



Spatial Spread Sampling Using Weakly Associated Vectors

Raphaël JAUSLIN^{ID} and Yves TILLÉ^{ID}

Geographical data are generally autocorrelated. In this case, it is preferable to select spread units. In this paper, we propose a new method for selecting well-spread samples from a finite spatial population with equal or unequal inclusion probabilities. The proposed method is based on the definition of a spatial structure by using a stratification matrix. Our method exactly satisfies given inclusion probabilities and provides samples that are very well spread. A set of simulations shows that our method outperforms other existing methods such as the generalized random tessellation stratified or the local pivotal method. Analysis of the variance on a real dataset shows that our method is more accurate than these two. Furthermore, a variance estimator is proposed.

Key Words: GRTS; Local pivotal method; Cube method; Stratification.

1. INTRODUCTION

In most natural resources surveys, data contain spatial coordinates. The process of selecting units from a population defined over a region of space is called spatial sampling. These kinds of data are usually autocorrelated, meaning that two close measurements are generally similar. In general, to estimate a total of a target variable, selecting the units spatially best spread collects more information and provides better estimation. An important problem of spatial sampling is thus to spread at best the sampled units in space. A well-spread sample is called spatially balanced. Grafström and Lundström (2013) and Grafström and Schelin (2014) give the formal definition of a representative sample and discuss the theoretical justification of taking a well-spread sample with unequal probabilities. Marker and Stevens (2009) and Hankin et al. (2019) present some example of studies where the population considered is in an environmental context such as lakes, wetlands, rangelands and forests. Vallée et al. (2015) discuss forest ecosystem evolution using a well-spread spatial sampling design. Tillé (2020, Chapter 8), Tillé and Wilhelm (2017), Benedetti et al. (2017) and Wang

R. Jauslin (✉) · Y. Tillé Institute of statistics, University of Neuchâtel, Av. de Bellevaux 51, 2000 Neuchâtel, Switzerland
(E-mail: raphael.jauslin@unine.ch).

© 2020 The Author(s)

Journal of Agricultural, Biological, and Environmental Statistics, Volume 25, Number 3, Pages 431–451
<https://doi.org/10.1007/s13253-020-00407-1>

et al. (2012) give a review of the main spatial sampling methods. Quenouille (1949) and Bellhouse (1977) showed that systematic sampling is the optimal design for autocorrelated data.

Generalized random tessellation stratified (GRTS) sampling is a spatial sampling method proposed by Stevens and Olsen (1999, 2003, 2004). They use a mapping by means of a quadrant-recursive function to map a finite subset of a multi-dimensional space into the real line. A one-dimension systematic sampling is then applied, possibly with unequal probabilities (see also Theobald et al. 2007; Brown et al. 2015; Kincaid et al. 2019). Robertson et al. (2018) have proposed a similar method called Halton iterative partitioning (HIP). It uses structural properties of the Halton sequence to draw a well-spread sample. Dickson and Tillé (2016) have simply used the travelling salesman problem (TSP) in order to map the population points in one dimension. Systematic sampling is then applied. Grafström (2011) has proposed spatially correlated Poisson sampling (SCPS). This method uses weights to create strong negative correlations between the inclusion probabilities of nearby units. Grafström et al. (2012) proposed the local pivotal method (LPM). It is a particular case of the splitting methods proposed by Deville and Tillé (1998). It consists of randomly choosing between two nearby units at each step and produces an automatic repulsion in the selection of the neighbour units. Grafström and Tillé (2013) have generalized the LPM to obtain spread samples that are also balanced on totals of auxiliary variables. All these methods are implemented in the `BalancedSampling` R package (Grafström and Lisic 2019).

Stevens and Olsen (2004) have proposed to compute the Voronoï polygons around the sampled units, after which they sum the inclusion probabilities of the population units belonging to each Voronoï polygon. The variance of these sums, called “spatial balance”, is an indicator of the quality of spreading. Tillé et al. (2018) have modified the index proposed by Moran (1950) so that it can be interpreted as a coefficient of correlation between the units and their neighbourhood. The index provides another measure of the quality of spreading.

Diggle et al. (2010) defined preferential sampling as a sample selection where the sampling method is not independent of the spatial process, and where unequal inclusion probabilities cannot be explained by auxiliary variables. It is important to emphasize here, that in this manuscript the inclusion probabilities are supposed to be established in advance. The sample selection is a random realization of the sampling model and is independent of all of the variables.

In this paper, we propose a new spatial sampling method. We start with the vector of inclusion probabilities. Like in the cube method (Deville and Tillé 2004; Tillé 2006), inclusion probabilities are randomly modified at each step. It can be seen as random walk that from the vector of inclusion probabilities ends up with a sample. By choosing well the modification direction at each step, the sample selected is very well spread.

The paper is organized as follows. Section 2 gives the notation and a basic set-up of the problem as well as the insight that a well-spread sample results in an Horvitz–Thompson estimator with a smaller variance. In Sect. 3, we introduce the new method that we propose and the process of sample selection. In Sect. 4, we describe the indices that enable to evaluate the quality of the spreading: the spatial balance index and the measure based on Moran’s I index. In Sect. 5, we present a variance estimator for our method. In Sect. 6, we give simulation results of the algorithm on artificial spatial configurations, while Sect. 7 is dedicated

to simulations on real data. We used the geo-referenced “Meuse” dataset available in the R package “sp” of Pebesma and Bivand (2005) with inclusion probabilities proportional to the “cadmium” variable. Simulations show that the proposed method surpasses LPM, GRTS and SCPS for the quality of the spreading and the estimation accuracy.

2. NOTATION

2.1. BASIC SET-UP

Consider a finite population U of size N whose units can be defined by labels $k \in \{1, 2, \dots, N\}$. Let $\mathcal{S} = \{s | s \subset U\}$ be the power set of U . These units are geo-referenced in a space that can have more than two dimensions. A sampling design is defined by a probability distribution $p(\cdot)$ on \mathcal{S} such that

$$p(s) \geq 0 \text{ for all } s \in \mathcal{S} \text{ and } \sum_{s \in \mathcal{S}} p(s) = 1.$$

A random sample S is a random vector that maps elements of \mathcal{S} to an N vector of 0 or 1 such that $P(S = s) = p(s)$. Define $a_k(S)$, for $k = 1, \dots, N$:

$$a_k = \begin{cases} 1 & \text{if } k \in S \\ 0 & \text{otherwise.} \end{cases}$$

Then, a sample can be denoted by means of a vector notation: $\mathbf{a}^\top = (a_1, a_2, \dots, a_N)$. For each unit of the population, the inclusion probability $0 \leq \pi_k \leq 1$ is defined as the probability that unit k is selected into sample S :

$$\pi_k = P(k \in S) = E(a_k) = \sum_{s \in \mathcal{S} | k \in s} p(s), \text{ for all } k \in U.$$

Let $\boldsymbol{\pi}^\top = (\pi_1, \dots, \pi_N)$ be the vector of inclusion probabilities. Then, $E(\mathbf{a}) = \boldsymbol{\pi}$. In many applications, inclusion probabilities are such that samples have a fixed size n . Let the set of all samples that have fixed size equal to n be defined by

$$\mathcal{S}_n = \left\{ \mathbf{a} \in \{0, 1\}^N \mid \sum_{k=1}^N a_k = n \right\}.$$

The sample is generally selected with the aim of estimating some population parameters. Let y_k denote a real number associated with unit $k \in U$, usually called the variable of interest. For example, the total

$$Y = \sum_{k \in U} y_k$$

can be estimated by using the classical Horvitz–Thompson estimator of the total defined by

$$\widehat{Y}_{HT} = \sum_{k \in U} \frac{y_k a_k}{\pi_k}. \tag{1}$$

Usually, some auxiliary information $\mathbf{x}_k^\top = (x_{k1}, x_{k2}, \dots, x_{kq}) \in \mathbb{R}^q$ regarding the population units is available. In the particular case of spatial sampling, a set of spatial coordinates $\mathbf{z}_k^\top = (z_{k1}, z_{k2}, \dots, z_{kp}) \in \mathbb{R}^p$ is supposed to be available, where p is the dimension of the considered space. A sampling design is said to be balanced on the auxiliary variables x_k if and only if it satisfies the balancing equations

$$\widehat{\mathbf{X}} = \sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} = \sum_{k \in U} \mathbf{x}_k = \mathbf{X}.$$

2.2. WELL-SPREAD SAMPLE

A sample is well spread “if the number of selected units is close to what is expected on average in any part of the space” (Grafström and Lundström 2013). We give in this section an insight that selecting a well-spread sample minimizes the variance of the Horvitz–Thompson estimator. Let suppose we are in general linear superpopulation model:

$$\mathbf{y}_k = \mathbf{x}_k^\top \beta + \varepsilon_k, \text{ for all } k \in U,$$

where \mathbf{x}_k is a column vector of values taken by q auxiliary variables on unit k , $\beta \in \mathbb{R}^q$ are q regression coefficients and ε_k is a random variable that satisfies $E_M(\varepsilon_k) = 0$ and $\text{var}_M(\varepsilon_k) = \sigma^2(\mathbf{x}_k) = \sigma_k^2$, with $\sigma^2(\cdot)$ a Lipschitz continuous function. Note that $E_M(\cdot)$ and $\text{var}_M(\cdot)$ are the expectation and the variance under the model. Let also

$$\text{cov}_M(\varepsilon_k, \varepsilon_\ell) = \sigma_k \sigma_\ell \rho_{k\ell}, \text{ with } k \neq \ell \in U,$$

where $\rho_{k\ell}$ is a function that decreases when the distance between two units increase. This notation shows that two close units are autocorrelated. Grafström and Tillé (2013) showed that

$$E_p E_M(\widehat{Y}_{HT} - Y)^2 = E_p \left[\left(\sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} - \sum_{k \in U} \mathbf{x}_k \right)^\top \beta \right]^2 + \sum_{k \in U} \sum_{\ell \in U} \sigma_k \sigma_\ell \rho_{k\ell} \frac{\pi_{k\ell} - \pi_k \pi_\ell}{\pi_k \pi_\ell}, \tag{2}$$

where E_p is the expectation of the design and $\pi_{k\ell} = E_p(a_k a_\ell)$ is the joint inclusion probabilities. From Eq. (2), we could see that the first term of the right-hand side is minimized if the sample is balanced on the auxiliary variables \mathbf{X} . The second term is minimized if $\pi_{k\ell}$ is small whenever $\rho_{k\ell}$ is large, meaning that choosing a well-spread sample (i.e. a sample where the $\pi_{k\ell}$ are small) minimizing Eq. (2). Grafström and Lundström (2013) showed that if the inclusion probabilities are set up proportional to the σ_k then (2) is even more minimized. As result, selecting a well-spread sample jointly used with the Horvitz–Thompson estimator is a very efficient procedure in terms of variance reduction.

3. WEAKLY ASSOCIATED VECTOR SAMPLING

3.1. GENERAL IDEA

Our sampling algorithm, Weakly Associated Vector (WAVE) sampling starts with the inclusion probability vector. At each step, this vector is randomly modified so that at least one of the components of the vector is replaced by a 0 or a 1. So, in at most N steps a sample is randomly selected. This idea is also used in the cube method proposed by [Deville and Tillé \(2004\)](#) to select balanced samples. The proposed method is different from the cube method by selecting in a completely different way the vector of modification of inclusion probabilities. By carefully choosing the direction of the modification of the working vector, we can ensure that the selection of the sample will be well spread. This choice is described in Sect. 3.4.

3.2. DISTANCE

In order to describe the spatial structure of the population, a distance is defined as a function m defined on the product set $U \times U$ such that

$$m : U \times U \rightarrow \mathbb{R}^+, \tag{3}$$

and satisfies the property of non-negativity, symmetry and triangular inequality. More specifically, for all $x, y, z \in U$ the following properties hold:

$$\begin{aligned} m(x, y) &\geq 0, \quad m(x, y) = 0 \iff x = y, \\ m(x, y) &= m(y, x), \\ m(x, z) &\leq m(x, y) + m(y, z). \end{aligned}$$

In most of applications, the usual Euclidean squared distance is used. It is defined by,

$$m_E^2(k, \ell) = (\mathbf{z}_k - \mathbf{z}_\ell)^\top (\mathbf{z}_k - \mathbf{z}_\ell), \tag{4}$$

where \mathbf{z}_k and \mathbf{z}_ℓ are the spatial coordinates of units $k, \ell \in U$. Sometimes it could be interesting to compute the distance on auxiliary variables. In this case, the Mahalanobis distance can be more appropriate,

$$m_M^2(k, \ell) = (\mathbf{x}_k - \mathbf{x}_\ell)^\top \mathbf{S}^{-1} (\mathbf{x}_k - \mathbf{x}_\ell),$$

where

$$\mathbf{S} = \frac{1}{N} \sum_{k \in U} (\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})^\top, \quad \bar{\mathbf{x}} = \frac{1}{N} \sum_{k \in U} \mathbf{x}_k.$$

When the population is distributed on a $N_1 \times N_2$ regular grid of \mathbb{R}^2 , a tore distance can be defined. We define a tore distance as the Euclidean metric calculated on a regular tore. An

advantage of using this distance is that the surface on which we working on has not anymore corners and edges. With this tore distance, two units on the same column (respectively, row) that are on the opposite side have a small distance. More precisely, a unit that is positioned at the right top corner of the grid will be equally distant to the left top corner and the right bottom corner. It is like seeing the grid curved such that it looks like a regular tore.

The distance is then defined by:

$$m_T^2(k, \ell) = \min[(z_{k1} - z_{\ell1})^2, (z_{k1} + N_1 - z_{\ell1})^2, (z_{k1} - N_1 - z_{\ell1})^2] + \min[(z_{k2} - z_{\ell2})^2, (z_{k2} + N_2 - z_{\ell2})^2, (z_{k2} - N_2 - z_{\ell2})^2]. \tag{5}$$

Example 3.1. Let $\{1, \dots, 9\}$ be on a regular grid of size 3×3 , then the squared distance matrices defined by Eqs. (4) and (5) are equal to

$$\mathbf{M}_E = \begin{pmatrix} 0 & 1 & 4 & 1 & 2 & 5 & 4 & 5 & 8 \\ 1 & 0 & 1 & 2 & 1 & 2 & 5 & 4 & 5 \\ 4 & 1 & 0 & 5 & 2 & 1 & 8 & 5 & 4 \\ 1 & 2 & 5 & 0 & 1 & 4 & 1 & 2 & 5 \\ 2 & 1 & 2 & 1 & 0 & 1 & 2 & 1 & 2 \\ 5 & 2 & 1 & 4 & 1 & 0 & 5 & 2 & 1 \\ 4 & 5 & 8 & 1 & 2 & 5 & 0 & 1 & 4 \\ 5 & 4 & 5 & 2 & 1 & 2 & 1 & 0 & 1 \\ 8 & 5 & 4 & 5 & 2 & 1 & 4 & 1 & 0 \end{pmatrix}, \mathbf{M}_T = \begin{pmatrix} 0 & 1 & 1 & 1 & 2 & 2 & 1 & 2 & 2 \\ 1 & 0 & 1 & 2 & 1 & 2 & 2 & 1 & 2 \\ 1 & 1 & 0 & 2 & 2 & 1 & 2 & 2 & 1 \\ 1 & 2 & 2 & 0 & 1 & 1 & 1 & 2 & 2 \\ 2 & 1 & 2 & 1 & 0 & 1 & 2 & 1 & 2 \\ 2 & 2 & 1 & 1 & 1 & 0 & 2 & 2 & 1 \\ 1 & 2 & 2 & 1 & 2 & 2 & 0 & 1 & 1 \\ 2 & 1 & 2 & 2 & 1 & 2 & 1 & 0 & 1 \\ 2 & 2 & 1 & 2 & 2 & 1 & 1 & 1 & 0 \end{pmatrix}. \tag{6}$$

In spatial configuration of a regular grid, some distances between points are equal. The rank of the nearest neighbours is then assigned and duplicated values appear. In order to obtain a different rank distance for each unit, a small random quantity is added to the coordinates so that it disturbs the given units and the distances are a little bit different from each other. Let $\boldsymbol{\varepsilon} \in \mathbb{R}^2$ and $\tilde{\mathbf{z}}_k = \mathbf{z}_k + \boldsymbol{\varepsilon}$ the shifted coordinates, Equation (5) is then replaced by,

$$m_S^2(k, \ell) = \min[(\tilde{z}_{k1} - z_{\ell1})^2, (\tilde{z}_{k1} + N_1 - z_{\ell1})^2, (\tilde{z}_{k1} - N_1 - z_{\ell1})^2] + \min[(\tilde{z}_{k2} - z_{\ell2})^2, (\tilde{z}_{k2} + N_2 - z_{\ell2})^2, (\tilde{z}_{k2} - N_2 - z_{\ell2})^2].$$

$\boldsymbol{\varepsilon}$ is called a “shift” and m_S the shifted version of m_T , for example if $\boldsymbol{\varepsilon} = (1/12, 1/4)$, the distance matrix \mathbf{M}_S becomes,

$$\mathbf{M}_S = \begin{pmatrix} 0 & 0.90 & 1.24 & 0.57 & 1.40 & 1.74 & 1.57 & 2.40 & 2.74 \\ 1.24 & 0 & 0.90 & 1.74 & 0.57 & 1.40 & 2.74 & 1.57 & 2.40 \\ 0.90 & 1.24 & 0 & 1.40 & 1.74 & 0.57 & 2.40 & 2.74 & 1.57 \\ 1.57 & 2.40 & 2.74 & 0 & 0.90 & 1.24 & 0.57 & 1.40 & 1.74 \\ 2.74 & 1.57 & 2.40 & 1.24 & 0 & 0.90 & 1.74 & 0.57 & 1.40 \\ 2.40 & 2.74 & 1.57 & 0.90 & 1.24 & 0 & 1.40 & 1.74 & 0.57 \\ 0.57 & 1.40 & 1.74 & 1.57 & 2.40 & 2.74 & 0 & 0.90 & 1.24 \\ 1.74 & 0.57 & 1.40 & 2.74 & 1.57 & 2.40 & 1.24 & 0 & 0.90 \\ 1.40 & 1.74 & 0.57 & 2.40 & 2.74 & 1.57 & 0.90 & 1.24 & 0 \end{pmatrix}. \tag{7}$$

The matrix is no longer a distance matrix since the symmetric axiom has been dropped. A distance that has an unsatisfied symmetry axiom is called a quasi-metric. Nevertheless,

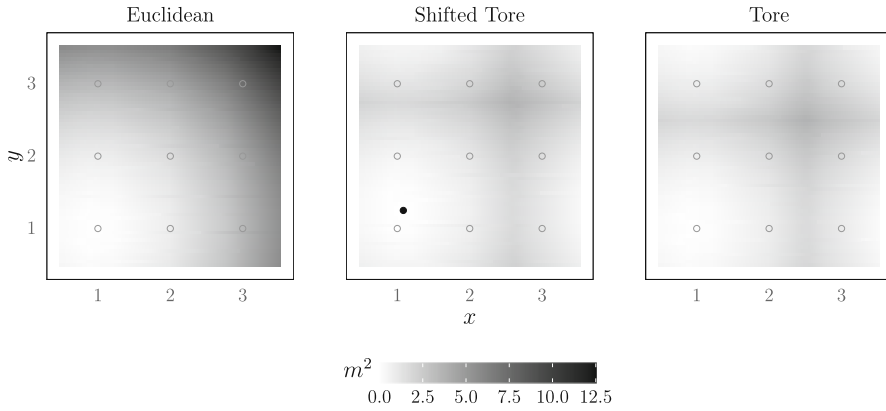


Figure 1. Simple example of a 3 × 3 regular grid set up on three different distances with a gradient calculated from the points (1, 1). The left one is the classical Euclidean distance (4), the right one is the tore distance given in (5) and the central graph is the shifted tore distance with a shift equal to (1/12, 1/4) (the black point on the graph). It illustrates the two different patterns and the values of the grid points corresponding to the entries of the first row of the three previous matrices (6).

if an epsilon value is added instead of (1/12, 1/4), then the values are almost the same and the order is preserved in each row. In Fig. 1, three simple configurations are presented: Euclidean, tore and shifted tore distance on a 3 × 3 regular grid. In shifted distance graph, all the distances from point (1, 1) to the other grid points are different.

3.3. THE STRATIFICATION MATRIX

Let $k \in U$ denote a unit in the population. The idea is to construct a strata G_k under some distance metric such that the elements in G_k are ranked in increasing order. Define G_k the set of the nearest neighbours of unit k , including k , such that their inclusion probabilities are greater or equal than one by only one unit. Denote g_k the number of elements inside G_k , the spatial weights are then defined as follows

$$w_{k\ell} = \begin{cases} \pi_\ell & \text{if unit } \ell \text{ is in the set of the } g_k - 1 \text{ nearest neighbour of } k, \\ \pi_\ell + 1 - \sum_{j \in G_k} \pi_k & \text{if unit } \ell \text{ is the } g_k \text{th nearest neighbour of } k, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

Let \mathbf{W} denote an $N \times N$ stratification matrix and each row of matrix \mathbf{W} represents a stratum. Each stratum is defined by a particular unit and its neighbouring units. Nearest neighbours are defined with a metric function (3). If the metric is such that there exists ties values, then we can divide the quantity w_{kl} into the different g_k nearest neighbours of the unit k that have the same distance. Or, a shifted metrics can be used [exemplified in matrix (7)] such that all the distances are different. Each row of matrix \mathbf{W} sum to 1. Thus, matrix \mathbf{W} is a right stochastic matrix. Most of the components of matrix \mathbf{W} are null. Matrix \mathbf{W} can thus be encoded as a sparse matrix.

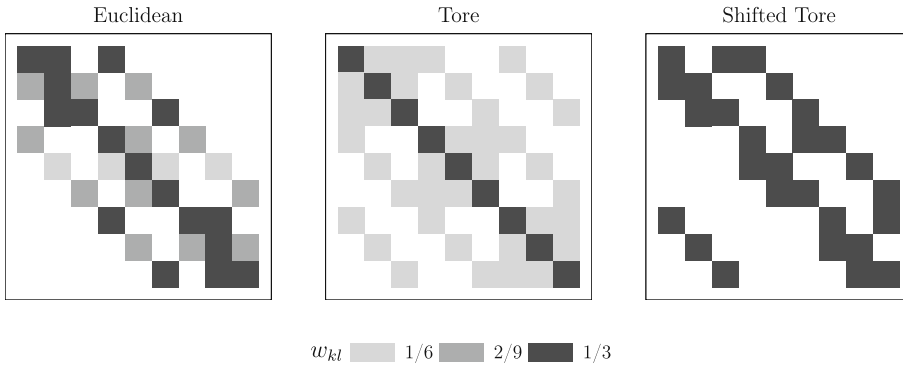


Figure 2. Sparsity pattern of three stratification matrices. Spatial coordinates are 3×3 regular grid and the inclusion probabilities are equal to $\pi = (1/3, \dots, 1/3)$. Depending on the way of defining the nearest neighbours in Eq. (8), different weight values are obtained. The left stratification matrix uses the classical Euclidean distance (4), the central one the tore distance (5) and the right one uses a shifted tore distance with a shift randomly generated from a random variable $\mathcal{N}(0, 1/100\mathbf{I})$.

Example 3.2. Let $U = \{1, 2, 3, 4, 5\}$ a population of 5 units. Suppose that the inclusion probabilities are equal to $\pi = (1/2, 1/3, 1/4, 1/5, 1/6)$ and that the order in terms of distance metric from the unit 1 is exactly equal to 1, 2, 3, 4, 5, meaning that the 5th unit is the farthest to the first. Then, $G_k = \{1, 2, 3\}$ because $1/2 + 1/3 + 1/4 \cong 1.084 > 1$ and $w_{13} = 1/4 + 1 - (1/2 + 1/3 + 1/4) = 1/6$.

Example 3.3. Let $\{1, \dots, 9\}$ be on a regular grid of size 3×3 with inclusion probabilities equal $\pi_k = 1/3$, for all $k \in U$. Figure 2 shows different stratification matrices corresponding to $\mathbf{M}_E, \mathbf{M}_T$ and \mathbf{M}_S with a shift randomly generated from a random variable $\mathcal{N}(0, 1/100\mathbf{I})$ where \mathbf{I} is the identity matrix.

Let now $\mathbf{D} = \text{diag}(\pi)$ the matrix with inclusion probabilities on the diagonal and define \mathbf{A} by

$$\mathbf{A} = \mathbf{W}\mathbf{D}^{-1} = \begin{pmatrix} w_{11}/\pi_1 & w_{12}/\pi_2 & \cdots & w_{1N}/\pi_N \\ \vdots & \vdots & \ddots & \vdots \\ w_{N1}/\pi_1 & w_{N2}/\pi_2 & \cdots & w_{NN}/\pi_N \end{pmatrix}. \tag{9}$$

Matrices \mathbf{W} and \mathbf{A} are square but not necessarily full rank. The sum of the rows of \mathbf{A} is equal or approximately equal to the number of elements in each stratum. The strata are represented by the rows and the contribution of a unit i in each stratum is represented by the i th column. Figure 3 shows the sparsity pattern of the two stratification matrices.

Example 3.4. Let U be a population of size $N = 250$ and inclusion probabilities equal to $\pi_k = 1/25$, for all $k \in U$. Suppose that spatial coordinates are generated independently from a uniform distribution on the square unit, so that with probability one there are no tied distance values, since all $1/\pi_k = 25$ the nonzero entries of \mathbf{A} are all equal to 1. Based on the definition (8), the weights are all equal to the inclusion probabilities or zero. Figure 3 shows the sparsity pattern of the stratification matrices and exemplifies some initial strata.

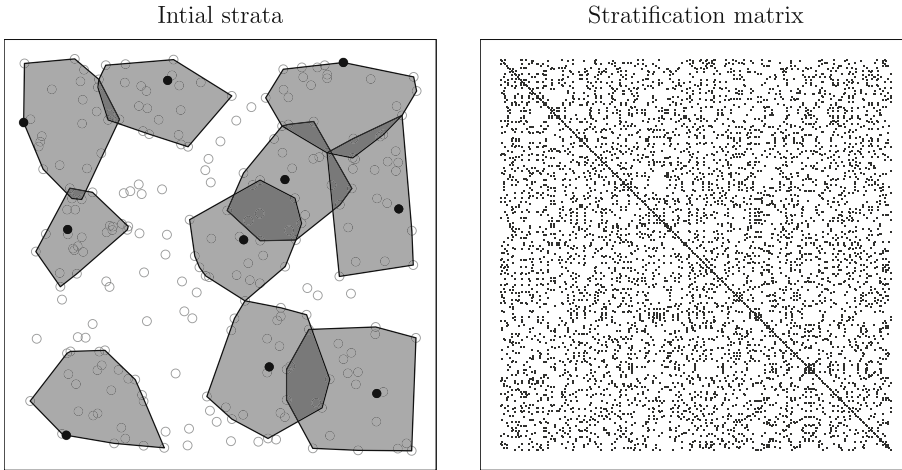


Figure 3. Representation of the strata defined by the spatial weights [Eq. (8)]. Spatial coordinates of the units are generated randomly from a uniform distribution on the square unit $[0, 1] \times [0, 1]$. The overall population size is equal to $N = 250$ and the inclusion probabilities are identical and equal to $\pi_k = 1/25 = 0.04$, meaning that the sample size is equal to $n = 10$. With these parameters, the expected number of units in each stratum is equal to $125/4 = 25$. The left graph shows the population and the selected units with its initial strata. On the right, it shows the sparsity pattern of the matrix (9). All entries of the matrix are equal to 1 .

3.4. IMPLEMENTATION

The method is described in detail in Algorithm 1. The main idea is derived from the cube method (Deville and Tillé 2004). At each step, vector π is randomly modified. To modify π , we choose a vector that spreads at best. Ideally, the aim consists of obtaining a sample \mathbf{a} such that the following equality is satisfied:

$$\mathbf{A}\mathbf{a} = \mathbf{A}\pi = \mathbf{1}.$$

This linear system defines an affine subspace of \mathbb{R}^N :

$$\mathcal{A} = \{\mathbf{a} \in \mathbb{R}^N \mid \mathbf{A}\mathbf{a} = \mathbf{A}\pi\}$$

which could also be rewritten:

$$\mathcal{A} = \pi + \text{Null}(\mathbf{A})$$

where

$$\text{Null}(\mathbf{A}) = \{\mathbf{v} \in \mathbb{R}^N \mid \mathbf{A}\mathbf{v} = \mathbf{0}\}.$$

Depending if matrix \mathbf{A} is full rank or not, the vector giving the direction is not selected in the same way. If matrix \mathbf{A} is not full rank, a vector that is contained in the right null space is selected. If matrix \mathbf{A} is full rank, we compute \mathbf{v}, \mathbf{u} a left and a right singular vectors associated with the smallest singular value σ of \mathbf{A} , i.e.

$$\mathbf{A}\mathbf{v} = \sigma\mathbf{u}, \quad \mathbf{A}^\top\mathbf{u} = \sigma\mathbf{v}.$$

By choosing the modification vector \mathbf{v} , we ensure that we select the vector which remains closest to the set \mathcal{A} . Vector \mathbf{v} is called the weakest associated vector to the matrix \mathbf{A} . Vector \mathbf{v} is then centred to ensure the fixed sample size. By using these weakest associated vectors, the initial spatial configurations are the least modified. At each step, some inclusion probabilities $\boldsymbol{\pi}$ are modified and at least one component is set to 0 or 1. Matrix \mathbf{A} is updated from the new inclusion probabilities. This step is repeated until there is only one component that is not equal to 0 or 1.

Algorithm 1 Algorithm for WAVE sampling

Let $\mathbf{A} = \mathbf{A}_0$ and $\boldsymbol{\pi}_0 = (\pi_1^{(0)}, \dots, \pi_N^{(0)}) = \boldsymbol{\pi}$ for the initialization step. For $t = 0, 1, 2, \dots$

1. From $\boldsymbol{\pi}_t$, extract $\tilde{\boldsymbol{\pi}}_t$ vector $\boldsymbol{\pi}_t$ restricted to the k such that $0 < \pi_k^{(t)} < 1$. Let J be the length of $\tilde{\boldsymbol{\pi}}_t$.
2. Compute the $J \times J$ matrix \mathbf{A}_t of Equation (9) using inclusion probabilities $\tilde{\boldsymbol{\pi}}_t$.
3. Calculate the rank r of matrix \mathbf{A}_t .
 - (a) If matrix \mathbf{A}_t does not have full rank, choose $\mathbf{v}_t = (v_1^{(t)}, \dots, v_J^{(t)}) \in \mathbb{R}^J$ a vector in the right null space of \mathbf{A}_t .
 - (b) If matrix \mathbf{A}_t has full rank, compute the singular value decomposition and seek for \mathbf{v}_t a right singular vector associated with the smallest singular value σ_t .
4. Next in order to ensure the fixed sample size, vector \mathbf{v}_t is centred:

$$\tilde{\mathbf{v}}_t = \mathbf{v}_t - \frac{1}{J} \sum_{i \in J} v_i^{(t)} \mathbf{1}_J,$$

where $\mathbf{1}_J$ is the $J \times 1$ vector of one.

5. Find λ_1 and λ_2 the largest positive real numbers such that all the $0 \leq \tilde{\pi}_k^{(t)} + \lambda_1 \tilde{v}_k^{(t)} \leq 1$ and $0 \leq \tilde{\pi}_k^{(t)} - \lambda_2 \tilde{v}_k^{(t)} \leq 1, k = 1, \dots, J$.
6. Compute

$$\boldsymbol{\pi}_{t+1} = \begin{cases} \tilde{\boldsymbol{\pi}}_t + \lambda_1 \tilde{\mathbf{v}}_t & \text{with probability } \lambda_2 / (\lambda_1 + \lambda_2) \\ \tilde{\boldsymbol{\pi}}_t - \lambda_2 \tilde{\mathbf{v}}_t & \text{with probability } \lambda_1 / (\lambda_1 + \lambda_2). \end{cases}$$

7. Return at 1. with $\boldsymbol{\pi}_{t+1}$ until no units k remains such that $0 < \pi_k^{(t+1)} < 1$.
-

Algorithm 1 is implemented in a R package, which uses the Armadillo C++ library into the R interface (Eddelbuettel and Sanderson 2014). The implementation uses the sparse matrix class. Indeed, depending on the inclusion probabilities, matrix \mathbf{A} given in (9) could be strongly sparse. Even if the function benefits from the C++ implementation, it could be quite time consuming as the size of the population N increases. Nevertheless, we will see in the next section that the algorithm performs better in terms of two spreading measures than those currently used for the spatial balanced sampling design.

4. SPATIAL BALANCE

4.1. VORONOÏ POLYGONS

Stevens and Olsen (2004) suggested the spatial balance of a sample consists of using the Voronoï polygons. The Voronoï polygon associated with the sample unit k is the set of all units of the population that are closer to k than to any other sample units. Let v_k be the sum of inclusion probabilities of the units belonging to the Voronoï polygon associated with the sample unit k . If the sample is perfectly spread, v_k should be equal to 1 for each k . Indeed, n units are selected in the sample, then

$$\sum_{k \in S} v_k = \sum_{k \in U} \pi_k = n,$$

and so

$$\frac{1}{n} \sum_{k \in S} v_k = 1.$$

The variance of the $E[v_k]$ could be approximated and give a good measure of the spatial balance of the sample. The spatial balance measure based on the Voronoï polygons is defined by

$$B(S) = \frac{1}{n} \sum_{k \in S} (v_k - 1)^2. \tag{10}$$

Two samples are compared in Fig. 4. The left one is selected with a simple random sampling without replacement and the right one is selected with WAVE sampling. The darker the Voronoï polygon, the more units it contains. An exactly well-spread sample should have all polygons of the same colour.

The measure B has some limitations. It does not vary from a fixed finite range. This does not allow a clear understanding if the sample is balanced or clustered (Tillé et al. 2018). Moreover, the measure behaves sometimes wrongly and suggest a well-spread sample although it is not the case. Examples are given in Supplementary Material Section. For these reasons, we suggest to use another measure based on Moran’s I index.

4.2. MORAN’S I INDEX

A second approach for measuring the spatial balance of a sampling design has been proposed by Tillé et al. (2018). Consider a $N \times N$ spatial weights matrix,

$$\mathbf{W} = \begin{pmatrix} 0 & w_{12} & \cdots & w_{1N} \\ w_{21} & 0 & \cdots & w_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ w_{N1} & w_{N2} & \cdots & 0 \end{pmatrix}.$$

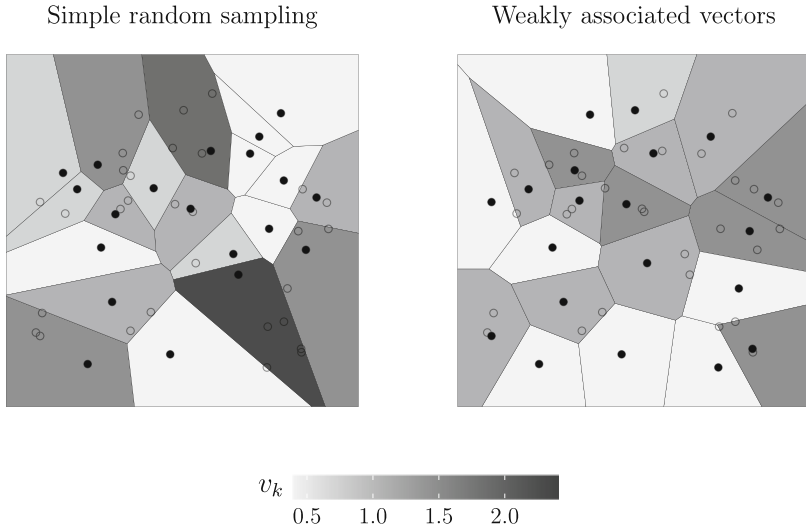


Figure 4. Illustrated example of how the spatial balance measure based on the Voronoi is performed. The population and sample sizes are, respectively, equal to $N = 50$ and $n = 20$, the inclusion probabilities are identical and equal to $\pi_k = 0.4$. The spatial coordinates are generated from two random uniform $\mathcal{U}(0, 1)$. Two sampling design are compared. The left one is the simple random sampling without replacement and the right one is the weakly associated vector sampling .

A large value of $w_{k\ell}$ indicates that ℓ is a neighbour of k . Matrix \mathbf{W} is not necessarily symmetric. The index proposed by Tillé et al. (2018) is defined by

$$I_B(\mathbf{a}) = \frac{(\mathbf{a} - \bar{\mathbf{a}}_w)^\top \mathbf{W}(\mathbf{a} - \bar{\mathbf{a}}_w)}{\sqrt{(\mathbf{a} - \bar{\mathbf{a}}_w)^\top \mathbf{D}(\mathbf{a} - \bar{\mathbf{a}}_w)(\mathbf{a} - \bar{\mathbf{a}}_w)^\top \mathbf{B}(\mathbf{a} - \bar{\mathbf{a}}_w)}}, \tag{11}$$

where \mathbf{a} is the sample and

$$\bar{\mathbf{a}}_w = \frac{\mathbf{a}^\top \mathbf{W} \mathbf{1}}{\mathbf{1}^\top \mathbf{W} \mathbf{1}},$$

\mathbf{D} is the diagonal matrix containing $w_{k.} = \sum_{\ell \in U} w_{k\ell}$ on its diagonal,

$$\mathbf{B} = \mathbf{C}^\top \mathbf{D} \mathbf{C}, \quad \mathbf{C} = \mathbf{D}^{-1} \mathbf{W} - \frac{\mathbf{1} \mathbf{1}^\top \mathbf{W}}{\mathbf{1}^\top \mathbf{W} \mathbf{1}},$$

and $\mathbf{1}$ is a column vector of N ones. Tillé et al. (2018) pointed out that I_B can be interpreted as weighted correlation between a_k and the average of the a_ℓ that are in the neighbourhood of k . We have that $-1 \leq I_B \leq 1$ and $I_B = -1$ when the sample is well spread. Tillé et al. (2018) have proposed to use the inverse of the inclusion probability $h_k = 1/\pi_k$ to define the neighbours of the unit k . More specifically, if the unit k is selected it seems natural to consider $h_k - 1$ neighbours in the population. Let $\lfloor h_k \rfloor$ and $\lceil h_k \rceil$ be, respectively, the inferior and superior integers of h_k . Spatial weights are then defined as follows,

$$w_{k\ell} = \begin{cases} 1 & \text{if unit } \ell \text{ is in the set of the } \lfloor h_k \rfloor \text{ nearest neighbour of } k \\ h_k - \lfloor h_k \rfloor & \text{if unit } \ell \text{ is the } \lceil h_k \rceil \text{th nearest neighbour of } k \\ 0 & \text{otherwise .} \end{cases} \tag{12}$$

For example, if a unit k has an inclusion probability of $\pi_k = 0.35$ then $h_k \cong 2.857$, meaning that the first nearest neighbour of k has a weight equal to 1 and the second has a weight of 0.857. In case there are units that are at equal distance from each other, Tillé et al. (2018) suggest to divide the spatial weights equally among them.

We propose a new way of defining the spatial weights. It consists of using spatial weights defined in (8) rather than the weights (12). We set $w_{kk} = 0$ for all $k \in U$. For the rest of the paper, I_{B1} will represent the measure based on the spatial weights (12) and I_B the one based on (8).

5. VARIANCE ESTIMATION

If the sampling design is of fixed size, the variance of the Horvitz–Thompson estimator of the total (1) is defined by

$$\text{var}(\widehat{Y}_{HT}) = -\frac{1}{2} \sum_{k \in U} \sum_{\ell \in U} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell} \right)^2 \Delta_{k\ell},$$

where $\Delta_{k\ell} = \pi_{k\ell} - \pi_k \pi_\ell$ and $\pi_{k\ell} = E(a_k a_\ell)$ is the joint inclusion probabilities. For complex sampling designs, quantities $\pi_{k\ell}$ are generally impossible to compute.

Many different estimators have been developed. Sen (1953) and Yates and Grundy (1953) proposed one classical estimator:

$$v_{SYG}(\widehat{Y}_{HT}) = -\frac{1}{2} \sum_{k \in S} \sum_{\ell \in S} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell} \right)^2 \frac{\Delta_{k\ell}}{\pi_{k\ell}}.$$

This estimator can take negative values, but it is non-negative when $\Delta_{k\ell} \leq 0$ for all $k \neq \ell \in U$. A common problem with spatially balanced sampling designs is that many joint inclusion probabilities are equal to zero. Indeed, the probability of selecting two close units is generally zero or very close to zero. In this case, v_{SYG} is not an unbiased estimator of $\text{var}(\widehat{Y}_{HT})$.

Tillé (2020, Chapter 5) gives a general estimator based on the variance estimator of the conditional Poisson sampling. It is equal to

$$v(\widehat{Y}_{HT}) = \sum_{k \in S} \frac{c_k}{\pi_k^2} (y_k - \hat{y}_k^*)^2,$$

where

$$\hat{y}_k^* = \pi_k \frac{\sum_{\ell \in S} c_\ell y_\ell / \pi_\ell}{\sum_{\ell \in S} c_\ell}.$$

Choosing $c_\ell = (1 - \pi_k)n/n - 1$, we obtain the Hájek–Rosén estimator (Hájek 1981) defined by

$$v_{\text{HAJ}}(\widehat{Y}_{\text{HT}}) = \frac{n}{n-1} \sum_{k \in S} (1 - \pi_k) \left\{ \frac{y_k}{\pi_k} - \frac{\sum_{\ell \in S} y_\ell (1 - \pi_\ell) / \pi_\ell}{\sum_{\ell \in S} (1 - \pi_\ell)} \right\}^2. \quad (13)$$

This variance estimator is simple to compute and has the advantage of using only the first-order inclusion probabilities. It is a good estimator for maximum entropy sampling design and simple random sampling without replacement. Grafström et al. (2012) pointed out that the estimator seems to overestimate the variance for spread sampling design. Grafström and Schelin (2014) proposed an estimator based on the nearest neighbour in the sample. It is called variance estimator for spatially balanced sampled and is defined as follows:

$$v_{\text{SB}}(\widehat{Y}_{\text{HT}}) = \frac{1}{2} \sum_{k \in S} \left(\frac{y_k}{\pi_k} - \frac{y_{\ell_k}}{\pi_{\ell_k}} \right)^2, \quad (14)$$

where ℓ_k is the nearest neighbour to the unit k in the sample. Stevens and Olsen (2003) proposed an estimator based on a local neighbourhood for each unit in the sample. It is called the local mean variance estimator and is given by

$$v_{\text{LM}}(\widehat{Y}_{\text{HT}}) = \sum_{k \in U} \sum_{\ell \in D_k} w_{k\ell} \left(\frac{y_k}{\pi_k} - \sum_{m \in D_k} w_{km} \frac{y_m}{\pi_m} \right)^2, \quad (15)$$

where the weights $w_{k\ell}$ are computed such that they vary inversely as π_ℓ and decrease as the distance between unit k and ℓ increases. Moreover, it satisfies the constraint $\sum_{k \in S} w_{k\ell} = \sum_{\ell \in S} w_{k\ell} = 1$. The set D_k is the neighbourhood of the unit k and is defined by the unit itself and the three neighbourhoods of the three nearest neighbours, meaning that D_k contains at least four units and at most thirteen. This variance estimator is implemented by function `localmean.var` in the R package “spsurvey” Kincaid et al. (2019). It produces a good estimator for the GRTS method. For the rest of the manuscript, we will adopt the following notation: $v_{\text{LM}_j}(\widehat{Y}_{\text{HT}})$ where j is the number of neighbours used in the calculations. In Sect. 7, we compare the previous estimators for different sampling designs.

6. SIMULATIONS ON ARTIFICIAL SPATIAL CONFIGURATIONS

In this section, we propose three artificial spatial configurations to study the performance of the WAVE sampling in terms of spreading measure. To generate the three population datasets, the expected size of the population is equal to $N = 144$.

1. The dataset is generated from the complete spatial randomness (CSR) that is a Poisson process with intensity equal to N , meaning that the expected number of points in the unit square is equal to N .
2. A Neyman–Scott cluster process (Neyman and Scott 1958) is generated with 12 circular discs of radius 0.055 with units uniformly distributed around the centre. Each cluster contains 12 units such that the population target size is equal to N .

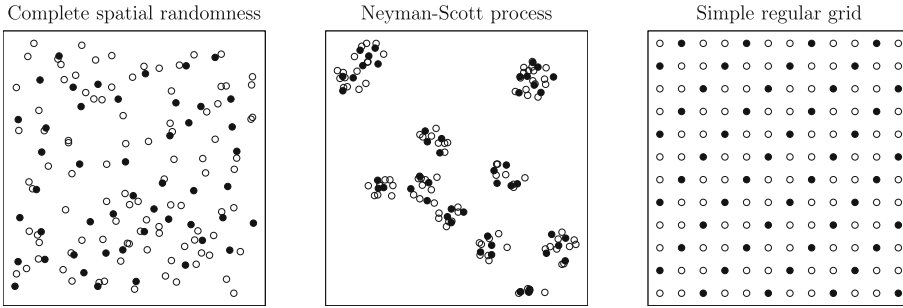


Figure 5. Example of a sample selection by the WAVE sampling on the three different spatial configurations, complete spatial randomness, Neyman–Scott and regular grid. For each of them, the inclusion probabilities are equal to $\pi_k = n/N$ for all $k \in U$.

3. Simple regular grid of size 12×12 .

Figure 5 shows a sample selection by the WAVE sampling design on the three different datasets. For the three configurations, the sample size is equal to $n = 3$ and the inclusion probabilities are all equal to $\pi_k = n/N$ for all $k \in U$. When units are regularly dispersed in the space and when the inverse of inclusion probabilities is equal to an integer that is a divisor of the population size N , the selected sample can be systematic, which is the optimal solution.

For each population, 10,000 samples of size n , respectively, equal to 25, 50 and 100 are selected. Two cases are considered for the inclusion probabilities. In the first case, all inclusion probabilities are equal

$$\text{for all } k \in U, \pi_k = \pi = \frac{n}{N}.$$

For the second case, the inclusion probabilities are unequal and sum up to n ,

$$\text{for all } k, \ell \in U, k \neq \ell \text{ we have } \pi_k \neq \pi_\ell \text{ and } \sum_{k \in U} \pi_k = n.$$

In each case, we calculate the spatial balance based on the Voronoi polygons (10) and measures based on Moran’s I index (11). The simulation results of the CRS dataset are given in Table 1. For the measures based on Moran’s I index, the WAVE sampling design performs better than the other algorithms. Moreover, for the classical measure based on the Voronoi polygons, the WAVE sampling design performs equally and sometimes better than the local pivotal method. This can be explained by the fact that the spatial balance measure based on the Voronoi polygons is less sensitive to observe a well-spread sample and sometimes suggest a well-spread sample although it is not the case (see Supplementary Material Section). For the equal probabilities designs, the measures I_{B_1} and I_B coincide. Indeed, the strata based on the inverse inclusion probabilities are the same as the ones considered such that the inclusion probabilities sum to 1. For unequal sampling designs, the differences are less marked with the measure based on the inverse inclusion probabilities (12). This result comes from the heterogeneity of the strata and the randomness of the algorithm. If the inclusion probabilities

Table 1. Spreading measures results based on 10,000 simulations on the complete spatial randomness dataset

	Sampling design										
	Equal probabilities						Unequal probabilities				
	Wave	lpm1	scps	grts	hip	srswor	Wave	lpm1	scps	grts	maxent
I_{B_1}											
$n = 16$	-0.530	-0.348	-0.370	-0.220	-0.259	-0.030	-0.028	-0.009	-0.012	0.027	0.093
$n = 32$	-0.693	-0.467	-0.464	-0.322	-0.392	-0.017	-0.125	-0.095	-0.085	-0.059	0.016
$n = 48$	-0.807	-0.583	-0.506	-0.375	-0.373	-0.015	-0.436	-0.344	-0.318	-0.229	-0.020
I_B											
$n = 16$	-0.530	-0.348	-0.370	-0.220	-0.259	-0.030	-0.459	-0.316	-0.331	-0.201	-0.028
$n = 32$	-0.693	-0.467	-0.464	-0.322	-0.392	-0.017	-0.548	-0.393	-0.373	-0.261	-0.013
$n = 48$	-0.807	-0.583	-0.506	-0.375	-0.373	-0.015	-0.621	-0.469	-0.424	-0.292	-0.029
B											
$n = 16$	0.115	0.117	0.108	0.164	0.135	0.338	0.123	0.124	0.118	0.177	0.345
$n = 32$	0.137	0.128	0.130	0.167	0.165	0.345	0.140	0.146	0.138	0.180	0.352
$n = 48$	0.158	0.137	0.149	0.177	0.195	0.337	0.165	0.151	0.158	0.189	0.319

The population size is equal to 144

of a unit are nearly zero, then the size of the strata will be very large. This effect can increase the spatial balance measure. Similar results for the two remaining datasets can be seen in the Supplementary Material Section. This analysis shows that the measure I_B should be preferred to I_{B_1} .

7. APPLICATION TO THE MEUSE DATASET

This section investigates the application of WAVE sampling on the dataset “Meuse” available in the R package “sp” of [Pebesma and Bivand \(2005\)](#). It is described as follows: “This data set gives locations and topsoil heavy metal concentrations, along with a number of soil and landscape variables at the observation locations, collected in a flood plain of the river Meuse, near the village of Stein (NL). Heavy metal concentrations are from composite samples of an area of approximately 15 m × 15 m”.

In order to see how the WAVE sampling performs in terms of spread measures, $m = 10,000$ samples of size, respectively, equal to 15, 30 and 50 are selected. As in the previous simulation with an artificial population, two cases are considered, equal and unequal probabilities. In the latter case, inclusion probabilities are set proportional to concentration of copper. Locations with high concentrations of copper were therefore more likely to be selected into the sample. Let Y be the total cadmium concentration over the whole population. To show that the variance of the estimated total with the WAVE sampling design is lower than the other method, we calculate the approximated variance with the following quantity:

$$v_{\text{SIM}}(\hat{Y}_{\text{HT}}) = \frac{1}{m} \sum_s \{ \hat{Y}_{\text{HT}}(s) - Y \}^2. \tag{16}$$

Figure 6 shows a sample selected with the WAVE sampling. The filled black circles are selected units, while the hollow circles are those that are not selected in the sample. We

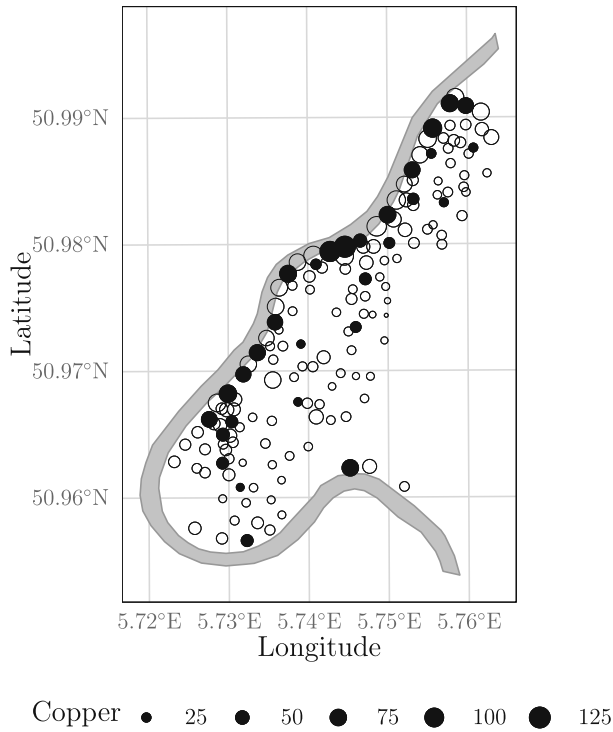


Figure 6. Example of WAVE sampling on the Meuse dataset. The overall population size is equal to 155. The inclusion probabilities are proportional to copper level variable and the sample size is equal to 30. Plotted sizes of the units are proportional to the copper concentration. The Meuse River is filled in light blue .

observe that the dataset is partially aggregated around the river showing a strong spatial correlation.

Results of the three spatial balanced measures on 10,000 simulated samples are given in Table 2. WAVE sampling performs better than other sampling designs in terms of I_B and I_{B_1} . In terms of spatial balance measure B , the algorithms are comparable to the artificial simulation, the differences are less marked.

Results of the simulations on the variance estimator in Table 3 show that the WAVE sampling strategy has a lower variance than the currently used method. This suggests that the method is more efficient in cases where there is a clear spatial correlation. A design-unbiased variance estimator does not exist for the Horvitz–Thompson estimator, but the spatially balanced estimator (14) seems to produce a good estimator for this dataset. Although the latter slightly overestimates the variance none of the other estimators seem to offer a better alternative. As there is no unbiased estimator we favour a slight overestimation of the variance. Table 4 shows the coverage rate as well as the ratio v_{SB}/v_{SIM} for all sampling methods.

Based on these simulation results, we are confident that we propose here a new method that allows to select a sample with a really strong degree of spreading. It performs better than the other sampling method. It can be generalized to higher dimensions and respects the unequal inclusion probabilities.

Table 2. Spreading measures results based on 10,000 simulations on the Meuse dataset. The population size is equal to 155

	Sampling design										
	Equal probabilities						Unequal probabilities				
	Wave	lpm1	scps	grts	hip	srswor	Wave	lpm1	scps	grts	maxent
I_{B_1}											
$n = 15$	-0.518	-0.338	-0.351	-0.226	-0.230	-0.030	-0.340	-0.250	-0.246	-0.165	-0.003
$n = 30$	-0.664	-0.427	-0.427	-0.266	-0.259	-0.019	-0.407	-0.298	-0.288	-0.172	0.024
$n = 50$	-0.796	-0.519	-0.473	-0.302	-0.248	-0.011	-0.466	-0.326	-0.285	-0.204	0.038
I_B											
$n = 15$	-0.518	-0.338	-0.351	-0.226	-0.230	-0.030	-0.354	-0.244	-0.247	-0.153	0.009
$n = 30$	-0.664	-0.427	-0.427	-0.266	-0.259	-0.019	-0.427	-0.290	-0.283	-0.154	0.048
$n = 50$	-0.796	-0.519	-0.473	-0.302	-0.248	-0.011	-0.455	-0.305	-0.263	-0.181	0.060
B											
$n = 15$	0.119	0.125	0.118	0.170	0.160	0.379	0.115	0.121	0.120	0.170	0.387
$n = 30$	0.118	0.123	0.126	0.164	0.159	0.359	0.120	0.121	0.120	0.162	0.345
$n = 50$	0.139	0.132	0.143	0.174	0.194	0.329	0.138	0.133	0.141	0.160	0.281

Table 3. Results of 10,000 simulations on Meuse dataset

	Sampling design										
	Equal probabilities						Unequal probabilities				
	Wave	lpm1	scps	grts	hip	srswor	Wave	lpm1	scps	grts	maxent
$n = 15$											
v_{SIM}	1.232	1.387	1.309	1.517	1.315	1.774	0.250	0.287	0.260	0.330	0.361
v_{SB}	1.847	1.670	1.635	1.596	1.701	1.455	0.393	0.362	0.371	0.333	0.321
v_{LM2}	0.962	0.889	0.889	0.855	0.930	0.786	0.224	0.206	0.209	0.194	0.183
v_{LM3}	1.301	1.256	1.261	1.230	1.308	1.147	0.293	0.279	0.282	0.269	0.259
v_{LM4}	1.463	1.445	1.452	1.430	1.487	1.352	0.325	0.315	0.319	0.306	0.299
v_{HAJ}	1.808	1.824	1.829	1.826	1.854	1.784	0.375	0.370	0.373	0.369	0.365
$n = 30$											
v_{SIM}	0.533	0.525	0.538	0.586	0.463	0.805	0.116	0.109	0.096	0.115	0.150
v_{SB}	0.692	0.687	0.670	0.657	0.639	0.634	0.154	0.153	0.152	0.150	0.143
v_{LM2}	0.382	0.373	0.370	0.362	0.356	0.348	0.094	0.090	0.090	0.089	0.082
v_{LM3}	0.555	0.543	0.543	0.534	0.534	0.512	0.130	0.127	0.127	0.126	0.118
v_{LM4}	0.654	0.649	0.649	0.641	0.652	0.616	0.150	0.148	0.148	0.147	0.140
v_{HAJ}	0.808	0.805	0.806	0.808	0.814	0.808	0.153	0.154	0.155	0.154	0.153
$n = 50$											
v_{SIM}	0.250	0.250	0.222	0.284	0.200	0.413	0.052	0.049	0.039	0.049	0.065
v_{SB}	0.380	0.375	0.385	0.353	0.337	0.344	0.081	0.078	0.080	0.080	0.080
v_{LM2}	0.214	0.208	0.213	0.196	0.190	0.190	0.050	0.048	0.049	0.048	0.045
v_{LM3}	0.308	0.294	0.298	0.284	0.280	0.276	0.068	0.068	0.069	0.068	0.065
v_{LM4}	0.358	0.349	0.351	0.340	0.337	0.330	0.079	0.080	0.081	0.081	0.078
v_{HAJ}	0.406	0.407	0.407	0.404	0.405	0.403	0.065	0.066	0.066	0.066	0.066

The population size is equal to 155. v_{SIM} (16) is equal to the variance approximated by the simulations. v_{SB} (14) is the variance estimator based on the nearest neighbours in the sample. v_{LM_j} is equal to the estimator (15) where the number of neighbouring units used is set to $j = 2, 3, 4$. v_{HAJ} (13) is the Hajek–Rosen estimator

Table 4. Results of 10,000 simulations on Meuse dataset

	Sampling design						Unequal probabilities				
	Equal probabilities						Wave	lpm1	scps	grts	maxent
	Wave	lpm1	scps	grts	hip	srswor					
v_{SIM}											
$n = 15$	1.232	1.387	1.309	1.517	1.315	1.774	0.250	0.287	0.260	0.330	0.361
$n = 30$	0.533	0.525	0.538	0.586	0.463	0.805	0.116	0.109	0.096	0.115	0.150
$n = 50$	0.250	0.250	0.222	0.284	0.200	0.413	0.052	0.049	0.039	0.049	0.065
v											
$n = 15$	1.847	1.670	1.635	1.596	1.701	1.784	0.393	0.362	0.371	0.333	0.365
$n = 30$	0.692	0.687	0.670	0.657	0.639	0.808	0.154	0.153	0.152	0.150	0.153
$n = 50$	0.380	0.375	0.385	0.353	0.337	0.403	0.081	0.078	0.080	0.080	0.066
Coverage of the 95% confidence interval											
$n = 15$	0.925	0.907	0.914	0.887	0.918	0.890	0.973	0.958	0.972	0.929	0.933
$n = 30$	0.953	0.943	0.942	0.929	0.963	0.924	0.971	0.972	0.983	0.966	0.942
$n = 50$	0.975	0.966	0.977	0.946	0.973	0.927	0.978	0.979	0.990	0.979	0.944
Ratio v/v_{SIM}											
$n = 15$	1.499	1.204	1.249	1.052	1.294	1.006	1.573	1.264	1.428	1.011	1.011
$n = 30$	1.298	1.307	1.246	1.121	1.380	1.003	1.323	1.400	1.588	1.308	1.016
$n = 50$	1.521	1.501	1.739	1.242	1.685	0.976	1.564	1.615	2.030	1.616	1.003

The population size is equal to 155. v_{SIM} is equal to the variance approximated by the simulations (16). v depends on the sampling design. For the srswor and maxent methods, we used the estimator v_{HAJ} (13), while for the other sampling designs, we use v_{SB} (14). Coverage rate of the 95% confidence intervals is computed as well as the ratio between averages of v and v_{SIM}

8. DISCUSSION

Environmental data are generally not uniformly distributed over a region of the space. Thus, it is generally justified to use unequal inclusion probabilities to overrepresent some parts of the population. As explained in Sect. 2.2, this reduces the variance of the Horvitz–Thompson estimator, a phenomenon also observed in Sect. 7 on the Meuse dataset.

In this manuscript, we present a sampling design that selects the units in a very well-spread configuration. We have shown on the Meuse dataset that on measurements of spatial spreading the method behaves very well. Moreover, the approximated variance of the Horvitz–Thompson estimator is lower with WAVE sampling than the other methods. Some second-order inclusion probabilities are null. It is thus impossible to estimate unbiasedly the variance of the estimator. However, we propose different estimators and compare their performance. We show that it is possible to estimate appropriately the variance and to construct confidence intervals that have good coverage rates, particularly when the sample size is large. All of these results indicate that our method is very efficient to select a well-spread sample and has better properties than the usual spatial sampling designs.

ACKNOWLEDGEMENTS

Open access funding provided by University of Neuchâtel. We would like to thank the associate editor and two reviewers for their conscientious reading and positive comments, which improved the quality of this manuscript.

We are grateful to Pierre-Yves Deléamont, Ziqing Dong, Esther Eustache, Cliona Jauslin and Lionel Qualité for their time spent to bring valuable comments at an early stage of this manuscript.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

[Received March 2020. Accepted July 2020. Published Online July 2020.]

REFERENCES

- Bellhouse, D. R. (1977). Some optimal designs for sampling in two dimensions. *Biometrika*, 64(3):605–611.
- Benedetti, R., Piersimoni, F., and Postiglione, P. (2017). Spatially balanced sampling: A review and a reappraisal. *International Statistical Review*, 85(3):439–454.
- Brown, J. A., Robertson, B. L., and McDonald, T. (2015). Spatially balanced sampling: application to environmental surveys. *Procedia Environmental Sciences*, 27:6–9.
- Deville, J.-C. and Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika*, 85:89–101.
- Deville, J.-C. and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91:893–912.
- Dickson, M. M. and Tillé, Y. (2016). Ordered spatial sampling by means of the traveling salesman problem. *Computational Statistics*, 31(4):1359–1372.
- Diggle, P. J., Menezes, R., and Su, T.-I. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(2):191–232.
- Eddelbuettel, D. and Sanderson, C. (2014). Rcpparmadillo: Accelerating R with high-performance c++ linear algebra. *Computational Statistics & Data Analysis*, 71:1054–1063.
- Grafström, A. and Lisic, J. (2019). *BalancedSampling: Balanced and Spatially Balanced Sampling*. R package version 1.5.5.
- Grafström, A. (2011). Spatially correlated Poisson sampling. *Journal of Statistical Planning and Inference*, 142:139–147.
- Grafström, A. and Lundström, N. L. P. (2013). Why well spread probability samples are balanced? *Open Journal of Statistics*, 3(1):36–41.
- Grafström, A., Lundström, N. L. P., and Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics*, 68(2):514–520.
- Grafström, A. and Schelin, L. (2014). How to select representative samples? *Scandinavian Journal of Statistics*, 41:277–290.
- Grafström, A. and Tillé, Y. (2013). Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. *Environmetrics*, 14(2):120–131.
- Hájek, J. (1981). *Sampling from a Finite Population*. Marcel Dekker, New York.
- Hankin, D., Mohr, M., and Newman, K. (2019). *Sampling Theory: For the Ecological and Natural Resource Sciences*. Oxford University Press, New York.
- Kincaid, T. M., Olsen, A. R., and Weber, M. H. (2019). *spsurvey: Spatial Survey Design and Analysis*. R package version 4.1.0.

- Marker, D. A. and Stevens Jr., D. L. (2009). Sampling and inference in environmental surveys. In *Sample surveys: design, methods and applications*, volume 29 of *Handbook of Statistics*, pages 487–512. Elsevier/North-Holland, New York, Amsterdam.
- Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23.
- Neyman, J. and Scott, E. L. (1958). Statistical approach to problems of cosmology. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(1):1–29.
- Pebesma, E. J. and Bivand, R. S. (2005). Classes and methods for spatial data in R. *R News*, 5(2):9–13.
- Quenouille, M. H. (1949). Problems in plane sampling. *The Annals of Mathematical Statistics*, 20:355–375.
- Robertson, B., McDonald, T., Price, C., and Brown, J. (2018). Halton iterative partitioning: spatially balanced sampling via partitioning. *Environmental and Ecological Statistics*, 25:305–323.
- Sen, A. R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 5:119–127.
- Stevens Jr., D. L. and Olsen, A. R. (1999). Spatially restricted surveys over time for aquatic resources. *Journal of Agricultural, Biological, and Environmental Statistics*, 4:415–428.
- Stevens Jr., D. L. and Olsen, A. R. (2003). Variance estimation for spatially balanced samples of environmental resources. *Environmetrics*, 14(6):593–610.
- Stevens Jr., D. L. and Olsen, A. R. (2004). Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, 99(465):262–278.
- Theobald, D. M., Stevens Jr., D. L., White, D. E., Urquhart, N. S., Olsen, A. R., and Norman, J. B. (2007). Using GIS to generate spatially balanced random survey designs for natural resource applications. *Environmental Management*, 40(1):134–146.
- Tillé, Y. (2006). *Sampling Algorithms*. Springer, New York.
- Tillé, Y. (2020). *Sampling and estimation from finite populations*. Wiley, New York.
- Tillé, Y., Dickson, M. M., Espa, G., and Giuliani, D. (2018). Measuring the spatial balance of a sample: A new measure based on the Moran's I index. *Spatial Statistics*, 23:182–192.
- Tillé, Y. and Wilhelm, M. (2017). Probability sampling designs: Balancing and principles for choice of design. *Statistical Science*, 32(2):176–189.
- 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.
- Wang, J.-F., Stein, A., Gao, B.-B., and Ge, Y. (2012). A review of spatial sampling. *Spatial Statistics*, 2:1–14.
- Yates, F. and Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, B15:235–261.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.