Published in Environmetrics 24, issue 2, 120-131, 2013
which should be used for any reference to this work

Doubly balanced spatial sampling with spreading and restitution of auxiliary totals

Anton Grafström[a]* and Yves Tillé[b]

A new spatial sampling method is proposed in order to achieve a double property of balancing. The sample is spatially balanced or well spread so as to avoid selecting neighbouring units. Moreover, the method also enables to satisfy balancing equations on auxiliary variables available on all the sampling units because the Horvitz–Thompson estimator is almost equal to the population totals for these variables. The method works with any definitio of distance in a multidimensional space and supports the use of unequal inclusion probabilities. The algorithm is simple and fast. Examples show that the method succeeds in using more information than the local pivotal method, the cube method and the Generalized Random-Tessellation Stratifie sampling method, and thus performs better. An estimator of the variance for this sampling design is proposed in order to lead to an inference that takes the effect of the sampling design into account.

Keywords: balanced sampling; pivotal method, spatially balanced sampling; spatial correlation

1. INTRODUCTION

Most of the samples are selected from space. This is the case in environmental studies, geology, geography, population biology and even in officia statistics. Establishments and households always have geographical coordinates. Statistical units selected from a territory are generally spatially correlated, which means that two neighbouring statistical units tend to be more similar than two distant statistical units. A large set of publications are dedicated to methods of spatial sampling that takes into account spatial correlation. The most usual methods are systematic sampling, spatial stratification Generalized Random-Tessellation Stratifie (GRTS) sampling (see among others Ripley, 1981; Thompson, 1992; Stevens and Olsen, 2003, 2004; Mandallaz, 2008; Marker and Stevens, 2009).

In this paper, we propose a new spatial sampling method that takes the spatial correlation into account but can also take advantage from auxiliary information available for the statistical units. Indeed, in many survey problems, auxiliary information is available for all the units of the population of interest under the form of a census or a register. The auxiliary information can be spatial coordinates and/or any other variables related to the variable of interest. Let $U = \{1, 2, \ldots, N\}$ denote the population of N units. We wish to estimate a total of some study variable y, which takes a fi ed value y_k for unit $k \in U$. A vector $\mathbf{x}_k = (x_{k1}, x_{k2}, \ldots, x_{kp})^T$ of the values taken by p auxiliary variables is supposed to be known for each unit of the population. The spatial coordinates of unit k are also supposed to be known.

We aim to combine two main ideas for constructing efficien sampling designs that make the best possible use of available auxiliary information. The firs main idea is the use of balanced sampling. Deville and Tillé (2004) introduced the cube method, which allows to select unequal or equal probability samples that are balanced or almost balanced on several auxiliary variables. Balanced sampling means that the Horvitz–Thompson (HT) estimator (Horvitz and Thompson, 1952) of the total of these auxiliary variables given by

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

is equal or almost equal to the known totals given by

$$\mathbf{X} = \sum_{k \in U} \mathbf{x}_k$$

that is,

$$\widehat{\mathbf{X}} \approx \mathbf{X}$$

* *Correspondence to: Anton Grafström, Department of Forest Resource Management, Swedish University of Agricultural Sciences, SE-90183 Umeå, Sweden. E-mail: anton.grafstrom@slu.se*

a *Department of Forest Resource Management, Swedish University of Agricultural Sciences, SE-90183 Umeå, Sweden*

b *Institute of Statistics, Faculty of Economics, University of Neuchâtel, Pierre à Mazel 7, 2000 Neuchâtel, Switzerland*

where S denotes the random sample and π_k the inclusion probability of unit k. Balanced sampling is very efficien when the study variable can be well approximated by a linear combination of the auxiliary variables (Nedyalkova and Tillé, 2009).

The second main idea is spatially balanced sampling, which means that the samples are well spread in the space so as to avoid selecting neighbouring units. Stevens and Olsen (2004) introduced GRTS sampling. Their method uses a specifi random mapping from two-dimensional or multidimensional locations to one dimension. The sample is then selected by a systematic design in one dimension and mapped back to two or more dimensions. This procedure guarantees that each sample is rather well spread over the population. Lister and Scott (2009) have used space-fillin curves in order to make sure that the sample locations are well spread over the space.

Grafström (2012) and Grafström *et al.* (2012) introduced new sampling methods that enable to select unequal probability samples that are well spread over the population. These methods are respectively called spatially correlated Poisson sampling and local pivotal method. Instead of a mapping, these methods use distance between units to create small joint inclusion probabilities for nearby units, forcing the samples to be well spread. An advantage of spatially correlated Poisson sampling and the local pivotal method is that the use of a distance measure makes it easy to spread the sample in any number of dimensions. Spatially balanced sampling is efficien when there are spatial trends within the population (e.g. Stevens and Olsen, 2004). Indeed, nearby locations or units usually have similarities. These similarities can be due to similar conditions in the environment. In this situation, it is efficien to make sure that the sample is well spread; that is, it is unwise to select nearby units. Spatially balanced sampling is commonly used for natural resources, which often exhibit spatial trends.

In this paper, we propose a method that is doubly balanced in the sense that it enables to select samples that are balanced on a number of auxiliary variables and at the same time are well spread for some variables, which can be topographical coordinates. The implementation supports the use of unequal inclusion probabilities. This new method is motivated by a quite general population model, for which we have good arguments that the method is close to optimality.

2. STRATEGY FOR BALANCED SAMPLING

If the population of interest is generated by a linear model with uncorrelated errors terms, Nedyalkova and Tillé (2009) have shown that the best model-assisted strategy is to firs randomly select a balanced sample with inclusion probabilities proportional to the standard deviations of the errors and then to use the HT estimator of total. When sampling from a territory, the units are often spatially correlated. This can be formalized by means of the following linear model:

$$y_k = \mathbf{x}_k^T \boldsymbol{\beta} + \varepsilon_k, \text{ for all } k \in U \tag{1}$$

where $\mathbf{x}_k$ is a column vector of the values taken by the p auxiliary variables on unit k, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of regression coefficients Moreover, the ε_k are random variables such that $\mathrm{E}_{\boldsymbol{M}}(\varepsilon_k) = 0$, $\mathrm{var}_{\boldsymbol{M}}(\varepsilon_k) = \sigma_k^2$, for all $k \in U$, and

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

where $\mathrm{E}_{\boldsymbol{M}}(\cdot)$, $\mathrm{var}_{\boldsymbol{M}}(\cdot)$ and $\mathrm{cov}_{\boldsymbol{M}}(\cdot)$ respectively denote the expectation, variance and covariance under model (1).

Usually, the closer the units are, the more correlated they are. The $\rho_{k\ell}$ are thus supposed to be decreasing in function of a distance that can be computed between k and ℓ. For instance, the correlations could be written as $\rho_{k\ell} = \rho^{d(k,\ell)}$, where $d(k, \ell)$ is a distance between units k and ℓ.

Let $p(s)$ be a sampling design on the population, S be the random sample with fi ed sample size n, π_k be the first-orde inclusion probability, $\pi_{k\ell}$ be the joint inclusion probability and

$$\widehat{Y} = \sum_{k \in S} \frac{y_k}{\pi_k}$$

be the HT estimator of the total

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

Let $\mathrm{E}_p(\cdot)$ be the expectation under the sampling design, and $\mathrm{E}_{\boldsymbol{M}}(\cdot)$ be the expectation under the model. The random sample S is supposed to be independent from the ε_k.

Our aim is to search for an optimal strategy in such a way as to ensure that the anticipated variance of the HT estimator is as small as possible. Consider the following result:

Proposition 1. *Under model (1), the anticipated variance of the HT estimator can be shown to be*

$$\mathrm{E}_p \mathrm{E}_{\boldsymbol{M}}(\widehat{Y} - Y)^2 = \mathrm{E}_p \left[\left(\sum_{k \in S} \frac{\boldsymbol{x}_k}{\pi_k} - \sum_{k \in U} \boldsymbol{x}_k \right)^T \boldsymbol{\beta} \right]^2 + \sum_{k \in U} \sum_{\ell \in U} \sigma_k \sigma_\ell \rho_{k\ell} \frac{\pi_{k\ell} - \pi_k \pi_\ell}{\pi_k \pi_\ell} \tag{2}$$

The proof is routine and is omitted.

If the sample is balanced on the x-variables, that is,

$$\sum_{k\in S}\frac{\mathbf{x}_k}{\pi_k}=\sum_{k\in U}\mathbf{x}_k \tag{3}$$

then the firs term of expression (2) vanishes. We should thus select a sample that is balanced on the independent variables of the model.

If a balanced sample is selected, the anticipated variance simplifie to

$$\mathrm{E}_p\mathrm{E}_M\left(\widehat{Y}-Y\right)^2=\sum_{k\in U}\sum_{\ell\in U}\sigma_k\sigma_\ell\rho_{k\ell}\frac{\pi_{k\ell}-\pi_k\pi_\ell}{\pi_k\pi_\ell} \tag{4}$$

This expression directly shows that the joint inclusion probabilities $\pi_{k\ell}$ must be chosen as small as possible when $\rho_{k\ell}$ is large. This confirm a well-known rule of spatial sampling that the joint selection of units that are positively correlated must be avoided. As the correlated units are in general geographically close, the sample must be well spread (or spatially balanced) on the territory.

If the sample is balanced on the independent variables and well spread, the diagonal becomes the dominant term of the variance given in Equation (4), that is,

$$\mathrm{E}_p\mathrm{E}_M\left(\widehat{Y}-Y\right)^2\approx\sum_{k\in U}\sigma_k^2\frac{1-\pi_k}{\pi_k} \tag{5}$$

With the constraint that the expected sample size is fi ed, that is,

$$\sum_{k\in U}\pi_k=n$$

and by using a Lagrangian function, we fin that the minimum in π_k of Equation (5) is given by

$$\pi_k=\frac{n\sigma_k}{\sum_{\ell\in U}\sigma_\ell}$$

provided that $n\sigma_k\leqslant\sum_{\ell\in U}\sigma_\ell$.

Thus, under model (1), a very efficien sampling design consists of the following:

1. using a balanced sampling design on the independent variable $\mathbf{x}_k$,
2. avoiding the selection of neighbouring units, that is, selecting a well-spread sample (or spatially balanced), and
3. using inclusion probabilities proportional to σ_k.

In the next sections, we describe a new method that enables us to meet these three requirements.

Notice that the use of the HT estimator under a design with these properties will be efficien if the population is close to a realization from the model but maintains desirable properties such as design unbiasedness and design consistency even if the model is false. The use of the HT estimator thus guarantees the robustness against a mis-specificatio of the model.

3. THE LOCAL PIVOTAL METHOD

The proposed sampling algorithm was developed by combining ideas from the local pivotal method and the cube method. We will start by giving a description of these methods. The local pivotal method (Grafström *et al.*, 2012) is an application of the pivotal method (Deville and Tillé, 1998; Chauvet, 2012) to spatial statistics. In the pivotal method, the inclusion probabilities are successively updated to become inclusion indicators. One step of the algorithm can be described as follows. Choose two units k and ℓ with $0<\pi_k<1$ and $0<\pi_\ell<1$. If $\pi_k+\pi_\ell<1$, then

$$\left(\pi_k',\pi_\ell'\right)=\begin{cases}(0,\pi_k+\pi_\ell) & \text{with probability } \frac{\pi_\ell}{\pi_k+\pi_\ell}\\ (\pi_k+\pi_\ell,0) & \text{with probability } \frac{\pi_k}{\pi_k+\pi_\ell}\end{cases}$$

and if $\pi_k+\pi_\ell\geqslant 1$, then

$$\left(\pi_k',\pi_\ell'\right)=\begin{cases}(1,\pi_k+\pi_\ell-1) & \text{with probability } \frac{1-\pi_\ell}{2-\pi_k-\pi_\ell}\\ (\pi_k+\pi_\ell-1,1) & \text{with probability } \frac{1-\pi_k}{2-\pi_k-\pi_\ell}\end{cases}$$

Now, replace (π_k,π_ℓ) with the updated values $\left(\pi_k',\pi_\ell'\right)$. Repeat the aforementioned step until the outcome is decided for all units. In each update, the two units k and ℓ can be arbitrarily chosen.

The local pivotal method was constructed to give small joint inclusion probabilities for nearby units. This is achieved by, in each step, selecting two nearby units k and ℓ for the update. When π_k and π_ℓ are updated, one is increased as much as possible and one is decreased, while keeping the sum fi ed. This makes the second-order inclusion probability small for units that are simultaneously updated. One of the suggestions given by Grafström *et al.* (2012) is to choose unit k randomly (with equal probabilities) and then choose its nearest neighbour ℓ for each update. Any distance measure can be used to fin the nearest neighbours. The local pivotal method thus avoids the selection of neighbouring units and selects a well-spread sample.

4. THE CUBE METHOD

4.1. Aim of the method

The cube method (Deville and Tillé, 2004; Tillé, 2006, 2011) is a class of sampling algorithms that randomly select a balanced sample in the sense of Equation (3) and exactly satisfy a set of given inclusion probabilities π_k. The method is based on a random transformation of the vector of inclusion probabilities $\boldsymbol{\pi} = (\pi_1 \cdots \pi_N)^T$ until a sample $\mathbf{s}$ is obtained such that

1. the inclusion probabilities are exactly satisfied and
2. the balancing equations given in Equation (3) are satisfie to the furthest extent possible.

The name of the method comes from the geometric representation of a sampling design. Indeed, a sample may be represented by a vector $\mathbf{s} = (I[1 \in s] \cdots I[k \in s] \cdots I[N \in s])'$, where $I[k \in s]$ takes value 1 if $k \in s$ and 0 if not. A sample may thus be viewed as a vertex of an N-cube as showed in Figure 1.

The cube method thus selects a random sample $\mathbf{s}$ such that $\mathrm{E}(\mathbf{s}) = \boldsymbol{\pi}$. This expectation is computed with respect to the sampling design

$$\mathrm{E}(\mathbf{s}) = \sum_{s \in \mathcal{S}} p(\mathbf{s})\mathbf{s} = \boldsymbol{\pi}$$

The balancing equations given in Equation (3) may also be written as

$$\sum_{k \in U} \frac{\mathbf{x}_k}{\pi_k} s_k = \sum_{k \in U} \mathbf{x}_k \text{ with } s_k \in \{0, 1\}, k \in U$$

The balancing equations can be viewed as a linear system in $s_1, \ldots, s_N$ and thus defin an affin subspace in $\mathbb{R}^N$ of dimension $N - p$ denoted by Q, where p is the dimension of $\mathbf{x}_k$.

The problem of selecting a balanced sample may thus be reformulated. A balanced sampling design consists of choosing a vertex of the N-cube (a sample) that remains on the linear subspace Q. Figure 2 shows two examples: the firs one is a constraint of fi ed sample size. The second is an example where the balancing equations cannot be exactly satisfied Indeed, the selection of a sample is an integer-number problem and the balancing equations cannot be exactly satisfie in most of the cases. When it is not possible to select an exactly balanced sample, we say that there is a 'rounding problem'. In this case, the cube method provides a sample that is as balanced as possible.

The cube method (Deville and Tillé, 2004) is divided into two phases: the fligh phase and the landing phase. The fligh phase is a random walk that begins at the vector of inclusion probabilities and remains in the intersection of the cube and the constraint subspace. This random walk stops at a vertex of the intersection of the cube and the constraint subspace. At the end of the fligh phase, if a sample is not obtained, the landing phase consists in selecting a sample that is as close as possible to the constraint subspace.

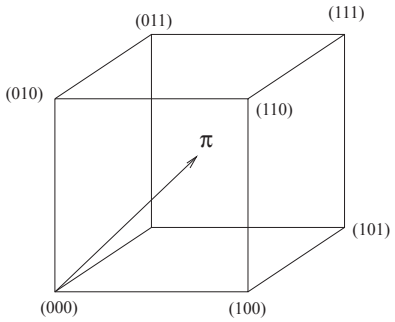

Figure 1. Possible samples in a population of size $N = 3$

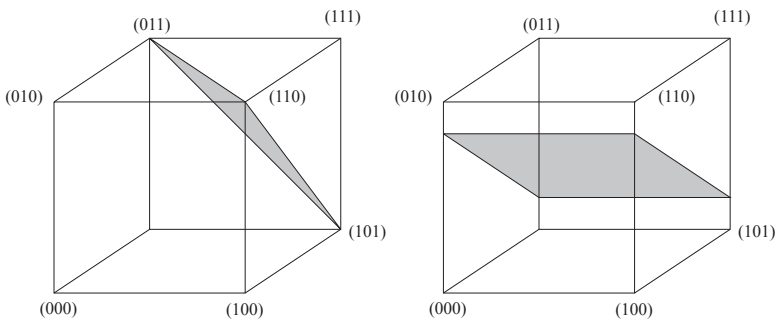

Figure 2. Both examples are in a population of size $N = 3$. The subspace of constraint cross the cube. In the firs one, the constraint is the fi ed sample size $n = 2$. In the second one, the constraint generates a rounding problem

4.2. The fligh phase

The fligh phase is a random walk in the intersection of the balancing subspace and of the cube. This random walk stops at a vertex of the intersection of the cube and the subspace. The fligh phase is a class of procedures define in Algorithm 1.

Algorithm 1 General algorithm of the fligh phase of the cube method.

First initialize with $\boldsymbol{\pi}(0) = \boldsymbol{\pi}$. Next, at time $t = 0, \ldots, T$,

1. Generate any vector $\mathbf{u}(t) = [u_k(t)] \neq 0$ such that
   (i) $\mathbf{u}(t)$ is in the kernel of matrix $\mathbf{A} = (\mathbf{x}_1/\pi_1, \ldots, \mathbf{x}_k/\pi_k, \ldots, \mathbf{x}_N/\pi_N)$), i.e. $\mathbf{Au}$(t)=0,
   (ii) $u_k(t) = 0$ if $\pi_k(t)$ is integer.

2. Compute $\lambda_1^*(t)$ and $\lambda_2^*(t)$, the largest values such that
   $0 \leqslant \pi_k(t) + \lambda_1(t)u_k(t) \leqslant 1, k \in U$
   $0 \leqslant \pi_k(t) - \lambda_2(t)u_k(t) \leqslant 1, k \in U$
3. Compute

$$\boldsymbol{\pi}(t+1) = \begin{cases} \boldsymbol{\pi}(t) + \lambda_1^*(t)\mathbf{u}(t) & \text{with probability } q_1(t) \\ \boldsymbol{\pi}(t) - \lambda_2^*(t)\mathbf{u}(t) & \text{with probability } q_2(t) \end{cases}$$

where $q_1(t) = \lambda_2^*(t)/\{\lambda_1^*(t) + \lambda_2^*(t)\}$ and $q_2(t) = 1 - q_1(t)\}$.

The fligh phase stops when it is no longer possible to fin a vector $\mathbf{u}(t) \neq 0$.

At each step, at least one component of $\boldsymbol{\pi}(t)$ is rounded to 0 or 1. This means that $\boldsymbol{\pi}(t)$ contains at least t values that are equal either to 0 or to 1. Thus, the algorithm cannot run more than N steps. Moreover, from the algorithm, we have

$$\mathrm{E}\left[\boldsymbol{\pi}(t+1)|\boldsymbol{\pi}(t)\right] = \boldsymbol{\pi}(t)$$

and thus by induction, we obtain

$$\mathrm{E}\left[\boldsymbol{\pi}(t)\right] = \boldsymbol{\pi}, \text{ for all } t = 0, 1, 2, \ldots$$

At each step, the inclusion probabilities are thus satisfied At each step, the balancing equations are also satisfied that is,

$$\sum_{k \in U} \frac{\mathbf{x}_k}{\pi_k} \pi_k(t) = \sum \mathbf{x}_k$$

because $\mathbf{u}(t)$ is in the kernel of matrix $\mathbf{A}$.

Chauvet and Tillé (2006) have shown that each step of the fligh phase can also be applied on a subset of the population provided that this subset contains more noninteger values $\pi_k(t)$ than the number p of balancing variables. They use this simple trick to construct a very fast algorithm because the application of the fligh phase on a subset requires much less computation time. This trick will be used below to defin a 'local' cube method.

4.3. Landing phase

Let $\boldsymbol{\pi}^* = \left[\pi_k^*\right]$ be the vector obtained at the last step of the fligh phase. When there is a rounding problem, it is not possible to fin a sample that is exactly balanced. In this case, some components of $\boldsymbol{\pi}$ are not integer. However, it is possible to prove (Deville and Tillé, 2004) that $\mathrm{card}(U^*) \leqslant p$, where $U^* = \left\{k \in U | 0 < \pi_k^* < 1\right\}$ and p is the number of balancing variables. The aim of the landing phase is to fin a sample $\mathbf{s}$ such that $\mathrm{E}(\mathbf{s}|\boldsymbol{\pi}^*) = \boldsymbol{\pi}^*$, which is almost balanced. There are two ways of selecting such a sample:

1. *The flight phase by linear programming* consists of considering all the possible samples of U^*. A cost is assigned to each sample. This cost is, for instance, the distance between the sample and the subspace of constraints. Next, one looks for a sampling design on U^* that minimizes the expected cost and that satisfie the inclusion probabilities $\boldsymbol{\pi}^*$. All the possible samples of U^* must thus be enumerated. This problem can be solved by linear programming because the number of samples to consider is reasonable because of the small size of U^*.
2. *The flight phase by suppression of variables* may be used when the number of balancing variables is too large for the linear program to be solved by a simplex algorithm, $p > 20$. With this method, an auxiliary variable is dropped at the end of the fligh phase. Next, we can return to the fligh phase until it is no longer possible to 'move' within the constraint subspace. The constraints are then relaxed successively according to an order of preference.

There are two SAS® implementations of the cube method available on the Web site of the University of Neuchâtel and on the Web site of the *Institut National de la Statistique et des Études Économiques*. A language R implementation is also available in the 'sampling' package (Tillé and Matei, 2011). The pivotal method can be seen as a particular case of the cube method when the only auxiliary variable is the intercept and when the fast implementation proposed by Chauvet and Tillé (2006) is used.

5. AN ALGORITHM FOR SPREAD AND BALANCED SAMPLING

In this section, we present the new algorithm used to select a sample that is balanced on p auxiliary x-variables and is well spread in some space. Distance between units can be measured in other variables than the x-variables on which we balance the sample. The sampling algorithm is a mixture of the cube method and a generalization of the local pivotal method. The basic idea is to repeatedly apply the fligh phase of the cube method on a cluster of $p+1$ nearby units. When the fligh phase is applied on such a cluster, the sampling outcome is decided for at least one of the units, while respecting the p balancing conditions. Because the updating of the inclusion probabilities is done locally, this procedure gives small joint inclusion probabilities for nearby units. When there are less than $p+1$ units left, for which the sampling outcome is undecided, the sample is finalize by applying the landing phase of the cube method. More precisely the procedure is described in Algorithm 2.

Algorithm 2 Algorithm for spread and balanced sampling.

- $\boldsymbol{\pi}(0)=\boldsymbol{\pi}$, $j=0$
- While there are at least $p+1$ units whose sampling outcome are undecided, i.e. $\#A(j)\geqslant p+1$, where $A(j)=\{k\in U|0<\pi_k(j)<1\}$.
  1. A subset $B(j)$ of $p+1$ neighbouring units is selected from $A(j)$ by means of Algorithm 3.
  2. A fligh phase of the cube method is applied on the $p+1$ selected units. This fligh phase transform in $B(j)$ the $\pi_k(j)$ to $\pi_k(j+1)$ and satisfie

$$\sum_{k\in B(j)}\frac{\mathbf{x}_k}{\pi_k}\pi_k(j+1)=\sum_{k\in B(j)}\frac{\mathbf{x}_k}{\pi_k}\pi_k(j)$$

     Notice that, for this fligh phase, the population of reference is $B(j)$, the balancing variables are $\pi_k(j)\mathbf{x}_k/\pi_k$ and the inclusion probabilities are $\pi_k(j)$. For the units of U that are not in $B(j)$, the values $\pi_k(j)$ remain unchanged, i.e. $\pi_k(j+1)=\pi_k(j)$.
  3. Compute $j=j+1$.
- A landing phase of the cube method is applied.

A cluster of $p+1$ nearby units is selected by Algorithm 3. With this procedure, the sample is as well balanced as with the usual cube method. The sample is also well spread. Indeed, at each step of Algorithm 2, a decision is once and for all taken for a statistical unit. If this statistical unit is taken, the inclusion probabilities of the other units of the cluster are generally decreased because the sum remains unchanged. Likewise, if the decision consists of not taking a unit, the inclusion probabilities of the other units of the cluster are generally increased. So, the method avoids the selection of neighbours. It is difficul to give a formal proof that the samples are well spread in the general case. However, with only an intercept used as auxiliary variable, the new method coincides with the local pivotal method. For that special case, Grafström *et al.* (2012) provided some theoretical results that supports that the resulting samples are very well spread.

Algorithm 3 Cluster selection algorithm.

1. Select with equal probabilities among the undecided units (i.e. from $A(j)$), one unit k randomly and then the p closest units to unit k.
2. Calculate the mean position of the $p+1$ units.
3. Select the nearest $p+1$ units to the mean position.
4. Repeat 2–3 while the sum of squares of the distances of the units of the cluster to their mean is decreasing.

Because the sampling outcome is decided for at least one unit in each step of the algorithm, there are at most $N-p$ steps until the landing phase can be applied and a sample is achieved. By using the R-package 'sampling', which includes functions for applying the fligh phase and the landing phase of the cube method, this algorithm is easily implemented in R. The R code of the new method is available on demand.

6. THE FIRST EXAMPLE

The firs example consists of selecting 400 points among a grid of $40\times 40=1600$ points with equal inclusion probabilities $\pi_k=0.25$, $k=1,\ldots,1600$. This example also introduces the concept of selecting samples that are balanced on the square of the coordinates. This means that the variance of these coordinates will be almost the same in the samples as in the population. We also have included the square of the distance to a set of points as additional balancing variables. This is done to show that it is also possible to use balanced sampling in order to force the samples to be spread.

The balancing variables are the following:

1. const: intercept,
2. coor_x: horizontal coordinate of the point that takes the values $\{1, \ldots, 40\}$,
3. coor_y: vertical coordinate of the point that takes the values $\{1, \ldots, 40\}$,
4. coor_x_2: square of variable coor_x,
5. coor_y_2: square of variable coor_y,
6. dist_10_10: square of the distance of the current point to the point (10,10),
7. dist_10_30: square of the distance of the current point to the point (10,30),
8. dist_30_10: square of the distance of the current point to the point (30,10),
9. dist_30_30: square of the distance of the current point to the point (30,30).

Four sampling designs are applied on the population:

1. Design 1 (spread and balanced): the design is balanced on the nine variables and spread in the coordinates. This is the method developed in this paper.
2. Design 2 (only balanced): the design is balanced by means of the cube method.
3. Design 3 (simple): simple random sampling without replacement.
4. Design 4 (only spread): the sample is only spread and is not balanced. This is the local pivotal method.

Figure 3 contains the representation of four samples selected by means of these sampling designs. The interest of the method is directly visible on Figure 3(a). The sample is well spread. The selection of contiguous units is avoided.

Table 1 shows the relative variances of the balancing variables multiplied by 100 000. These variances are estimated by 5000 simulations.

$$RV_j = 100\,000 \times \mathrm{E}\left(\frac{\widehat{X}_j - X_j}{X_j}\right)^2$$

Table 1 clearly shows that Design 1 (spread and balanced) and Design 2 (only balanced) are better balanced on the auxiliary variables. With respect to simple random sampling, the local pivotal method naturally balances on the coordinates because the samples are well spread. Nevertheless, Designs 1 and 2 are much better. This example shows that the accuracy of a method of spatial balancing (or spreading) can strongly be improved by balancing the sample on auxiliary totals. Design 1 is preferable because it spreads the sample and it balances on the totals of auxiliary variables at the same time.

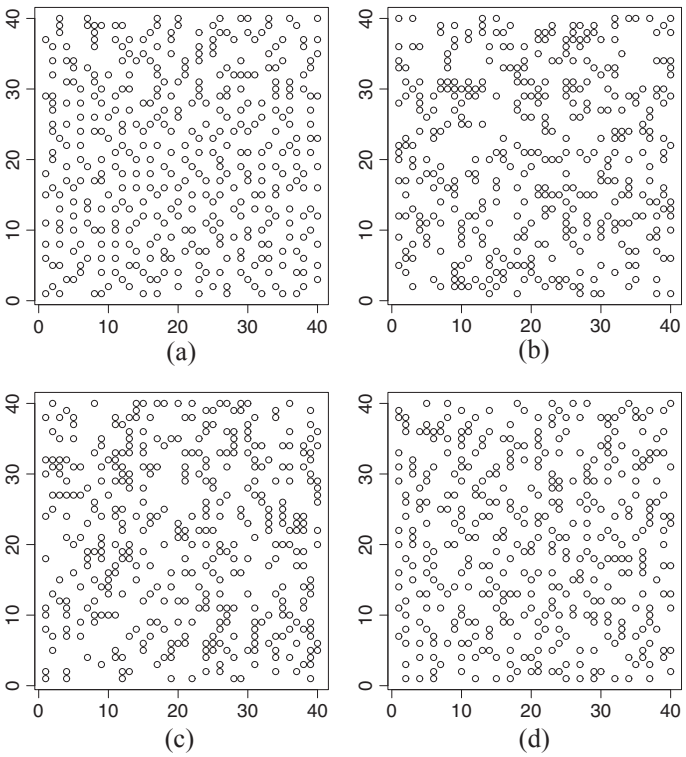

Figure 3. Selection of a sample with (a) Design 1 (spread and balanced), (b) Design 2 (only balanced), (c) Design 3 (simple), and (d) Design 4 (only spread)

Table 1. Relative variances of the balancing variables multiplied by 100 000 estimated by 5000 simulations

	Design 1 Spread and balanced	Design 2 Only balanced	Design 3 Simple	Design 4 Only spread
coor_x	0.030	0.024	5.889	0.252
coor_y	0.030	0.025	5.807	0.077
coor_x_2	0.074	0.058	14.362	0.680
coor_y_2	0.074	0.060	14.095	0.211
dist_10_10	0.021	0.016	3.949	0.055
dist_10_30	0.019	0.016	4.098	0.145
dist_30_10	0.021	0.016	4.086	0.063
dist_30_30	0.018	0.016	4.011	0.132

7. VARIANCE ESTIMATION

With a sampling method that spreads the sample, the selection of neighbouring units is avoided, which means that a large part of the joint inclusion probabilities are null. Under this sampling design, it is thus impossible to estimate the variance of the total estimator without bias. We thus propose an estimator based on a heuristic reasoning by noting that spreading is a kind of local stratification Such estimators are commonly used to take the effect of autocorrelated population into account for systematic sampling (see among others Wolter, 1984; Bellhouse and Sutradhar, 1988). This idea has been generalized by Stevens and Olsen (2003) for spatial sampling that avoids the selection of neighbours. Indeed, they derived a local mean variance estimator for spatially balanced sampling. They used a weighted mean in a neighbourhood of four points, reflectin the GRTS design which always produces well-spread samples.

If the samples are only balanced and not spread on p auxiliary x-variables, then an estimator of variance of $\widehat{Y}$ can be constructed by using the residuals of the regression fi of y by the p balancing variables (Deville and Tillé, 2005). Let e_k be the residuals given by

$$e_k = y_k - \mathbf{x}_k^T \widehat{\boldsymbol{\beta}}$$

where

$$\widehat{\boldsymbol{\beta}} = \left[\sum_{\ell \in S} (1-\pi_\ell) \frac{\mathbf{x}_\ell}{\pi_\ell} \frac{\mathbf{x}_\ell^T}{\pi_\ell} \right]^{-1} \sum_{\ell \in S} (1-\pi_\ell) \frac{\mathbf{x}_\ell}{\pi_\ell} \frac{y_\ell}{\pi_\ell}$$

One of the suggested variance estimators for $\widehat{Y}$ under balanced sampling (the second estimator in Deville and Tillé, 2005) is

$$\widehat{\text{var}}_B \left(\widehat{Y} \right) = \frac{n}{n-p} \sum_{k \in S} (1-\pi_k) \left(\frac{e_k}{\pi_k} \right)^2 \tag{6}$$

Estimator (6) is based on a variance approximation for balanced sampling through the Poisson design, conditioned on $\widehat{\mathbf{X}} = \mathbf{X}$ (Deville and Tillé, 2005, for details). If $\mathbf{x}_k = \pi_k$, Expression (6) gives the Hájek (1981) estimator for unequal probability sampling. If $x_k = n/N$ is the only balancing variable, then Equation (6) corresponds to the usual variance estimator for simple random sampling.

To account for the doubly balanced effect, we simply suggest combining the estimator proposed by Stevens and Olsen (2003) for spread spatial sampling with the estimator proposed by Deville and Tillé (2005) for balanced sampling. We thus suggest the following estimator for samples that are both spread and balanced:

$$\widehat{\text{var}}_{SB} \left(\widehat{Y} \right) = \frac{n}{n-p} \frac{p+1}{p} \sum_{k \in S} (1-\pi_k) \left(\frac{e_k}{\pi_k} - \bar{e}_k \right)^2 \tag{7}$$

where

$$\bar{e}_k = \frac{\sum_{\ell \in G_k} (1-\pi_\ell) \frac{e_\ell}{\pi_\ell}}{\sum_{\ell \in G_k} (1-\pi_\ell)}$$

and G_k is the set of the $p+1$ closest units of k in the sample (including k itself). Thus, $\bar{e}_k$ is a local mean computed in a neighbourhood of k.

8. EXAMPLE WITH THE MEUSE DATA SET

The full 'Meuse' data set is available in the package 'gstat' of the R language. Pebesma (2011) gives the following description: 'This data set gives locations and top soil heavy metal concentrations (ppm), along with a number of soil and landscape variables, collected in a floo plain of the river Meuse, near the village Stein. Heavy metal concentrations are bulk sampled from an area of approximately 15 m x 15 m.'

The following variables are used:

1. x: x-topographical map coordinate,
2. y: y-topographical map coordinate,
3. cadmium: topsoil cadmium concentration,
4. copper: topsoil copper concentration,
5. lead: topsoil lead concentration,
6. zinc: topsoil zinc concentration,
7. elev: relative elevation,
8. om: organic matter, as percentage.

The simulations consist of selecting a sample of size 50 among the 164 locations by using the balancing variables: copper, elev and om to predict the variables zinc, lead and cadmium. Obviously, the sample is also spread across the topographical map coordinates. The variables related to the concentrations in heavy metals are highly correlated. Moreover, there is an important spatial correlation. As shown in Figure 4, there is also a strong heteroscedasticity.

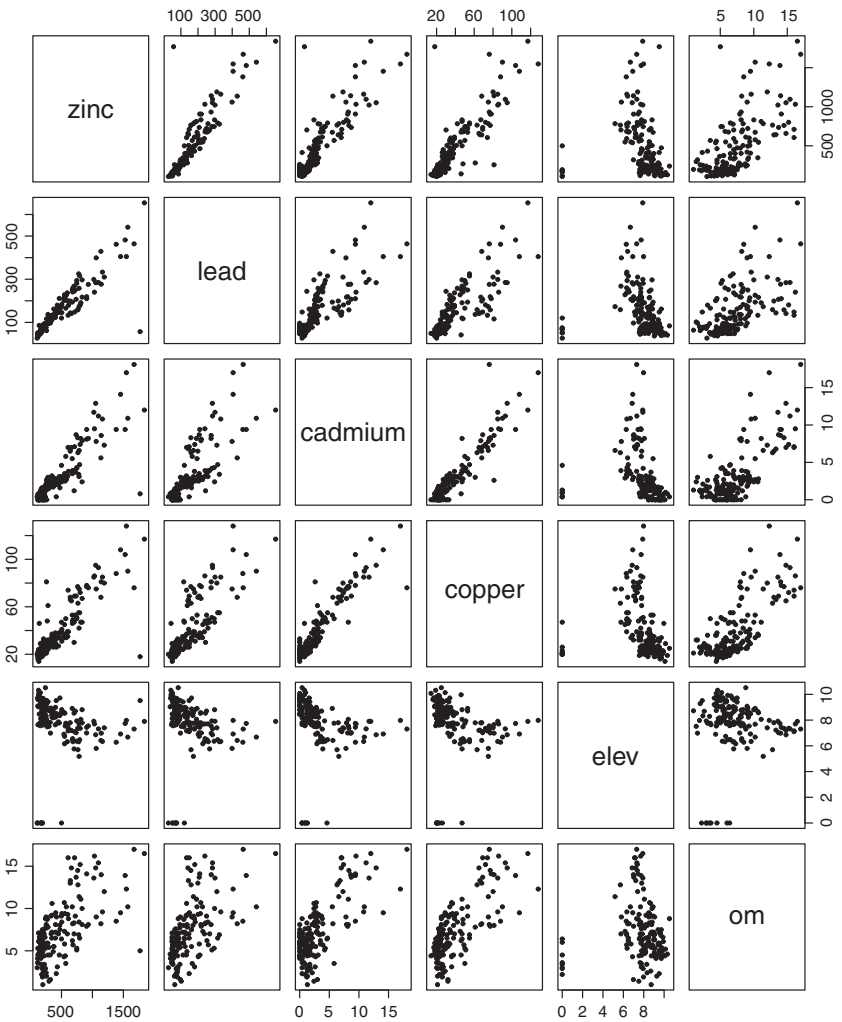

Figure 4. Relations between the balancing variables (copper, elev and om) and the interest variables (zinc, lead and cadmium)

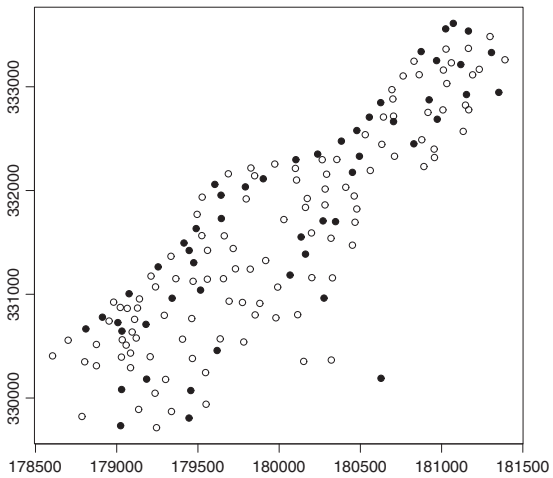

Figure 5. Sample of 50 units selected in the 'Meuse' data set. The sample was selected with unequal probabilities proportional to the copper concentration. The sampling design is spread and balanced. The selected units are the fille circles

Table 2. Results of 10 000 simulations with equal inclusion probabilities with the five sampling designs

	Variance approximated by the simulations		
	zinc	lead	cadmium
Spread and balanced	12 116 860	964 038	715
Cube method	15 386 870	1 594 422	783
Simple random sampling	51 896 830	4 448 526	4438
Local pivotal method	33 179 400	2 640 523	3001
GRTS	37 565 209	2 932 201	3256
	Variance approximated by the simulations in relation to simple random sampling (%)		
	zinc	lead	cadmium
Spread and balanced	23.35	21.67	16.11
Cube method	29.65	35.84	17.64
Simple random sampling	100.00	100.00	100.00
Local pivotal method	63.93	59.36	67.62
GRTS	72.38	65.91	73.37
	Coverage rate of the estimated 95% confidenc interval (%)		
	zinc	lead	cadmium
Spread and balanced	91.67	95.37	89.09
Cube method	90.57	92.40	88.87
Simple random sampling	93.36	93.10	93.39
Local pivotal method	92.44	91.63	92.57
GRTS	89.42	88.90	90.23
	Ratio between averages of the estimated variances and the variances given by the simulations		
	zinc	lead	cadmium
Spread and balanced	1.06	1.14	0.74
Cube method	0.91	0.89	0.75
Simple random sampling	0.99	0.99	1.00
Local pivotal method	0.99	0.91	0.95
GRTS	0.83	0.79	0.84

Five sampling designs are compared:

1. Spread and balanced sampling. This is the method developed in this paper.
2. Balanced sampling by the cube method. In this case, the sample is not spread and the topographical coordinates are not taken into account.
3. Unequal probability sampling without replacement. If the inclusion probabilities are equal, this reduces to simple random sampling without replacement.
4. The local pivotal method. Spread sampling, but the balancing variables are not used.
5. GRTS. Spread sampling, but the balancing variables are not used.

Two sets of inclusions probabilities are used:

1. Equal inclusion probabilities.
2. Unequal inclusion probabilities proportional to the copper concentration.

Figure 5 shows of a sample selected by means of spread and balanced sampling with unequal inclusion probabilities. We ran 10 000 simulations to compare the variance of each sampling design and to check the proposed estimator (7) of variance. The variance of each sampling design was compared with the variance obtained under simple random sampling. The estimator of variance was evaluated by constructing a ratio of the expectation of the estimator of variance on the variance approximated by the simulations. The results are presented in Table 2 for equal probability sampling and in Table 3 for unequal probability sampling.

Table 3. Results of 10 000 simulations with unequal inclusion probabilities with the f ve sampling designs

Variance approximated by the simulations			
	zinc	lead	cadmium
Spread and balanced	19 483 080	501 241	328
Cube method	22 022 460	854 072	400
Unequal probability sampling	21 547 960	901 392	779
Local pivotal method	19 915 120	571 500	601
GRTS	19 502 579	575 623	586
Variance approximated by the simulations in relation to simple random sampling (%)			
	zinc	lead	cadmium
Spread and balanced	37.54	11.27	7.39
Cube method	42.44	19.20	9.01
Unequal probability sampling	41.52	20.26	17.55
Local pivotal method	38.37	12.85	13.54
GRTS	37.58	12.94	13.20
Coverage rate of the estimated 95% confidenc interval (%)			
	zinc	lead	cadmium
Spread and balanced	88.77	96.77	94.31
Cube method	89.70	95.15	93.10
Unequal probability sampling	91.17	94.22	94.30
Local pivotal method	83.65	92.16	90.20
GRTS	83.30	91.70	90.92
Ratio between averages of the estimated variances and the variances given by the simulations			
	zinc	lead	cadmium
Spread and balanced	1.00	1.27	1.10
Cube method	0.94	1.04	0.96
Unequal probability sampling	1.00	1.00	0.99
Local pivotal method	0.88	0.87	0.83
GRTS	0.89	0.86	0.86

The results mainly show that selecting samples that are well spread and balanced decreases the variances. The new method that combines spreading and balanced sampling is always the most accurate one, and the gain in accuracy can be quite high. The use of unequal probability sampling markedly improves accuracy except for the zinc variable. The local pivotal method and the GRTS give very similar results, which is expected because both designs produce well-spread samples. The local pivotal method is however much easier to implement.

The proposed estimator of variance sometimes overestimates the variance and sometimes underestimates it. In particular, with equal inclusion probabilities, the variance of variable 'cadmium' is underestimated for the cube method and the new method. This is probably because the variable 'cadmium' is very well explained by the balancing variables. The gain of accuracy is the most important for 'cadmium'. In this case, the rounding problem of balanced sampling becomes an important part of the variance and is then difficul to catch (Breidt and Chauvet, 2011). The coverage rates of the 95% confidenc intervals are in general less accurate for the methods that uses more auxiliary information than for simple random sampling or unequal probability sampling. This phenomenon is quite typical in statistics. The more accurate estimator a design gives, the more difficul it generally is to estimate the variance. However, the coverage rates of the 95% confidenc intervals show that the proposed estimator of variance leads to a relatively good inference.

9. DISCUSSION

The Meuse example clearly shows that the new method is more efficien than using only balanced sampling because of the remaining spatial trends in residual terms. The new method is also more efficien than a design that only spreads the samples in the topographical space. Because the new method can both spread and balance the samples, it enables one to use more information than other alternatives. Hence, it performs better. The example also shows that spreading and balanced sampling can be efficien even if the relationships between the x-variables and the study variables are not exactly linear. Possibly remaining trends appear in the residuals of the regression model. These trends are neutralized because the sample is well spread.

Even though we justify the method by using a superpopulation model, the inference is based on the sampling design. The HT estimator will be efficien if the population is close to a realization from model (1), but the estimator maintains desirable properties like design unbiasedness and design consistency even if the model is not properly specified

Acknowledgements

The authors are grateful to Lionel Qualité and two reviewers for their useful comments that helped to improve this manuscript.

REFERENCES

Bellhouse DR, Sutradhar BC. 1988. Variance estimation for systematic sampling when autocorrelation is present. *The Statistician* **37**: 327–332.
Breidt FJ, Chauvet G. 2011. Improved variance estimation for balanced samples drawn via the Cube method. *Journal of Statistical Planning and Inference* **141**: 479–487.
Chauvet G. 2012. On a characterization of ordered pivotal sampling. *Bernoulli* **18**: 1320–1340.
Chauvet G, Tillé Y. 2006. A fast algorithm of balanced sampling. *Journal of Computational Statistics* **21**: 9–31.
Deville J-C, Tillé Y. 1998. Unequal probability sampling without replacement through a splitting method. *Biometrika* **85**: 89–101.
Deville J-C, Tillé Y. 2004. Efficien balanced sampling: the cube method. *Biometrika* **91**: 893–912.
Deville J-C, Tillé Y. 2005. Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference* **128**: 569–591.
Grafström A. 2012. Spatially correlated Poisson sampling. *Journal of Statistical Planning and Inference* **142**: 139–147.
Grafström A, Lundström NLP, Schelin L. 2012. Spatially balanced sampling through the Pivotal method. *Biometrics* **68**: 514–520.
Hájek J. 1981. *Sampling from a Finite Population*. Marcel Dekker: New York.
Horvitz DG, Thompson DJ. 1952. A generalization of sampling without replacement from a finit universe. *Journal of the American Statistical Association* **47**: 663–685.
Lister AJ, Scott CT. 2009. Use of space-fillin curves to select sample locations in natural resource monitoring studies. *Environmental Monitoring and Assessment* **149**: 71–80.
Mandallaz D. 2008. *Sampling Techniques for Forest Inventories*. Chapman & Hall/CRC: Boca Raton, FL.
Marker DA, Stevens DL Jr. 2009. Sampling and inference in environmental surveys. In *Sample Surveys: Design, Methods and Applications*, Vol. 29, Handbook of Statist. Elsevier/North-Holland: Amsterdam; 487–512.
Nedyalkova D, Tillé Y. 2009. Optimal sampling and estimation strategies under the linear model. *Biometrika* **95**: 521–537.
Pebesma E. 2011. Reference manual for R-package 'gstat'. Available from: http://cran.r-project.org/web/packages/gstat/gstat.pdf [Accessed on 27/10/2011].
Stevens DL Jr, Olsen AR. 2003. Variance estimation for spatially balanced samples of environmental resources. *Environmetrics* **14**: 593–610.
Stevens DL Jr, Olsen AR. 2004. Spatially balanced sampling of natural resources. *Journal of the American Statistical Association* **99**: 262–278.
Ripley BD. 1981. *Spatial Statistics*. John Wiley & Sons: New York.
Thompson SK. 1992. *Sampling*. Wiley: New York.
Tillé Y. 2006. *Sampling Algorithms*. Springer: New York.
Tillé Y. 2011. Ten years of balanced sampling with the cube method: an appraisal. *Survey Methodology* **3**: 215–226.
Tillé Y, Matei A. 2011. Reference manual for the R-package 'sampling'. Available from: http://cran.r-project.org/web/packages/sampling/sampling.pdf [Accessed on 18/2/2011].
Wolter KM. 1984. An investigation of some estimators of variance for systematic sampling. *Journal of the American Statistical Association* **79**: 781–790.