Spatially Correlated Poisson Sampling

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

Abstract

A new method for sampling from a finite population that is spread in one, two or more dimensions is presented. Weights are used to create strong negative correlations between the inclusion indicators of nearby units. The method can be used to produce unequal probability samples that are well spread over the population in every dimension, without any spatial stratification. Since the method is very general there are numerous possible applications, especially in sampling of natural resources where spatially balanced sampling has proven to be efficient. Two examples show that the method gives better estimates than other commonly used designs.

Keywords: Correlated Poisson Sampling, Generalized Random Tessellation Stratified design, negative correlation, spatial sampling, spatially balanced sampling, unequal probability sampling

1 Introduction

This paper presents a new and very general sampling method that can be used to draw unequal probability samples from a population that is spatially spread in one, two or more dimensions. Many applications of spatial sampling can be found in environmental studies, where the population is distributed over space. Forest inventory is one example, where remote sensing data may be available. Such information can be included in the sampling design to make field inventories more efficient. The proposed method is called Spatially

Correlated Poisson Sampling (SCPS) and is a modification of Correlated Poisson Sampling (CPS), introduced by Bondesson & Thorburn (2008). With SCPS it is possible to select samples that are spatially balanced. The sample locations are then well distributed over the population, which has proven to be efficient for sampling natural resources.

One way to produce unequal probability samples that are well spread over the population is to make a spatial stratification. A problem with spatial stratification is that it is seldom clear how to stratify in a good way. If the inclusion probabilities are equal it is often very efficient to use maximal stratification, i.e. to select one or two units per stratum. It is not straightforward how to generalize this concept to unequal inclusion probabilities. Bondesson & Grafström (2010) extended Sampford's (1967) sampling design to make it possible to coordinate the selection of an unequal probability sample over many small strata when the inclusion probabilities have non-integer sums within strata. If the population is spatially stratified into small strata with inclusion probability sums greater than or equal to 1, then it is possible to make sure that the sample is well spread over the population. However, the stratification is somewhat arbitrary.

In one dimension (units along a line) it is possible to use systematic πps sampling (see e.g. Brewer & Hanif, 1983, pp. 21-22) to produce unequal probability samples that are well spread over the population. This idea was generalized to two dimensions by Stevens & Olsen (2004) who introduced the Generalized Random-Tessellation Stratified (GRTS) design. The idea behind GRTS is to map the two-dimensional locations into one dimension while preserving some spatial order. The sample is then selected in one dimension, using systematic πps sampling, and mapped back into two dimensions.

In order to use the new SCPS-design it is required that some kind of distance can be calculated between units. The focus is on introducing negative correlation between the inclusion indicators of units that are close in distance. In that way units that are close seldom appear simultaneously in the sample. The samples produced will then be well spread over the population. This can be achieved without spatial stratification. Before the design can be described, we need to introduce the sampling situation and some notation. Let $U = \{1, 2, ..., N\}$ be a finite spatial population of N units. The goal is usually to estimate the population total of some study variable with value y_i for unit i. We assume throughout that we want to estimate the total $Y = \sum_{i=1}^{N} y_i$.

If we have access to some auxiliary information, other than the location of the units, that information may be used. One example is when we know the value of a variable, z_i, for each unit i and $z \propto y$ holds approximately. Then units will be sampled with a probability proportional to z, i.e. the inclusion probability π_i for unit i will be chosen as $\pi_i = cz_i$, where c is a positive constant. If such information is not available, sampling is made with equal inclusion probabilities for all units.

When a sample has been selected, the total Y can be estimated by the unbiased Horvitz-Thompson estimator

$$\hat{Y}_{HT} = \sum_{i=1}^{N} \frac{y_i}{\pi_i} I_i, \tag{1}$$

where $I_i, i = 1, 2, ..., N$, are inclusion indicators, i.e. $I_i = 1$ if unit i is in the sample and $I_i = 0$ otherwise. If the sample size is random, then it is often better to use the nearly unbiased Horvitz-Thompson ratio estimator

$$\hat{Y}_R = Z \frac{\sum_{i=1}^{N} \frac{y_i}{\pi_i} I_i}{\sum_{i=1}^{N} \frac{z_i}{\pi_i} I_i}, \tag{2}$$

where $Z = \sum_{i=1}^{N} z_i$. If $\pi_i = cz_i$, then $\hat{Y}_R$ simplifies to

$$\hat{Y}_R = \frac{Z}{|\mathbf{I}|} \sum_{i=1}^{N} \frac{y_i}{z_i} I_i,$$

where $|\mathbf{I}| = \sum_{i=1}^{N} I_i$ is the sample size. If the sample size is fixed and $\pi_i = cz_i$, then (2) and (1) are identical.

In section 2, the CPS-design is described. Then, in section 3, the generalization to spatial sampling is presented. Two simulation examples are provided in section 4 and variance estimation is discussed in section 5. Concluding comments are given in section 6.

2 Correlated Poisson Sampling

Correlated Poisson Sampling was introduced by Bondesson & Thorburn (2008) as a method suitable for real time sampling with unequal inclusion probabilities. In real time sampling the units of the population are visited by the sampler one by one in some order. The sampler must decide at the visit whether or not the unit should be sampled. There is no possibility to re-visit units at a later time. The method can be used to create correlations as desired between the inclusion indicators. In real time sampling it is often good to have negative correlation between the inclusion indicators of units that are close in the order they are visited. The method is very flexible. In fact, every without replacement design with prescribed inclusion probabilities can be implemented by CPS, cf. Bondesson & Thorburn (2008). The method is list sequential, i.e. it is applied to a list of units and the sampling outcome is first decided for unit 1, then for unit 2 etc. After each sampling decision, the inclusion probabilities for the remaining units in the list are updated according to a specific updating rule.

Each unit i has a prescribed inclusion probability π_i, $i = 1, 2, ..., N$, with $\sum_{i=1}^{N} \pi_i = n$. Thus the expected sample size is n. We will assume that n is an integer, but that is not required for CPS to work. However it simplifies comparisons between different approaches. The method works as follows. First unit 1 is included with probability $\pi_1^{(0)} = \pi_1$. If unit 1 was included, we set $I_1 = 1$ and otherwise $I_1 = 0$. Generally at step j, when the values for $I_1, ..., I_{j-1}$ have been recorded, unit j is included with probability $\pi_j^{(j-1)}$. Then the inclusion probabilities are updated for the units $i = j+1, ..., N$, according to

$$\pi_i^{(j)} = \pi_i^{(j-1)} - (I_j - \pi_j^{(j-1)}) w_j^{(i)}, \tag{3}$$

where $w_j^{(i)}$ are weights given by unit j to the units $i = j+1, j+2, ..., N$ and $\pi_i^{(0)} = \pi_i$. For unit j it is convenient to let $\pi_j^{(k)} = I_j$ for $k \geq j$. The updating can then be illustrated as

$$\begin{array}{rcccccc}
\boldsymbol{\pi}^{(0)}: & \pi_1 & \pi_2 & \pi_3 & \pi_4 & \cdots & \pi_N \\
\boldsymbol{\pi}^{(1)}: & I_1 & \pi_2^{(1)} & \pi_3^{(1)} & \pi_4^{(1)} & \cdots & \pi_N^{(1)} \\
\boldsymbol{\pi}^{(2)}: & I_1 & I_2 & \pi_3^{(2)} & \pi_4^{(2)} & \cdots & \pi_N^{(2)} \\
\boldsymbol{\pi}^{(3)}: & I_1 & I_2 & I_3 & \pi_4^{(3)} & \cdots & \pi_N^{(3)} \\
\vdots\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\boldsymbol{\pi}^{(N)}: & I_1 & I_2 & I_3 & I_4 & \cdots & I_N
\end{array} .$$

Thus we gradually update the inclusion probability vector in N steps, until it becomes the vector of inclusion indicators.

2.1 The weights

Since every without replacement design with prescribed inclusion probabilities can be implemented by CPS, it all comes down to finding weights that give the desired design or properties. The weight $w_j^{(i)}$, $j < i$, determines how the inclusion probability for unit i should be affected by the sampling outcome for unit j. More precisely, the weight $w_j^{(i)}$, $j < i$, may depend on the previous sampling outcome $I_1, I_2, ..., I_{j-1}$ but not on the future outcomes $I_j, I_{j+1}, ..., I_N$. From (3), we see that the weights should also satisfy the following restrictions

$$-\min\left(\frac{1-\pi_i^{(j-1)}}{1-\pi_j^{(j-1)}}, \frac{\pi_i^{(j-1)}}{\pi_j^{(j-1)}}\right) \leq w_j^{(i)} \leq \min\left(\frac{\pi_i^{(j-1)}}{1-\pi_j^{(j-1)}}, \frac{1-\pi_i^{(j-1)}}{\pi_j^{(j-1)}}\right) \tag{4}$$

in order for $0 \leq \pi_i^{(j)} \leq 1$, $i = j+1, j+2, ..., N$, to hold. Besides these restrictions the weights may be chosen freely. The unconditional inclusion probabilities are not affected by the weights since

$$E(\pi_i^{(i-1)}) = E(E(\pi_i^{(i-1)} \mid \pi_i^{(i-2)})) = E(\pi_i^{(i-2)}) = \cdots = \pi_i.$$

Thus the method always gives the prescribed inclusion probabilities $\pi_i, i = 1, 2, ..., N$. As can be seen from the restrictions (4) it is possible to have negative weights. Usually positive weights give negative correlations between the inclusion indicators and negative weights give positive correlation. In this paper, the focus is on positive weights.

Bondesson & Thorburn (2008) gave some examples on how to choose the weights to get some specific designs. They also derived a general expression for the weights needed to implement any specific design. However, such weights are in general not practical to calculate.

Another important result about the weights, given by Bondesson & Thorburn (2008), is that if the weights $w_j^{(i)}$, $j < i$, sum to 1 for each $j < N$, then a fixed sample size is obtained if the π_is sum to an integer. Since a fixed sample size often is desirable, we try to construct weights that sum to 1.

Different strategies for choosing the weights for the real time sampling situation were investigated by Grafström (2010a). Those strategies depended solely on the order of the units in the list. More weight was given to units that followed close in the order. For some different sampling situations, those strategies gave very promising results in terms of low variance for the Horvitz-Thompson ratio estimator. The gain was the greatest when there existed some trend in the ratios y_i/z_i over the ordered population.

2.2 The probability function

A sampling design and all its properties are given by its probability function. The probability function may be of theoretical interest and can give a better understanding of the sampling method. A random sample is described by the vector of inclusion indicators $\mathbf{I} = (I_1, I_2, ..., I_N)$ and a sample, which is the outcome of $\mathbf{I}$ is here denoted by $\mathbf{x}$. By using the fact that the implementation is list sequential and that the updated inclusion probabilities are known for a generated sample $\mathbf{x}$, the probability function for CPS can be written as

$$\Pr(\mathbf{I} = \mathbf{x}) = \prod_{i=1}^{N} \left(\pi_i^{(i-1)}\right)^{x_i} \left(1 - \pi_i^{(i-1)}\right)^{1-x_i}, \quad \mathbf{x} \in \{0,1\}^N, \tag{5}$$

cf. Grafström (2010b). Notice that $\pi_i^{(i-1)}$ is a function of the first $i-1$ components of $\mathbf{x}$ and can be recursively calculated by using the updating rule (3).

3 Spatially Correlated Poisson Sampling

It is possible to adapt the CPS method by introducing a distance between units, rather than just having an ordered list of the units. The new method, SCPS, works exactly as CPS and it is applied to an ordered list of units. The only additional requirement is that the distance between all units must be known. The SCPS-design is a set of strategies for choosing weights for CPS. We need a distance function $d(i,j)$ that gives a distance between units i and j. The distance $d(i,j)$ may be the Euclidean distance or some other general distance measure. In some cases it may suffice to rank the distances to determine which unit is the closest, the next closest etc. Two different strategies for choosing the weights are given here.

3.1 Maximal weights

One approach to choose weights is that unit j first gives as much weight as possible to the closest unit (in distance) among the units $i = j+1, j+2, ..., N$, then as much weight as possible to the second closest unit etc. with the restriction that the weights sum to 1. This strategy is called the maximal weight strategy. If distances can be equal, then the weight is distributed equal on those units that have equal distance if possible. The first priority is that weight is not put on a unit if it is possible to put the weight on a closer unit. The maximal weight strategy always produces samples of fixed size if the inclusion probabilities sum to an integer.

Example 1. A small population of size $N = 4$ is used to illustrate how the weights are distributed with the maximal weight strategy. The units are shown in Figure 1. All distances between the units are equal to 1 except for the distance between unit 2 and 4, which is 1.73. Let $\pi_1 = 0.8$, $\pi_2 = 0.2$, $\pi_3 = 0.8$ and $\pi_4 = 0.2$, which gives a sample size of $n = 2$. The visiting order is chosen to be $1, 2, 3, 4$ and thus the sampling starts with unit 1, which has 3 closest units. According to the restrictions (4) on the weights, unit 1 can give a maximal weight of 1 to unit 2, a maximal weight of 0.25 to unit 3 and a maximal weight of 1 to unit 4. Now, we always begin to give weight to the unit with the smallest maximal weight among units at equal distance. We try to distribute the weight as equal as possible, which means that we try to give unit 3 weight 1/3. Since it is not possible to give unit 3 weight 1/3, we give it the maximal weight 0.25. Unit 1 now has a total weight of 0.75 left to give to the units 2 and 4. Since the maximal weights for unit 2 and unit 4 are equal we can start with unit 2. Then we try to give unit 2 the weight $0.75/2 = 0.375$, which is possible. The remaining weight 0.375 is put on unit 4. We set $I_1 = 1$ with probability 0.8 and $I_1 = 0$ with probability 0.2. If we assume that $I_1 = 1$, then we get the following new probabilities $\pi_2^{(1)} = 0.125$, $\pi_3^{(1)} = 0.75$ and $\pi_4^{(1)} = 0.125$. Since the outcome is decided for unit 1 it is not considered further and we move on to unit 2, which is easier to handle since there are no units at equal distance any more. The closest unit to unit 2 is now unit 3 and the maximal weight to give unit 3 is $0.75/0.875$ so it receives that weight. Unit 4 receives the remaining weight $0.125/0.875$. We set $I_2 = 1$ with probability $\pi_2^{(1)} = 0.125$ and $I_2 = 0$ with probability 0.875. If the outcome is $I_2 = 1$ the sampling is finished since we would have $\pi_3^{(2)} = 0$ and $\pi_4^{(2)} = 0$. If the outcome is $I_2 = 0$, the procedure continues similarly.

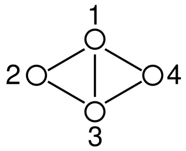

Figure 1: A small population used to illustrate the maximal weight strategy.

A property of the maximal weight strategy is that it balances the sample size locally, like a form of loose stratification without strict borders. For any subset A of all units within some subregion, the number of selected units $\sum_{i \in A} I_i$ will be close to the sum of the inclusion probabilities $\sum_{i \in A} \pi_i$. In a sense it is not possible to make the local sample sizes more balanced. In some cases stratification, with fixed sample size within strata, automatically appear as the following theorem shows.

Theorem 1. *Let the population consist of two separated regions A and B, such that the within region distances are always less than the distance between units in different regions. If $\sum_{i \in A} \pi_i = n_A$ and $\sum_{i \in B} \pi_i = n_B$, where n_A and n_B are positive integers, then the maximal weight strategy produces samples of fixed sizes, n_A and n_B respectively. The sample sizes will be fixed regardless of the order of the sampling.*

Proof. Unit j (in the arbitrary given order) must distribute weights that sum to 1 in order to keep the sum of the updated inclusion probabilities constant. By construction of the maximal weight strategy, unit j will first put weight on the closest unit among the units $j+1, ..., N$. Hence the weight is put on units in the same region first. It remains to show that it is always possible to distribute weights with sum 1 within the same region as unit j. If we assume the opposite, that no such solution is possible with maximal weights, then it is not possible to select a fixed size sample from this region and respect the inclusion probabilities using any other design. Since we know that if e.g. $\sum_{i \in A} \pi_i = n_A$ and n_A is integer it is always possible to select a fixed size sample that respect the inclusion probabilities, by using e.g. systematic πps design, a solution exist for the maximal weight strategy as well. □

Remark 1. Theorem 1 can be directly extended to any number of similarly separated regions with integer valued inclusion probability sums.

3.2 Gaussian preliminary weights

Another strategy is to choose preliminary weights with sum 1 that are controlled by a Gaussian distribution centred at the position of unit j. The weights may then be chosen as

$$w_j^{(i)*} \propto \exp\left(-\left(d(i,j)/\sigma\right)^2\right), \quad i = j+1, j+2, ..., N. \tag{6}$$

Here σ is a parameter that can be used to control the spread of the weights. Evidently more weight is put on units that are close and all units on the same distance receive the same preliminary weight. How to choose the scaling parameter σ depends on the distance between units. One option is to choose σ as the average (or median) of the distances between each unit and its closest neighbour. The weights are called preliminary since the restrictions, (4), may interfere and cut off some weights. Thus the weights that are used may be smaller than the preliminary weights. This fact will generally cause a very small variation in the sample size.

Remark 2. For both maximal weights and Gaussian preliminary weights, the order of the units in the list does affect the design. If the order is changed, then the second order inclusion probabilities may be changed. Nonetheless the design has the same general properties no matter how the units are ordered since the weights depend on distances. Thus it is not so important how the units are ordered. The samples will always be well spread over the population, so the efficiency of the method is practically independent of the ordering.

4 Simulation studies

In this section the SCPS-design is tested in different sampling situations by use of simulation. The approach of Voronoi polygons, suggested by Stevens & Olsen (2004), is used to compare spatial balance. We assume that $n = \sum_{i=1}^{N} \pi_i$ is a positive integer. For a sample of units $\{u_1, u_2, ..., u_n\}$, the Voronoi polygon for the sample unit u_i include all population units closer to u_i than to any other sample unit u_j. See Figure 2 for an example. Let v_i be the sum of the inclusion probabilities of all units in the ith Voronoi polygon. If a population unit has equal distance to two or more sample units, then it is included in more than one polygon. The inclusion probability of that unit is then divided equally to each polygon it is included in.

For a randomly chosen sample unit u_i, we have $E(v_i) = 1$ since there are n units in the sample and $\sum_{i=1}^{n} v_i = \sum_{i=1}^{N} \pi_i = n$. For a spatially balanced sample, all the v_is should be close to 1. Let $\mathbf{v} = (v_1, v_2, ..., v_n)$. Then the variance

$$\text{var}(\mathbf{v}) = \frac{1}{n}\sum_{i=1}^{n} (v_i - 1)^2 \tag{7}$$

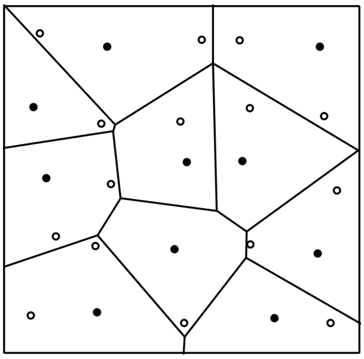

Figure 2: Example of Voronoi polygons. The locations of the sample units used to construct the Voronoi polygons are marked with black dots and the locations of the non-sampled units are marked with circles. For a balanced sample there should be approximately equal amount of probability mass in each polygon.

is used as a measure of spatial balance for a sample. To compare different sampling designs, the mean of (7) is computed over many repeated samples.

Example 2. In this example we consider a 10x10 grid where each cell has length and width 1 and correspond to one unit. We wish to estimate the total of a variable y that has value y_i for unit i. The units in row 1 are labelled 1-10 consecutively, and the units in row 2 are labelled 11-20 etc. The x-position for unit i is given by its column number and the y-position by its row number. The auxiliary information z is given by

$$z = \begin{pmatrix} 1 & 1 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\ 1 & 1 & 2 & 3 & 3 & 4 & 4 & 4 & 3 & 3 \\ 1 & 2 & 3 & 4 & 4 & 5 & 5 & 5 & 4 & 4 \\ 2 & 3 & 4 & 5 & 6 & 7 & 7 & 7 & 6 & 5 \\ 2 & 3 & 4 & 6 & 7 & 8 & 9 & 8 & 7 & 6 \\ 2 & 4 & 5 & 7 & 8 & 9 & 10 & 9 & 8 & 7 \\ 2 & 4 & 5 & 7 & 9 & 10 & 10 & 10 & 9 & 7 \\ 2 & 4 & 5 & 7 & 8 & 9 & 10 & 9 & 8 & 7 \\ 2 & 3 & 4 & 6 & 7 & 8 & 9 & 8 & 7 & 6 \\ 2 & 3 & 4 & 5 & 6 & 7 & 7 & 7 & 6 & 5 \end{pmatrix}.$$

There is a strong spatial trend in z with a peak at z_{67} (row 7, column 7). The auxiliary information may be the result of remote sensing and correspond to average intensity for each cell. To also have a spatial trend in the ratios, we have assumed the following relationship for the ratios y_i/z_i

$$y_i/z_i = 1.1 - 0.2 \times \frac{z_i - \min_i(z_i)}{\max_i(z_i) - \min_i(z_i)}.$$

Then y_i/z_i has a perfect linear trend in z_i and hence also a trend in space. The ratios vary from 1.1 for the smallest z_i to 0.9 for the largest z_i. This corresponds to a situation where small intensities tend to overestimate the target variable and larger intensities tend to underestimate the target variable, which is common with remote sensing data. Of course, an error term can be added to the ratios to make the situation more realistic, but adding noise makes it more difficult to see which methods really can capture the spatial trends.

From the population of size $N = 100$ we shall sample $n = 25$ units. The inclusion probabilities are chosen as $\pi_i = cz_i$ with $\sum_{i=1}^{N} \pi_i = n$ for the unequal probability designs. The following sampling designs are compared.

- SCPS with maximal weights (SCPS 1), which gives spatially balanced samples of fixed size with prescribed inclusion probabilities.

- SCPS with weights chosen as (6) with Euclidean distance and $\sigma = 1$ (SCPS 2), which gives spatially balanced samples of random size with prescribed inclusion probabilities. The variation in the sample size is generally very small.

- Generalized Random-Tessellation Stratified (GRTS), which gives spatially balanced samples of fixed size with prescribed inclusion probabilities.

- Maximum stratification (MSTRAT) with selection of one unit per stratum with equal inclusion probabilities. The grid is stratified into 25 strata of size 2x2 and one unit is selected with simple random sampling within each stratum.

- Sampford sampling (SAMPF), which gives samples of fixed size and prescribed inclusion probabilities but does not produce spatially balanced samples.

- Poisson sampling (POISS), which gives samples of varying size and prescribed inclusion probabilities but does not produce spatially balanced samples.

- Simple random sampling (SRS) with equal inclusion probabilities, which does not produce spatially balanced samples.

For all designs, the HT-ratio estimator (2) is used. For SRS and MSTRAT the estimator is used with equal inclusion probabilities ($\pi_i = 1/4$). Hence the strong auxiliary information is utilized even for SRS and MSTRAT, but in the estimation stage instead of being directly included in the sampling design.

A total of $n(s) = 1000$ samples were selected with each design and the mean square error (MSE) was estimated as

$$MSE_{Sim}(\hat{Y}) = \frac{1}{n(s)} \sum_{s} (\hat{Y}(s) - Y)^2. \tag{8}$$

The results for this simulation are found in Table 1 and as we can see, the best result is obtained for SCPS 1 followed by SCPS 2. A bit surprising is that MSTRAT is more efficient than SAMPF since MSTRAT does not use unequal probabilities. Thus it is more important to spread the sample over space than to use unequal inclusion probabilities and ignore spatial trends. The GRTS design performs well and is better than MSTRAT as expected. We notice that the HT-ratio estimator seems to be slightly biased under the MSTRAT design.

Example 3. In this example a population of size $N = 20$ is used. The population can be seen in Figure 3 and the details are listed in Appendix. Before the sampling is performed we have only information about the location of the units. Thus sampling is made with equal inclusion probabilities. It is not clear how to stratify on location for this population and hence SCPS will be compared to GRTS and SRS. For SCPS we use maximal weights (SCPS 1) and Gaussian preliminary weights with $\sigma = 1$ (SCPS 2). Samples of size $n = 8$ are selected and we want to estimate the total of a variable y that has value y_i for unit i. A total of $n(s) = 1000$ samples were selected and the MSE was estimated by (8). The estimator used is the HT-ratio estimator (2) with $\pi_i = 8/20$ and $z_i = 1$ for all units. The results are listed in Table 2. We see that SCPS 1 is the most efficient design followed closely by SCPS 2. Both these designs succeed in capturing the spatial trend. The GRTS design is more efficient than the SRS design, but not as efficient as SCPS.

Table 1: Simulation results for Example 2. The true total is $Y = 505.59$. The presented measure of spatial balance is the mean of (7) over all simulated samples. A low value indicates a high degree of spatial balance. For MSTRAT and SRS, (7) is calculated with equal inclusion probabilities and the spatial balance of these designs should not be directly compared to the unequal probability designs.

Design	mean($\hat{Y}$)	$MSE_{Sim}(\hat{Y})$	Spatial balance
SCPS 1	505.70	2.7	0.062
SCPS 2	505.65	4.7	0.071
GRTS	505.53	6.8	0.096
MSTRAT	508.61	15.3	0.063
SAMPF	505.61	22.1	0.219
POISS	505.37	23.0	0.243
SRS	506.19	26.1	0.238

Remark 3. Of course, a computer program is needed to select SCPS samples. All designs were implemented in the R statistical programming language, except for the GRTS design for which the R package spsurvey (Kincaid, 2009) was used. SCPS can be used for fairly large populations. Selecting a sample from a population of size 1000 takes less than 1 second and selecting a sample from a population of size 10000 takes about 30 seconds (on a Dell Latitude E6410). A population size of 100000 is also feasible and selecting a sample takes about one hour. For really large populations, a rough initial spatial stratification is recommended to have feasible population sizes. Such a stratification does not significantly affect the overall spatial balance.

5 Variance estimation

Variance estimation can be a bit tricky for SCPS in general. If the maximal weight strategy is used, then many of the second order inclusion probabilities will be zero. This makes it impossible to make a design-based unbiased estimator of the variance. For other weight strategies, that spread the weights more, it is possible to approximate the second order inclusion probabilities and obtain an approximately unbiased design-based variance estimator cf. Bondesson & Thorburn (2008). However, if the weights are more spread

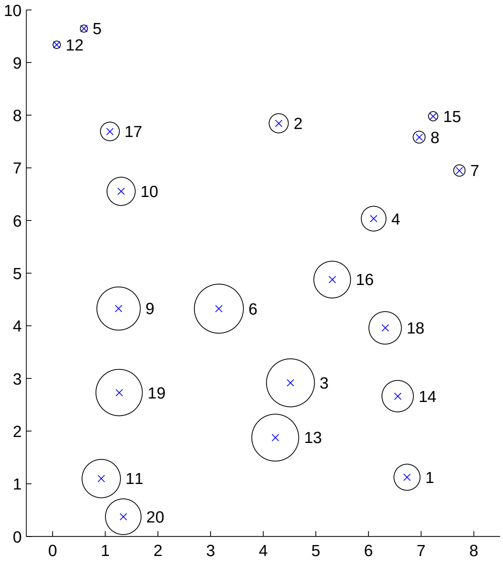

Figure 3: The spatial population used in Example 3. The centre of each circle marks the position of the units and the area of the circles correspond to the study variable y, which has strong spatial trends. The labels of the units are shown to the right of each unit.

out, the samples will be less spatially balanced and the estimator will be less efficient.

It is a common problem for spatially balanced sampling that the second order inclusion probabilities may be zero or very close to zero for units (or points) that are close in distance. There are different solutions to this problem. One may use an estimator that overestimates the variance. Such an estimator can be constructed by pretending that the sample has been selected by another design with non-zero second order inclusion probabilities. In such a situation it is hard to say how much the variance is overestimated, but at least it is possible to get a conservative variance estimate.

Stevens & Olsen (2003) introduced a local neighbourhood variance estimator for the GRTS design that seemed to produce good variance estimates. It may be possible to develop a similar variance estimator for the SCPS design, or perhaps use the same estimator. Finding a good variance estimator

Table 2: Simulation results for Example 3. The true total is $Y = 5.79$. The presented measure of spatial balance is the mean of (7) over all simulated samples. A low value indicates a high degree of spatial balance.

Design	mean($\hat{Y}$)	$MSE_{Sim}(\hat{Y})$	Spatial balance
SCPS 1	5.79	0.47	0.134
SCPS 2	5.77	0.54	0.134
GRTS	5.80	0.79	0.179
SRS	5.79	1.66	0.306

requires some further work and such an estimator needs to be evaluated in different settings. This may be the topic of a subsequent paper.

6 Final comments

It is interesting that there is a noticeable difference between GRTS and SCPS and the reason is that SCPS samples are even more balanced than GRTS samples. The mapping used in GRTS is not perfect in the sense that units that are close in distance may be mapped rather far apart in the one dimensional space where the sampling is made. Nonetheless GRTS produces samples that are much more evenly distributed over space than an ordinary unequal probability design, such as the Sampford design, does. The Sampford design has high entropy, which means that the probability mass is well distributed on all possible samples of the given size. In spatial sampling, high entropy is in general not a good property, it is better to focus the probability mass on samples that are well spread. The main strength of SCPS lies in the ability to use unequal inclusion probabilities and that the method produces samples that are well spread over the population. The latter part is very important if the relationship between the auxiliary variable z and the study variable y vary over space, i.e. if there exists some trend in the ratios y_i/z_i over space. If the design is used with equal inclusion probabilities, then it is good in cases where the study variable has trends over space. A big advantage with SCPS is that it does not require a spatial stratification. In order to make good spatial stratification some knowledge is needed about the spatial trends, otherwise there is a risk that the within stratum variances are greater than the population variance. With SCPS there is no such risk as long as nearby units tend to be more similar than units further apart. SCPS is rather easy to implement and gives samples well spread over the population for any number of dimensions, and can be used with equal or unequal inclusion probabilities. Two different weight strategies have been presented. Maximal weights that seem to always produce the most efficient estimator in situations where nearby units have similar values. With Gaussian preliminary weights it is possible to adjust how much to spread out the weights. Gaussian preliminary weights are a bit easier to implement. Of course, the weights can be chosen in other ways than described here. SCPS can then, at least in theory, implement any other spatial without replacement sampling design with prescribed inclusion probabilities.

Acknowledgements

The Author is thankful to Lennart Bondesson, Göran Ståhl and Anna Hedström Ringvall for valuable comments. The author also thanks two anonymous reviewers for their valuable comments and suggestions that improved this paper.

References

Bondesson, L. & Thorburn, D. (2008). A list sequential sampling method suitable for real-time sampling. *Scand. J. Statistics.* **35**, 466-483

Bondesson, L. & Grafström, A. (2010). An extension of Sampford's method for unequal probability sampling. *Scand. J. Statist.* **38**, 377-392.

Brewer, K.R.W, & Hanif, M. (1983). *Sampling with unequal probabilities.* Lecture notes in statistics, Springer-Verlag, New York.

Grafström, A. (2010a). On a generalization of Poisson sampling. *J. Statist. Plann. Inference* **140**, 982-991.

Grafström, A. (2010b). Entropy of unequal probability sampling designs. *Statist. Methodol.* **7**, 84-97.

Kincaid, T. (2009). User guide for spsurvey, version 2.0, probability survey design and analysis. Retrieved 18/2/2011, from http://www.epa.gov/nheerl/arm/analysispages/software.htm.

Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika* **54**, 499-513.

Stevens, D.L. Jr. & Olsen A.R. (2003). Variance estimation for spatially balanced samples of environmental resources. *Environmetrics* **14**, 593-610.

Stevens, D.L. Jr. & Olsen A.R. (2004). Spatially balanced sampling of natural resources. *Journal of the American Statistical Association* **99**, 262-278.

A Details for the population used in Example 3

Table 3: The population used in Example 3

Unit	x-position	y-position	Target variable (y)
1	6.73	1.12	0.19
2	4.30	7.84	0.11
3	4.52	2.92	0.65
4	6.10	6.04	0.17
5	0.59	9.64	0.01
6	3.16	4.32	0.68
7	7.73	6.95	0.04
8	6.96	7.58	0.04
9	1.25	4.33	0.53
10	1.30	6.55	0.23
11	0.92	1.10	0.42
12	0.08	9.34	0.02
13	4.23	1.87	0.63
14	6.56	2.66	0.28
15	7.23	7.98	0.03
16	5.31	4.88	0.39
17	1.09	7.69	0.10
18	6.32	3.96	0.30
19	1.26	2.73	0.61
20	1.34	0.37	0.36