (54)$$

$$V_{\{\mathcal{F}_0\}}(\hat{t}_{y\pi}) \geq C_2 \{1 - f_1\}^2 \frac{N^2}{n}. \quad (55)$$

From equations (53) and (55) we have

$$\sum_{i=1}^n E_{\{y_U\}}(\eta_i^4) = O(n^{-1}), \quad (56)$$

and equation (13) holds with $\delta = 2$. From equations (55) and (57):

$$V_{\{\mathcal{F}_0\}} \left\{ \sum_{i=1}^n V_{\mathcal{F}_{i-1}}(\eta_i) \right\} = O(n^{-1}), \quad (57)$$

and since $E_{\{\mathcal{F}_0\}} \left\{ \sum_{i=1}^n V_{\mathcal{F}_{i-1}}(\eta_i) \right\} = 1$, we obtain (14).

E Proof of Theorem 2

E.1 Preliminary Lemmas

The proofs of Lemmas 6, 7 and 9 are given in the Supplementary Material. The proof of Lemma 8 follows from straightforward computations, and is therefore omitted.

Lemma 6. *We have*

$$V_{\{\mathcal{F}_{i-1}\}}(\xi_i) \leq \left\{ 1 - \frac{a_i b_i}{1 - a_i} \right\} \sum_{k \in U'_i} \alpha_{ik} (\check{y}_k - \bar{y}_{U'_i})^2 \text{ where } \bar{y}_{U'_i} = \sum_{k \in U'_i} \alpha_{ik} \check{y}_k. \quad (58)$$

Lemma 7. *Let m denote some quantity which is non-random, conditionally on L_{i-1} . Then:*

$$V_{\{\mathcal{F}_{i-1}\}}(\xi_i) + E_{\{\mathcal{F}_{i-1}\}} V_{\{\mathcal{F}_i\}}(\xi_{i+1}) \leq \{1 + c_i b_i\} \left\{ \sum_{k \in U'_i} \alpha_{ik} (\check{y}_k - m)^2 + \sum_{k \in U_{i+1}} \alpha_{i+1,k} (\check{y}_k - m)^2 \right\}, \quad (59)$$

where $c_i = \frac{a_i b_i}{(1-a_i)(1-b_i)}$.

Lemma 8. *Conditionally on L_{i-1} , $\{F_i, F_{i+1}\}$ is a pivotal sample selected in $U'_i \cup U_{i+1}$ with inclusion probabilities*

$$\pi'_k = \begin{cases} \alpha_{ik}(1 - c_i c_{i+1}) & \text{for } k \in U'_i \setminus \{k_i\}, \\ a_i + b_i - \frac{(1-a_i-b_i)b_i c_{i+1}}{1-b_i} & \text{for } k = k_i, \\ \alpha_{i+1,k}(1 - c_{i+1}) & \text{for } k \in U_{i+1} \setminus \{k_i, k_{i+1}\}, \\ \frac{a_{i+1}}{1-b_{i+1}} & \text{for } k = k_{i+1}. \end{cases} \quad (60)$$

Lemma 9. *We have*

$$E_{\{\mathcal{F}_{i-1}\}} \{(\check{y}_{F_i} - \check{y}_{F_{i+1}})^2\} \geq \{1 - c_{i+1}\} \left\{ \sum_{k \in U'_i} \alpha_{ik} (\check{y}_k - m_1)^2 + \sum_{k \in U_{i+1}} \alpha_{i+1,k} (\check{y}_k - m_1)^2 \right\}, \quad (61)$$

where

$$m_1 = \frac{1}{2} \sum_{k \in U'_i \cup U_{i+1}} y_k \frac{\pi'_k}{\pi_k}, \quad (62)$$

and where the π'_k 's are given in (60).

E.2 Proof of Theorem 2

We can use equation (50) to write

$$\begin{aligned}
 V_{\{\mathcal{F}_0\}}(\hat{t}_{y\pi}) &= E_{\{\mathcal{F}_0\}} \left\{ \sum_{i=1}^{\lfloor n/2 \rfloor} V_{\{\mathcal{F}_{i-1}\}}(\xi_i) + E_{\{\mathcal{F}_{i-1}\}} V_{\{\mathcal{F}_i\}}(\xi_{i+1}) \right\} \\
 &+ E_{\{\mathcal{F}_0\}} V_{\{\mathcal{F}_{n-1}\}}(\xi_n) 1(n \text{ is odd}).
 \end{aligned} \tag{63}$$

If n is even, equation (23) follows directly from Lemmas 7 and 9. If n is odd, we note that $f(x) = E_{\{\mathcal{F}_{n-1}\}}(\check{y}_{F_n} - x)^2$ is minimized for $x = E_{\{\mathcal{F}_{n-1}\}}(\check{y}_{F_n})$, so that

$$V_{\{\mathcal{F}_{n-1}\}}(\xi_n) \leq E_{\{\mathcal{F}_{n-1}\}}(\check{y}_{F_n} - \check{y}_{F_{n-1}})^2, \tag{64}$$

which completes the proof.

References

- Benedetti, R., Piersimoni, F., Postiglione, P., et al. (2015). *Sampling spatial units for agricultural surveys*. Springer.
- Berger, Y. G. (1998). Rate of convergence to normal distribution for the horvitz-thompson estimator. *Journal of Statistical Planning and Inference*, 67(2):209–226.
- Berger, Y. G. (2011). Asymptotic consistency under large entropy sampling designs with unequal probabilities. *Pakistan Journal of Statistics*, 27(4):407–426.
- Boistard, H., Lopuhaä, H. P., Ruiz-Gazen, A., et al. (2012). Approximation of rejective sampling inclusion probabilities and application to high order correlations. *Electronic Journal of Statistics*, 6:1967–1983.
- Breidt, F. J. and Opsomer, J. D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28(4):1026–1053.
- Chauvet, G. (2012). On a characterization of ordered pivotal sampling. *Bernoulli*, 18(4):1320–1340.
- Chauvet, G. (2017). A comparison of pivotal sampling and unequal probability sampling with replacement. *Statistics & Probability Letters*, 121:1–5.

- Deville, J.-C. and Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika*, 85(1):89–101.
- Deville, J.-C. and Tillé, Y. (2004). Efficient balanced sampling: the cube method. *Biometrika*, 91(4):893–912.
- Dickson, M. M., Benedetti, R., Giuliani, D., and Espa, G. (2014). The use of spatial sampling designs in business surveys. *Open Journal of Statistics*, 4(05):345.
- Dickson, M. M. and Tillé, Y. (2016). Ordered spatial sampling by means of the traveling salesman problem. *Computational Statistics*, 31(4):1359–1372.
- Fattorini, L., Corona, P., Chirici, G., and Pagliarella, M. C. (2015). Design-based strategies for sampling spatial units from regular grids with applications to forest surveys, land use, and land cover estimation. *Environmetrics*, 26(3):216–228.
- Favre-Martinoz, C. and Merly-Alpa, T. (2017). Constitution et tirage d’unités primaires pour des sondages en mobilisant de l’information spatiale. In *Proceedings of the 49th Meeting of the French Statistical Society*.
- Gabler, S. (1981). A comparison of sampford’s sampling procedure versus unequal probability sampling with replacement. *Biometrika*, pages 725–727.
- Gabler, S. (1984). On unequal probability sampling: sufficient conditions for the superiority of sampling without replacement. *Biometrika*, 71(1):171–175.
- Grafström, A., Lundström, N. L., and Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics*, 68(2):514–520.
- Grafström, A. and Ringvall, A. H. (2013). Improving forest field inventories by using remote sensing data in novel sampling designs. *Canadian Journal of Forest Research*, 43(11):1015–1022.
- Grafström, A., Saarela, S., and Ene, L. T. (2014). Efficient sampling strategies for forest inventories by spreading the sample in auxiliary space. *Canadian Journal of Forest Research*, 44(10):1156–1164.
- Grafström, A. and Tillé, Y. (2013). Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. *Environmetrics*, 24(2):120–131.

- Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35:1491–1523.
- Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77(377):89–96.
- Le Gleut, R. (2017). Analyse factorielle et sondages - utilisation de méthodes d'échantillonnage spatial. In *Proceedings of the 49th Meeting of the French Statistical Society*.
- Mark, D. M. (1990). Neighbor-based properties of some orderings of two-dimensional space. *Geographical Analysis*, 22(2):145–157.
- Ohlsson, E. (1986). Asymptotic normality of the Rao-Hartley-Cochran estimator: an application of the martingale CLT. *Scandinavian Journal of Statistics*, 13(1):17–28.
- Qualité, L. (2008). A comparison of conditional poisson sampling versus unequal probability sampling with replacement. *Journal of Statistical Planning and Inference*, 138(5):1428–1432.
- Rao, J. N. K., Hartley, H. O., and Cochran, W. G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society, Series B*, 24:482–491.
- Rosén, B. (1972). Asymptotic theory for successive sampling with varying probabilities without replacement. I, II. *Annals of Mathematical Statistics*, 43:373–397; *ibid.* 43 (1972), 748–776.
- Stevens, D. L. and Olsen, A. R. (2004). Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, 99(465):262–278.
- Tillé, Y. (2006). *Sampling algorithms*. Springer, New York.
- Vallée, A.-A., Ferland-Raymond, B., Rivest, L.-P., and Tillé, Y. (2015). Incorporating spatial and operational constraints in the sampling designs for forest inventories. *Environmetrics*, 26(8):557–570.