Spatial Spread Sampling
Using Weakly Associated Vectors

Raphaël Jauslin and Yves Tillé

Abstract
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 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 (GRTS) or the Local Pivotal Method (LPM). Analysis of the variance on a real dataset shows that our method is more accurate than these two. Furthermore, a variance estimator is proposed.

Keywords: GRTS, local pivotal method, cube method, stratification

1 Introduction
Geo-referenced data are usually autocorrelated, meaning that two close measurements are often similar. By spreading the measurements in the space, more information is collected. An important problem of spatial sampling is thus to spread at best the sampled units in space. Tillé and Wilhelm (2017), Wang et al. (2012), Benedetti et al. (2017) and Tillé (2020, Chapter 8) give a review of the main spatial sampling methods. The simplest method is systematic sampling. This method provides the best spread of the sample. Geographical stratification with one or a small number of sampled units in each stratum is also an interesting solution. Unfortunately, these two methods cannot be applied when units are selected with unequal probabilities or when units of the population are not evenly distributed in space.

Generalized Random Tessellation Stratified (GRTS) sampling is a spatial sampling method proposed by Stevens Jr. 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). Dickson and Tillé (2016) have simply used the Traveling 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 indicators of nearby units. The pivotal method is one of the particular cases of the splitting method proposed by Deville and Tillé (1998). It consists of randomly confronting two units at each step. Grafström et al. (2012) proposed the Local Pivotal Method (LPM), where two neighbouring units are confronted at each step, which 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 BalanceSampling R package (Grafström and Lisic, 2019).

Stevens Jr. 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.

In this paper, we propose a new spatial sampling method based on a completely different approach. The algorithm starts 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 well choosing 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 setup of the problem. In Section 3, we introduce the new method that we propose and the process of sample selection. In Section 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 Section 5, we present a variance estimator for our method. In Section 6, we give simulation results of the algorithm on artificial spatial configurations while Section 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 Basic setup and notation

Consider a finite population $U$ of size $N$ whose units can be defined by labels $k \in \{1, 2, \ldots, N\}$. Let $S = \{s | s \subset U\}$ 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 $S$ such that

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

A random sample $S$ is a random variable that takes elements of $S$ as value such that

$$P(S = s) = p(s).$$

Define $a_k(S)$, for $k = 1, \ldots, N$:

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

Then a sample can be denoted by means of a vector notation: $a^\top = (a_1, a_2, \ldots, a_N)$. 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 $S$:

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

Let $\pi^\top = (\pi_1, \ldots, \pi_N)$ be the vector of inclusion probabilities. Then, $E(a) = \pi$. 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

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

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

$$\hat{Y}_{HT} = \sum_{k \in U} \frac{y_k a_k}{\pi_k}. \quad (1)$$

Usually, some auxiliary information $x_k^\top = (x_{k1}, x_{k2}, \ldots, x_{kp})$ regarding the population units is available. In the particular case of spatial sampling, a set of spatial coordinates $z_k^\top = (z_{k1}, z_{k2}, \ldots, z_{kp}) \in \mathbb{R}^p$ 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

$$\hat{X} = \sum_{k \in S} \frac{x_k}{\pi_k} = \sum_{k \in U} x_k = X.$$ 

Grafström and Lundström (2013) showed that well spread probability samples are relatively well balanced.

3 WAVE sampling

3.1 General idea

Our sampling algorithm 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. 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 describe in sections 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

$$m : U \times U \to \mathbb{R}^+,$$  

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

$m(x, y) \geq 0$, $m(x, y) = 0 \iff x = y$, $m(x, y) = m(y, x)$, $m(x, z) \leq m(x, y) + m(y, z)$.

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

$$m^2_E(k, \ell) = (z_k - z_\ell) ^\top (z_k - z_\ell),$$  

where $z_k$ and $z_\ell$ 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,

$$m^2_M(k, \ell) = (x_k - x_\ell) ^\top S^{-1} (x_k - x_\ell),$$

where

$$S = \frac{1}{N} \sum_{k \in U} (x_k - \bar{x})(x_k - \bar{x}) ^\top, \quad \bar{x} = \frac{1}{N} \sum_{k \in U} x_k.$$
When the population is distributed on a $N_1 \times N_2$ regular grid of $\mathbb{R}^2$, a tore distance can be defined. With a tore distance, two units on the same column (respectively row) that are on the opposite side have a small distance. It is like seeing the grid curved such that it looks like a regular tore. The distance is then defined by:

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

(4)

**Example 3.1:** Let \{1, \ldots, 9\} be on a regular grid of size $3 \times 3$, then the squared distance matrices defined by Equations (3) and (4) are equal to

$$M_E = \begin{pmatrix}
0 & 1 & 4 & 1 & 2 & 5 & 4 & 5 & 8 \\
1 & 0 & 1 & 2 & 1 & 2 & 5 & 4 & 5 \\
4 & 1 & 0 & 5 & 2 & 1 & 8 & 5 & 4 \\
1 & 2 & 5 & 0 & 1 & 4 & 1 & 2 & 5 \\
2 & 1 & 2 & 1 & 0 & 1 & 2 & 1 & 2 \\
5 & 2 & 1 & 4 & 1 & 0 & 5 & 2 & 1 \\
4 & 5 & 8 & 1 & 2 & 5 & 0 & 1 & 4 \\
5 & 4 & 5 & 2 & 1 & 2 & 1 & 0 & 1 \\
8 & 5 & 4 & 5 & 2 & 1 & 4 & 1 & 0
\end{pmatrix}, \quad M_T = \begin{pmatrix}
0 & 1 & 1 & 1 & 2 & 2 & 1 & 2 & 2 \\
1 & 0 & 1 & 2 & 1 & 2 & 1 & 2 & 2 \\
1 & 1 & 0 & 2 & 2 & 1 & 1 & 2 & 2 \\
1 & 2 & 2 & 0 & 1 & 1 & 1 & 2 & 2 \\
2 & 1 & 2 & 1 & 0 & 1 & 2 & 1 & 2 \\
2 & 2 & 1 & 1 & 1 & 0 & 2 & 2 & 1 \\
1 & 2 & 2 & 1 & 2 & 0 & 1 & 1 & 1 \\
2 & 1 & 2 & 2 & 1 & 2 & 1 & 0 & 1 \\
2 & 2 & 2 & 2 & 1 & 1 & 1 & 0 & 0
\end{pmatrix}.$$  

(5)

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 $\varepsilon \in \mathbb{R}^2$ and $\tilde{z}_k = z_k + \varepsilon$ the shifted coordinates, Equation (4) is then replaced by,

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

$\varepsilon$ is called a “shift” and $M_S$ the shifted version of $M_T$, for example if $\varepsilon = (1/12, 1/4)$, the distance matrix $M_S$ becomes,

$$M_S = \begin{pmatrix}
0 & 0.90 & 1.24 & 0.57 & 1.40 & 1.74 & 1.57 & 2.40 & 2.74 \\
1.24 & 0.90 & 1.74 & 0.57 & 1.40 & 2.74 & 1.57 & 2.40 & \ \\
0.90 & 1.24 & 0 & 1.40 & 1.74 & 0.57 & 2.40 & 2.74 & 1.57 \\
1.57 & 2.40 & 2.74 & 0 & 0.90 & 1.24 & 0.57 & 1.40 & 1.74 \\
2.74 & 1.57 & 2.40 & 1.24 & 0 & 0.90 & 1.74 & 0.57 & 1.40 \\
2.40 & 2.74 & 1.57 & 0.90 & 1.24 & 0 & 1.40 & 1.74 & 0.57 \\
0.57 & 1.40 & 1.74 & 2.40 & 1.57 & 2.40 & 0 & 0.90 & 1.24 \\
1.74 & 0.57 & 1.40 & 2.74 & 1.57 & 2.40 & 1.24 & 0 & 0.90 \\
1.40 & 1.74 & 0.57 & 2.40 & 2.74 & 1.57 & 0.90 & 1.24 & 0
\end{pmatrix}.$$  

(6)

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, 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 Figure 1, three simple configurations are presented: Euclidean, tore and shifted tore distance on a \(3 \times 3\) regular grid. In shifted distance graph, all the distances from point \((1, 1)\) to the other grid points are different.

Fig. 1: Simple example of a \(3 \times 3\) 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 (3), the central graph is the tore distance given in (4) and the right one is the shifted tore distance with a shift equal to \((1/12, 1/4)\) (the red 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 (5).

### 3.3 The stratification matrix

Let \(k \in U\) a unit in the population. 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 \(g_k = \#G_k\), the spatial weights are then defined as follows

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

(7)
\( W \) denote an \( N \times N \) stratification matrix and each row of matrix \( W \) represents a stratum. Each stratum is defined by a particular unit and its neighbouring units. Nearest neighbours are defined with a metric function (2). 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 (6)) such that all the distances are different. Each row of matrix \( W \) sum to 1. Thus matrix \( W \) is a right stochastic matrix. Most of the components of matrix \( W \) are null. \( W \) can thus be encoded as a sparse matrix.

**Example 3.2:** Let \( \{1, \ldots, 9\} \) be on a regular grid of size \( 3 \times 3 \) with inclusion probabilities equal \( \pi_k = 1/3 \), for all \( k \in U \). Figure 2 shows different stratification matrices corresponding to \( M_E, M_T \) and \( M_S \) with a shift randomly generated from a random variable \( \mathcal{N}(0, 1/100I) \).

![Fig. 2: Sparsity pattern of three stratification matrices. Spatial coordinates are 3 × 3 regular grid and the inclusion probabilities are equal to \( \pi = (1/3, \ldots, 1/3) \). Depending on the way of defining the nearest neighbours in Equation (7), different weight values are obtained. The left stratification matrix uses the classical Euclidean distance (3), the central one the tore distance (4) and the right one uses a shifted tore distance with a shift randomly generated from a random variable \( \mathcal{N}(0, 1/100I) \).](image-url)
define $\mathbf{A}$ by

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

(8)

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 3.3: 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 ties distance value. Since all $1/\pi_k = 25$ the non-zero entries of $\mathbf{A}$ are all equal to 1. Based on the definition (7), 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.

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 $\pi$ is randomly modified. To modify $\pi$, 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:

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

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

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

which could also be rewrite:

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

where

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

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 to the smallest singular value $\sigma$ of $\mathbf{A}$ i.e,

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

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
Fig. 3: Representation of the strata defined by the spatial weights Equation (7).
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 (8). All entries of the matrix are equal to 1.

A. Vector $v$ 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 $\pi$ are modified and at least one component is set to 0 or 1. Matrix $A$ 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 is implemented in the R package WaveSampling (Jauslin and Tillé, 2019), 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 $A$ given in (8) 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 measure than those currently used for the spatial balanced sampling design.
Algorithm 1 Algorithm for WAVE sampling

Let $A = A_0$ and $\pi_0 = (\pi_1^{(0)}, \ldots, \pi_N^{(0)}) = \pi$ for the initialization step. For $t = 0, 1, 2, \ldots$

1. From $\pi_t$, extract $\tilde{\pi}$, vector $\pi_t$ restricted to the $k$ such that $0 < \pi_k^{(t)} < 1$. Let $J$ be the length of $\tilde{\pi}$.

2. Compute the $J \times J$ matrix $A_t$ of Equation (8) using inclusion probabilities $\tilde{\pi}_t$.

3. Calculate the rank $r$ of matrix $A_t$.
   
   (a) If matrix $A_t$ does not have full rank, choose $v_t = (v_1^{(t)}, \ldots, v_J^{(t)}) \in \mathbb{R}^J$ a vector in the right null space of $A_t$.

   (b) If matrix $A_t$ has full rank, compute the singular value decomposition and seek for $v_t$ a right singular vector associated to the smallest singular value $\sigma_t$.

4. Next in order to ensure the fixed sample size, vector $v_t$ is centered:

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

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

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

6. Compute

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

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

4 Spreading measures

4.1 Voronoï polygons

Stevens Jr. and Olsen (2004) suggested the spatial balance of a sample consists of using the Voronoï polygons. The Voronoï polygons associated to 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
4. Spreading measures

$v_k$ 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 spreaded, $v_k$ 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

$$
B(S) = \frac{1}{n} \sum_{k \in S} (v_k - 1)^2. \tag{9}
$$

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 polygons, 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 in Fig. 7 and Fig. 8. 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,

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

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

$$
I_B(a) = \frac{(a - \bar{a}_w)^\top W (a - \bar{a}_w)}{\sqrt{(a - \bar{a}_w)^\top D(a - \bar{a}_w)(a - \bar{a}_w)^\top B(a - \bar{a}_w)}} \tag{10}
$$

where $a$ is the sample and

$$
\bar{a}_w = \frac{a^\top W 1}{1^\top W 1}.
$$
Fig. 4: Illustrated example of how the spatial balance measure based on the Voronoï 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.

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

$$
\mathbf{B} = \mathbf{C}^\top \mathbf{D} \mathbf{C}, \quad \mathbf{C} = \mathbf{D}^{-1} \mathbf{W} - \frac{11^\top \mathbf{W}}{1^\top \mathbf{W} 1},
$$

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 $a_\ell$ that are in the neighbouring 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 neighbouring 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 $[h_k]$ and $\lceil h_k \rceil$ be respectively the inferior and superior integers of $h_k$. Spatial weights are then defined as follows,

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

(11)
For example, if a unit $k$ has an inclusion probability of $\pi_k = 0.35$ then $h_k \approx 2.857$. 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) suggests 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 (7) rather than the weights (11). We set $w_{kk} = 0$ for all $k \in U$. For the rest of the paper, $I_B$ will represent the measure based on the spatial weights (11) and $I_B^1$ the one based on (7).

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

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

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:

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

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 $\text{var}(\hat{Y}_{HT})$.

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

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

where

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

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

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

(12)
6 Simulations on artificial spatial configurations

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. Stevens Jr. and Olsen (2003) proposed an estimator based on a local neighbourhood for each units in the sample. It is called the local mean variance estimator and is given by

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

(13)

where the weights $w_{k\ell}$ are computed such that they vary inversely as $\pi_\ell$ and decrease as the distance between unit $k$ and $\ell$ 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. As the WAVE sampling seems to select a more spatially spread sample than the GRTS method, we observe that $v_{LM}(\hat{Y}_{HT})$ overestimates the variance. We then proposed to decrease the neighbourhood size to two (three with the included unit) for the WAVE sampling. This should produce better estimates of the variance.

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 $N = 225$.

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 15 circular discs of radius 0.055 with units uniformly distributed around the centre. Each cluster contains 15 units such that the population target size is equal to $N$.

3. Simple regular grid of size $15 \times 15$.

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 $n = 75$ 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.
Simulations on artificial spatial configurations

Complete spatial randomness  Neyman-Scott process  Simple regular grid

Fig. 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 \( k \in U \).

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 Voronoï polygons (9) and measures based on Moran’s \( I \) index (10). The simulations results of the CRS dataset are given in the 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 Voronoï 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 Voronoï 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 (11). This result comes from the heterogeneity of the strata and the randomness of the algorithm. If the inclusion probabilities of a unit is nearly zero,
Tab. 1: Spreading measures results based on 10000 simulations on the Complete spatial randomness dataset. The population size is equal to 219.

| Sampling design | Equal probabilities | Unequal probabilities |
|-----------------|---------------------|-----------------------|
|                 | wave    | lpm1  | scps | grts | srswhr | wave    | lpm1  | scps | grts | maxent |
| IB₁             |         |       |      |      |        |         |       |      |      |        |
| n = 25          | -0.565  | -0.375 | -0.400 | -0.235 | -0.020 | -0.180  | -0.145 | -0.143 | -0.085 | 0.028   |
| n = 50          | -0.713  | -0.509 | -0.493 | -0.321 | -0.012 | -0.222  | -0.169 | -0.162 | -0.106 | 0.055   |
| n = 100         | -0.844  | -0.614 | -0.532 | -0.404 | -0.006 | -0.353  | -0.307 | -0.275 | -0.220 | -0.025  |
| IB             |         |       |      |      |        |         |       |      |      |        |
| n = 25          | -0.565  | -0.375 | -0.400 | -0.235 | -0.020 | -0.493  | -0.341 | -0.353 | -0.223 | -0.024  |
| n = 50          | -0.713  | -0.509 | -0.493 | -0.321 | -0.012 | -0.587  | -0.419 | -0.407 | -0.292 | -0.012  |
| n = 100         | -0.844  | -0.614 | -0.532 | -0.404 | -0.006 | -0.497  | -0.375 | -0.343 | -0.269 | -0.032  |
| B               |         |       |      |      |        |         |       |      |      |        |
| n = 25          | 0.106   | 0.107 | 0.108 | 0.156 | 0.357  | 0.121   | 0.119 | 0.126 | 0.173 | 0.365   |
| n = 50          | 0.135   | 0.131 | 0.134 | 0.179 | 0.370  | 0.141   | 0.141 | 0.146 | 0.176 | 0.356   |
| n = 100         | 0.170   | 0.159 | 0.170 | 0.200 | 0.345  | 0.167   | 0.170 | 0.177 | 0.190 | 0.288   |

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 IB should be prefered to IB₁.

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 x 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 75 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:

\[ v_{SIM}(\hat{Y}_{HT}) = \frac{1}{m} \sum_s \left( \hat{Y}_{HT}(s) - Y \right)^2. \] (14)
Fig. 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.

Figure 6 shows 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 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 is 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 shows 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-Tompson estimator, but the local mean variance (13) with only three neighbours seems to produce a good estimator for this dataset.
Tab. 2: Spreading measures results based on 10000 simulations on the Meuse dataset. The population size is equal to 155.

| Sampling design | Equal probabilities | Unequal probabilities |
|-----------------|----------------------|-----------------------|
|                 | wave | lpm | scps | grts | srswo | wave | lpm | scps | grts | maxen |
| I_B_1           |      |     |      |      |       |      |     |      |      |       |
| n = 15          | -0.518 | -0.337 | -0.352 | -0.224 | -0.030 | -0.340 | -0.250 | -0.246 | -0.162 | -0.004 |
| n = 30          | -0.663 | -0.428 | -0.429 | -0.267 | -0.018 | -0.407 | -0.297 | -0.288 | -0.172 | 0.024  |
| n = 75          | -0.926 | -0.624 | -0.508 | -0.365 | -0.007 | -0.537 | -0.372 | -0.328 | -0.226 | 0.033  |
| I_B_2           |      |     |      |      |       |      |     |      |      |       |
| n = 15          | -0.518 | -0.337 | -0.352 | -0.224 | -0.030 | -0.355 | -0.244 | -0.248 | -0.151 | 0.006  |
| n = 30          | -0.663 | -0.428 | -0.429 | -0.267 | -0.018 | -0.427 | -0.290 | -0.284 | -0.153 | 0.047  |
| n = 75          | -0.926 | -0.624 | -0.508 | -0.365 | -0.007 | -0.532 | -0.369 | -0.315 | -0.216 | 0.040  |
| B               |      |     |      |      |       |      |     |      |      |       |
| n = 15          | 0.119 | 0.126 | 0.118 | 0.170 | 0.382 | 0.115 | 0.121 | 0.120 | 0.171 | 0.383  |
| n = 30          | 0.119 | 0.123 | 0.125 | 0.163 | 0.357 | 0.120 | 0.120 | 0.120 | 0.162 | 0.345  |
| n = 75          | 0.162 | 0.148 | 0.153 | 0.184 | 0.295 | 0.115 | 0.113 | 0.120 | 0.135 | 0.210  |

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.

8 Acknowledgments

We would like to thank Pierre-Yves Deléamont, Ziqing Dong, Esther Eustache, Cliona Jauslin and Lionel Qualité for their time spent to bring valuable comments on this manuscript.
Tab. 3: Results of 10000 simulations on Meuse dataset. The population size is equal to 155. $v_{\text{sim}}$ is equal to the variance approximated by the simulations (14). $v$ depends on the sampling design. For the srswor and maxent methods, we used the estimator $v_{HAJ}$ (12) while for the other sampling designs, we use $v_{LM}$ (13). In the wave sampling design, the parameter for the neighbouring is set to three instead of four. Coverage rate of the 95% confidence intervals are computed as well as the ratio between averages of $v$ and $v_{\text{SIM}}$.

| Sampling design | Equal probabilities | Unequal probabilities |
|-----------------|----------------------|----------------------|
|                 | wave | lpm1 | scps | grts | srswor | wave | lpm1 | scps | grts | maxent |
| $v_{\text{SIM}}$ |      |      |      |      |        |      |      |      |      |        |
| n = 15          | 1.270 | 1.350 | 1.320 | 1.525 | 1.824 | 0.257 | 0.283 | 0.259 | 0.331 | 0.371 |
| n = 30          | 0.537 | 0.530 | 0.539 | 0.580 | 0.830 | 0.117 | 0.109 | 0.100 | 0.114 | 0.152 |
| n = 75          | 0.099 | 0.113 | 0.101 | 0.128 | 0.205 | 0.017 | 0.015 | 0.015 | 0.018 | 0.020 |
| $v$             |      |      |      |      |        |      |      |      |      |        |
| n = 15          | 0.970 | 1.431 | 1.440 | 1.412 | 1.816 | 0.223 | 0.316 | 0.320 | 0.306 | 0.366 |
| n = 30          | 0.385 | 0.652 | 0.651 | 0.641 | 0.795 | 0.095 | 0.148 | 0.148 | 0.148 | 0.153 |
| n = 75          | 0.140 | 0.214 | 0.215 | 0.210 | 0.206 | 0.033 | 0.057 | 0.057 | 0.057 | 0.020 |
| Coverage of the 95% confidence interval |      |      |      |      |        |      |      |      |      |        |
| n = 15          | 0.856 | 0.904 | 0.912 | 0.884 | 0.892 | 0.917 | 0.956 | 0.975 | 0.931 | 0.929 |
| n = 30          | 0.879 | 0.946 | 0.944 | 0.933 | 0.914 | 0.904 | 0.971 | 0.985 | 0.966 | 0.943 |
| n = 75          | 0.970 | 0.987 | 0.993 | 0.982 | 0.941 | 0.989 | 0.999 | 0.999 | 0.998 | 0.945 |
| Ratio $v/v_{\text{SIM}}$ |      |      |      |      |        |      |      |      |      |        |
| n = 15          | 0.764 | 1.060 | 1.091 | 0.926 | 0.996 | 0.868 | 1.117 | 1.235 | 0.923 | 0.988 |
| n = 30          | 0.718 | 1.230 | 1.209 | 1.104 | 0.958 | 0.808 | 1.355 | 1.478 | 1.295 | 1.013 |
| n = 75          | 1.423 | 1.895 | 2.127 | 1.641 | 1.003 | 1.902 | 3.771 | 3.702 | 3.180 | 0.998 |
Supplementary Material

This section presents the additional results of the Section 6. For each spatial configurations presented in the paper, the three spatial measures are given. For each population, 10,000 samples of size respectively equal to 25, 50 and 100 are selected. Two cases are considered for the inclusion probabilities: equal and unequal. In terms of spatial measures, the WAVE sampling method performs better than the others do. Examples where the spatial balance (9) behaves badly are given in Fig. 7 and Fig. 8.
Tab. 4: Results of 10000 simulations of Moran’s $I_B$ spatial measure (11). The three spatial configurations of the Section 6 are taken into account.

| Sampling design | Equal probabilities | Unequal probabilities |
|-----------------|---------------------|-----------------------|
| $I_B$           | wave lpm1 scps srswor grts | wave lpm1 scps maxent grts |
| Complete spatial randomness | | |
| $n = 25$        | -0.565 -0.375-0.400 -0.235 -0.020 | -0.180 -0.145 -0.143 -0.085 0.028 |
| $n = 50$        | -0.713 -0.509 -0.493 -0.321 -0.012 | -0.222 -0.169 -0.162 -0.106 0.055 |
| $n = 100$       | -0.844 -0.614 -0.532 -0.404 -0.006 | -0.353 -0.307 -0.275 -0.220 0.025 |
| Neyman-Scott process | | |
| $n = 25$        | -0.561 -0.350 -0.321 -0.250 -0.020 | -0.260 -0.222 -0.194 -0.159 -0.007 |
| $n = 50$        | -0.696 -0.456 -0.409 -0.323 -0.012 | -0.265 -0.216 -0.184 -0.156 0.015 |
| $n = 100$       | -0.820 -0.548 -0.479 -0.368 -0.005 | -0.351 -0.283 -0.239 -0.214 0.025 |
| Simple regular grid | | |
| $n = 25$        | -0.569 -0.341 -0.383 -0.211 -0.017 | -0.023 0.005 0.000 0.039 0.103 |
| $n = 50$        | -0.613 -0.397 -0.469 -0.311 -0.011 | -0.376 -0.245 -0.282 -0.197 0.008 |
| $n = 100$       | -0.825 -0.442 -0.583 -0.333 -0.008 | -0.423 -0.264 -0.323 -0.203 0.009 |

Tab. 5: Results of 10000 simulations of Moran’s $I_B$ spatial measure (7). The three spatial configurations of the Section 6 are taken into account.

| Sampling design | Equal probabilities | Unequal probabilities |
|-----------------|---------------------|-----------------------|
| $I_B$           | wave lpm1 scps srswor grts | wave lpm1 scps maxent grts |
| Complete spatial randomness | | |
| $n = 25$        | -0.565 -0.375 -0.400 -0.235 -0.020 | -0.493 -0.341 -0.353 -0.223 -0.024 |
| $n = 50$        | -0.713 -0.509 -0.493 -0.321 -0.012 | -0.587 -0.419 -0.407 -0.292 -0.012 |
| $n = 100$       | -0.844 -0.614 -0.532 -0.404 -0.006 | -0.497 -0.375 -0.343 -0.269 -0.032 |
| Neyman-Scott process | | |
| $n = 25$        | -0.561 -0.350 -0.321 -0.250 -0.020 | -0.451 -0.309 -0.273 -0.216 -0.017 |
| $n = 50$        | -0.696 -0.456 -0.409 -0.323 -0.012 | -0.573 -0.402 -0.339 -0.286 -0.008 |
| $n = 100$       | -0.820 -0.548 -0.479 -0.368 -0.005 | -0.565 -0.394 -0.321 -0.273 -0.006 |
| Simple regular grid | | |
| $n = 25$        | -0.569 -0.341 -0.383 -0.211 -0.017 | -0.446 -0.308 -0.337 -0.197 -0.017 |
| $n = 50$        | -0.613 -0.397 -0.469 -0.311 -0.011 | -0.539 -0.345 -0.394 -0.277 -0.010 |
| $n = 100$       | -0.825 -0.442 -0.583 -0.333 -0.008 | -0.509 -0.286 -0.340 -0.211 0.006 |
Tab. 6: Results of 10000 simulations of the spatial balance measure \( B \) (9). The three spatial configurations of the Section 6 are taken into account.

| \( B \) | Equal probabilities | | Unequal probabilities | |  |
|---|---|---|---|---|---|
| | wave | lpm1 | scps | srs | grts | wave | lpm1 | scps | maxent | grts |
| Complete spatial randomness | | | | | | | | | |
| \( n = 25 \) | 0.106 | 0.107 | 0.108 | 0.156 | 0.357 | 0.121 | 0.119 | 0.126 | 0.173 | 0.365 |
| \( n = 50 \) | 0.135 | 0.131 | 0.134 | 0.179 | 0.370 | 0.141 | 0.141 | 0.146 | 0.176 | 0.356 |
| \( n = 100 \) | 0.170 | 0.159 | 0.170 | 0.200 | 0.345 | 0.167 | 0.170 | 0.177 | 0.190 | 0.288 |
| Neyman-Scott process | | | | | | | | | |
| \( n = 25 \) | 0.155 | 0.153 | 0.172 | 0.198 | 0.539 | 0.156 | 0.150 | 0.174 | 0.207 | 0.528 |
| \( n = 50 \) | 0.159 | 0.154 | 0.174 | 0.195 | 0.512 | 0.151 | 0.144 | 0.169 | 0.194 | 0.493 |
| \( n = 100 \) | 0.185 | 0.172 | 0.189 | 0.215 | 0.428 | 0.171 | 0.170 | 0.190 | 0.211 | 0.382 |
| Simple regular grid | | | | | | | | | |
| \( n = 25 \) | 0.042 | 0.068 | 0.054 | 0.102 | 0.279 | 0.061 | 0.080 | 0.068 | 0.112 | 0.284 |
| \( n = 50 \) | 0.049 | 0.065 | 0.049 | 0.078 | 0.234 | 0.074 | 0.084 | 0.071 | 0.099 | 0.251 |
| \( n = 100 \) | 0.027 | 0.048 | 0.036 | 0.053 | 0.153 | 0.063 | 0.075 | 0.070 | 0.085 | 0.167 |

Fig. 7: Let \( U = \{1, \ldots, 36\} \) be on a regular grid \( 6 \times 6 \), inclusion probabilities are all equal to \( \pi_k = \pi = 1/6 \). Measure \( B \) given in Equation (9) is equal to 0 for the clustered sample while it should be a great value. \( B = 0.106 \) for the spreaded sample. On the other hand the Moran index given in Equation (10) shows a correct behaviour with the two samples. For the clustered sample \( I_B = 0.71 \) while \( I_B = -0.718 \) for the spreaded sample.
Fig. 8: Let $U = \{1, \ldots, 36\}$. On the left figure, population is on a regular grid $6 \times 6$. On the right figure, population is clustered around four units. Inclusion probabilities are all equal to $\pi_k = \pi = 1/9$ (respectively $\pi_k = \pi = 1/12$). Measure $B$ (9) for the both example is equal to 0 while it should be a great value. On the other hand the Moran index (10) shows a correct behaviour with the two samples. For the Example 1 $I_B = 0.732$ while $I_B = 0.232$ for Example 2.
References

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.

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 Líšic, 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 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.

Jauslin, R. and Tillé, Y. (2019). *WaveSampling: Weakly Associated Vectors (WAVE) Sampling*. R package version 0.1.0.

Kincaid, T. M., Olsen, A. R., and Weber, M. H. (2019). *spsurvey: Spatial Survey Design and Analysis*. R package version 4.1.0.

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.

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.

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.