Open Journal of Statistics, **2013, 3, 36-41**
http://dx.doi.org/10.4236/ojs.2013.31005 Published Online February 2013 (http://www.scirp.org/journal/ojs)

Why Well Spread Probability Samples Are Balanced

Anton Grafström[1], Niklas L. P. Lundström[2]
[1]Department of Forest Resource Management, Swedish University of Agricultural Sciences, Umeå, Sweden
[2]Department of Mathematics and Mathematical Statistics, Umeå University, Umeå, Sweden
Email: anton.grafstrom@slu.se, niklas.lundstrom@math.umu.se

Received November 22, 2012; revised December 24, 2012; accepted January 8, 2013

ABSTRACT

When sampling from a finite population there is often auxiliary information available on unit level. Such information can be used to improve the estimation of the target parameter. We show that probability samples that are well spread in the auxiliary space are balanced, or approximately balanced, on the auxiliary variables. A consequence of this balancing effect is that the Horvitz-Thompson estimator will be a very good estimator for any target variable that can be well approximated by a Lipschitz continuous function of the auxiliary variables. Hence we give a theoretical motivation for use of well spread probability samples. Our conclusions imply that well spread samples, combined with the Horvitz-Thompson estimator, is a good strategy in a varsity of situations.

Keywords: Balanced Sample; Local Pivotal Method; Spatial Balance; Spatially Correlated Poisson Sampling; Voronoi Polytopes

1. Introduction

In many fields there has been a great interest in selecting samples that are well spread or spatially balanced. Such samples are considered to produce good estimates for target variables that exhibit spatial trends, see e.g. [1,2]. The focus in this paper is to explain the connection between a well spread sample and a balanced sample. Roughly speaking, a sample is well spread if the number of selected units is close to what is expected on average, in every part of the auxiliary space. A sample is balanced on a variable if the Horvitz-Thompson (HT) estimator of the total of that variable agree exactly with the known population total of the variable. In fact, with a short analysis, this paper clarifies why a well spread sample is approximately balanced. We also explain that, if the sample is well spread, the variance of commonly used estimators is usually low.

It is well known that samples that are balanced or approximately balanced on the auxiliary variables may be selected by using the cube method, see e.g. [3]. Sampling methods for selection of well spread samples in a general auxiliary space, by utilizing a distance function, are more recent and less well known than the cube method. Two such methods are the local pivotal method (LPM) and spatially correlated Poisson sampling (SCPS). The LPM design, based on the pivotal method [4], was first introduced in [5]. The other method, SCPS, was first introduced in [6] and it is a special case of the method described in [7].

In many areas, such as forest inventories, environmental studies, and even in official statistics, different forms of stratification are commonly used to obtain samples that are well spread geographically or in other available information. Often, stratification is used as a variance reduction technique without particular interest in the different strata. Constructing a stratified sampling design is often not straightforward, especially if several mixed auxiliary variables are available. It is not uncommon that statisticians try to stratify using several variables, but crossing all strata of all variables usually results in cells that are too small. In such situations it may be preferable and less complicated to define a distance measure in the auxiliary space, and then use a sampling method that in general avoid selection of nearby units, thus forcing the sample to be well spread.

In Section 2, a theoretical motivation for the balancing effect of well spread samples is given. In Section 3, we give arguments indicating that using well spread samples provides a small anticipated variance for the HT-estimator under a very general super-population model. Some sampling methods for selecting well spread samples are briefly discussed in Section 4. Final comments are provided in Section 5.

2. Main Results

We start by introducing some notation and assumptions. Let $U=\{1,2,\cdots,N\}$ be a population of N units. We wish to select a probability sample s of size n in order to

estimate some characteristics of U. It is assumed that we have access to auxiliary information on unit level, *i.e.* the values of q auxiliary variables $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{iq}) \in R^q$ are known for each unit $i \in U$. We also assume that it is possible to calculate the distance $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ between two units i and j in the auxiliary space. Usually the total $Y = \sum_{i \in U} y_i$ of one or more target variables are the parameters we wish to estimate. It is assumed that each population unit i is included in the sample with a known probability π_i, $i \in U$, with $\sum_{i \in U} \pi_i = n$, where n is the sample size. In this case the unbiased and commonly used HT-estimator [8] of Y is

$$\hat{Y} = \sum_{i \in s} \frac{y_i}{\pi_i}.$$

We are now ready to formalize what a well spread sample is. As suggested in [2], we use Voronoi polytopes to measure how well spread a sample is. The Voronoi polytope p_i, for $i \in s$, includes all population units j satisfying $d(\boldsymbol{x}_i, \boldsymbol{x}_j) \le d(\boldsymbol{x}_k, \boldsymbol{x}_j)$ for all sample units $k \in s$. Let n_i denote the number of units in p_i, with the correction that if a unit j is included in m_j polytopes, then j is counted as $1/m_j$. Next, let v_i be the sum of the inclusion probabilities in p_i. Again, if a unit j is included in m_j polytopes, then its inclusion probability is divided equally (π_j / m_j) to each of the m_j polytopes. Hence,

$$n_i = \sum_{j \in p_i} \frac{1}{m_j} \text{ and } v_i = \sum_{j \in p_i} \frac{\pi_j}{m_j}.$$

Note that $\sum_{i \in s} n_i = N$ and $\sum_{i \in s} v_i = n$. We are now ready to give the definition of a well spread sample.

Definition 1 *A sample is said to be well spread* (*or spatially balanced*) *with respect to the inclusion probabilities if each v_i is equal or close to* 1.

As a measure of how well spread a sample is, we may use

$$B = \frac{1}{n} \sum_{i \in s} (v_i - 1)^2, \tag{1}$$

see e.g. [2]. A small value of B indicates a very well spread sample. The mean of B over repeated samples is an indicator of how well spread samples a design produces. We next define a balanced and an approximately balanced sample.

Definition 2 *We say that a sample s is balanced on the auxiliary* $\boldsymbol{x}$*-variables if*

$$\sum_{i \in U} \boldsymbol{x}_i = \sum_{i \in s} \frac{\boldsymbol{x}_i}{\pi_i}.$$

Moreover, a sample is said to be approximately balanced if $\sum_{i \in U} \boldsymbol{x}_i$ *is close to* $\sum_{i \in s} \frac{\boldsymbol{x}_i}{\pi_i}$.

In order to show that well spread samples are balanced we start by making three quite strong assumptions on the sample and the population. Later we will relax these assumptions a bit and then show that a larger class of well spread samples are approximately balanced. We first assume the following.

(A.0) In each polytope the inclusion probabilities sum to 1, *i.e.*

$$v_i = 1 \text{ for all } i \in s.$$

(A.1) In each polytope the inclusion probabilities are equal, *i.e.* for every $i \in s$, we assume

$$\pi_j = \pi_i \text{ for all } j \in p_i.$$

(A.2) In each polytope the auxiliary variables are equal for all units, *i.e.* for every $i \in s$,

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = 0 \text{ for all } j \in p_i.$$

The assumptions (A.0) and (A.1) tell us that the size n_i of the polytope p_i is equal to $1/\pi_i$ and (A.2) tells us that $f(\boldsymbol{x}_j) = f(\boldsymbol{x}_i)$, for $j \in p_i$. Note that π_j, $\boldsymbol{x}_j$ and hence $f(\boldsymbol{x}_j)$ are allowed to vary between polytopes. Under the three assumptions it follows that

$$\begin{aligned} \sum_{i \in U} f(\boldsymbol{x}_i) &= \sum_{i \in s} \sum_{j \in p_i} \frac{f(\boldsymbol{x}_j)}{m_j} \\ &= \sum_{i \in s} f(\boldsymbol{x}_i) \sum_{j \in p_i} \frac{1}{m_j} = \sum_{i \in s} \frac{f(\boldsymbol{x}_i)}{\pi_i}. \end{aligned} \tag{2}$$

Thus the sample is balanced on any function $f(\boldsymbol{x})$ and in particular, it is balanced on the auxiliary variables if we put $f(\boldsymbol{x}) = \boldsymbol{x}$.

The next step consists of introducing the following three new and less restrictive assumptions.

$(A'.1)$ For each polytope p_i, the inclusion probabilities satisfies

$$|\pi_i - \pi_j| \le \gamma_i \text{ for all } j \in p_i \text{ and some } 0 \le \gamma_i < \pi_i.$$

$(A'.2)$ In each polytope p_i, we have

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) \le r_i \text{ for all } j \in p_i \text{ and some } r_i \ge 0.$$

$(A'.3)$ The target is a Lipschitz continuous function of the auxiliary variables, *i.e.*

$$|f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j)| \le c_i d(\boldsymbol{x}_i, \boldsymbol{x}_j)$$
$$\text{for all } j \in p_i \text{ and some } c_i \ge 0.$$

Remark 1 *Concerning the validity of assumption* $(A'.1)$, *the inclusion probabilities are* (*if unequal*) *supposed to be derived from the auxiliary* $\boldsymbol{x}$*-variables, perhaps they are chosen proportional to one of the* $\boldsymbol{x}$*-vari-*

ables, so they should not vary much within a polytope. Remember that the polytopes are constructed by grouping together units with similar ***x****-values.*

We are now ready to state and prove our main result.

Theorem 1 *Let* s *be a well spread sample satisfying* $|v_i - 1| \le \epsilon_i$ *for all* $i \in s$ *and for some* $\epsilon_i \ge 0$. *Assume also that* s *is from a population satisfying assumptions* $(A'.1)$-$(A'.3)$. *Then* s *is approximately balanced. In particular*,

$$\sum_{i\in s}\frac{f(\boldsymbol{x}_i)-c_i r_i}{\pi_i+\gamma_i}(1-\epsilon_i)\le\sum_{i\in U}f(\boldsymbol{x}_i)\le\sum_{i\in s}\frac{f(\boldsymbol{x}_i)+c_i r_i}{\pi_i-\gamma_i}(1+\epsilon_i).$$

By sending $\epsilon_i, c_i r_i, \gamma_i \to 0$ we obtain exact balance on the target since

$$\sum_{i\in s}\frac{f(\boldsymbol{x}_i)}{\pi_i}=\sum_{i\in U}f(\boldsymbol{x}_i).$$

Note also that if we put $f(\boldsymbol{x})=\boldsymbol{x}$ then we get a balanced sample, see Definition 2. Besides that Theorem 1 tells us that when $\epsilon_i, r_i c_i, \gamma_i$ are small, the sample will be approximately balanced on $f(\boldsymbol{x})$ and $\boldsymbol{x}$, it also gives bounds for the target parameter $\sum_{i\in U} f(\boldsymbol{x}_i)$. We can however do better than the bounds in Theorem 1. For instance, we have

$$\sum_{i\in s}\sum_{j\in p_i}\frac{f(\boldsymbol{x}_i)-c_i d(\boldsymbol{x}_i,\boldsymbol{x}_j)}{m_j}\le\sum_{i\in U}f(\boldsymbol{x}_i)\le\sum_{i\in s}\sum_{j\in p_i}\frac{f(\boldsymbol{x}_i)+c_i d(\boldsymbol{x}_i,\boldsymbol{x}_j)}{m_j},$$

but these bounds are constructed by applying a worst case senario, within each polytope, so we cannot expect the bounds to be very good.

Proof of Theorem 1. By assumption $(A'.1)$ and since $|v_i - 1| \le \epsilon_i$ we have, for all $i \in s$,

$$\begin{aligned}1+\epsilon_i &\ge \sum_{j\in p_i}\frac{\pi_j}{m_j}\ge n_i(\pi_i-\gamma_i)\ \text{and}\\ 1-\epsilon_i &\le \sum_{j\in p_i}\frac{\pi_j}{m_j}\le n_i(\pi_i+\gamma_i).\end{aligned}\tag{3}$$

The inequalities in (3) give

$$\frac{1-\epsilon_i}{\pi_i+\gamma_i}\le n_i\le\frac{1+\epsilon_i}{\pi_i-\gamma_i}.\tag{4}$$

Moreover, from assumptions $(A'.2)$ and $(A'.3)$ we see that

$$\left|f(\boldsymbol{x}_j)-f(\boldsymbol{x}_i)\right|\le c_i d(\boldsymbol{x}_i,\boldsymbol{x}_j)\le c_i r_i.\tag{5}$$

Theorem 1 now follows from (4) and (5). In particular,

$$\begin{aligned}\sum_{i\in U}f(\boldsymbol{x}_i)&=\sum_{i\in s}\sum_{j\in p_i}\frac{f(\boldsymbol{x}_j)}{m_j}\ge\sum_{i\in s}n_i\left(f(\boldsymbol{x}_i)-c_i r_i\right)\\ &\ge\sum_{i\in s}\frac{f(\boldsymbol{x}_i)-c_i r_i}{\pi_i+\gamma_i}(1-\epsilon_i)\end{aligned}$$

and

$$\begin{aligned}\sum_{i\in U}f(\boldsymbol{x}_i)&=\sum_{i\in s}\sum_{j\in p_i}\frac{f(\boldsymbol{x}_j)}{m_j}\le\sum_{i\in s}n_i\left(f(\boldsymbol{x}_i)+c_i r_i\right)\\ &\le\sum_{i\in s}\frac{f(\boldsymbol{x}_i)+c_i r_i}{\pi_i-\gamma_i}(1+\epsilon_i).\end{aligned}$$

The proof is complete. □

3. Variance under a General Model

It is interesting to see how well spread samples perform under a general super-population model. Following [9], but here with a possibly non-linear model, we assume

$$y_i = f(\boldsymbol{x}_i)+\epsilon_i \text{ for all } i\in U,\tag{6}$$

where $f(\cdot)$ is a Lipschitz continuous function, $E_M(\epsilon_i)=0$, $V_M(\epsilon_i)=\sigma^2(\boldsymbol{x}_i)=\sigma_i^2$, and $\sigma^2(\cdot)$ is a Lipschitz continuous function. Moreover,

$$\operatorname{cov}_M(\epsilon_i,\epsilon_j)=\sigma_i\sigma_j\rho_{ij} \text{ with } i\ne j,$$

where $E_M(\cdot)$, $V_M(\cdot)$ and $\operatorname{cov}_M(\cdot,\cdot)$ are the expectation, variance and covariance under the model. The correlations ρ_{ij} are supposed to be decreasing in function of the distance between the units *i* and *j*.

With some routine calculations, the anticipated variance of the HT-estimator under model (6) can be shown to be

$$\begin{aligned}E_p E_M\left(\hat{Y}-Y\right)^2 &= E_p\left(\sum_{i\in s}\frac{f(\boldsymbol{x}_i)}{\pi_i}-\sum_{i\in U}f(\boldsymbol{x}_i)\right)^2\\ &\quad+\sum_{i\in U}\sum_{j\in U}\sigma_i\sigma_j\rho_{ij}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j},\end{aligned}\tag{7}$$

where $E_p(\cdot)$ is the expectation under the design. Now, if we study expression (7), it becomes evident that we want the samples to be as balanced as possible on $f(\boldsymbol{x})$ to minimize the term

$$E_p\left(\sum_{i\in s}\frac{f(\boldsymbol{x}_i)}{\pi_i}-\sum_{i\in U}f(\boldsymbol{x}_i)\right)^2.$$

We also want to make sure that π_{ij} is small whenever ρ_{ij} is large in order to minimize the term

$$\sum_{i\in U}\sum_{j\in U}\sigma_i\sigma_j\rho_{ij}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}.\tag{8}$$

If the samples are selected to be well spread (*i.e.* small joint inclusion probabilities for nearby units), then both terms in (7) becomes small. However, if the model standard deviation σ_i is known, it is possible to also choose the inclusion probabilities to minimize further. The diagonal term of (8) is dominant, *i.e.*

$$\sum_{i\in U}\sum_{j\in U}\sigma_i\sigma_j\rho_{ij}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\approx\sum_{i\in U}\sigma_i^2\frac{1-\pi_i}{\pi_i}. \qquad (9)$$

With the constraint of fixed sample size $\sum_{i\in U}\pi_i = n$ and by using a Lagrangian function, it follows that the minimum in π_i of the right hand side of (9) is

$$\pi_i = \frac{n\sigma_i}{\sum_{j\in U}\sigma_j}, \qquad (10)$$

if each $n\sigma_i < \sum_{j\in U}\sigma_j$. As a result, a very efficient sampling design under this general model is to select samples that are well spread in the $\boldsymbol{x}$-space with inclusion probabilities given by (10). The requirements needed in order for the samples to be approximately balanced on $f(\boldsymbol{x})$ are then fulfilled. The inclusion probabilities will not vary much within the polytopes since $\sigma(\boldsymbol{x})$ is supposed to be a Lipschitz continuous function of $\boldsymbol{x}$. Hence, with this strategy, the anticipated variance of the HT-estimator becomes small.

It is not possible to balance the sample directly on $f(\boldsymbol{x})$ since the function is obviously not known in advance. Probably, the best we can do in practise is to make sure the samples are well spread in $\boldsymbol{x}$ to have a balancing effect on the unknown function $f(\boldsymbol{x})$, and hence also have small π_{ij} when ρ_{ij} may be large.

If we use well spread probability samples together with the HT-estimator, the estimator will be very efficient (*i.e.* have a small variance) if the population is close to a realization of the model (6). Note also that the approach is purely design based, and the estimator maintains design unbiasedness and design consistency even if the model is false.

Example 2, given in the next section, supports the above statements. In particular, the example compares different sampling methods with respect to variance and spatial balance. It is clear that methods obtaining well spread samples are more balanced and hence produce smaller variance.

4. Some Methods for Selecting Well Spread Samples

Besides spatial stratification, one of the first more novel designs for selecting a well spread sample is called generalized random tessellation stratified (GRTS), and was introduced in [2]. The GRTS design uses a specific random mapping to map two (or more) dimensions to one dimension. Basically the units are re-ordered to a list and units close in the list tend to also be close in the auxiliary space. Then a systematic πps sample is selected from the list, making sure the sample becomes well spread in the list and hence also in the auxiliary space. A drawback of GRTS is that a lot of information is lost in the mapping, especially if the space has many dimensions (*i.e.* many auxiliary variables). However, for two dimensions, the GRTS produces rather well spread samples.

Another idea is to map dimensions to one by use of space-filling curves, and one such design was presented and evaluated in [10]. However, we believe that mapping several dimensions to one is not the best way to achieve a well spread sample. Too much information is lost in such a mapping.

A more recent idea to achieve well spread samples is to first define a distance measure in the auxiliary space. To do so, let $\boldsymbol{x}\in R^q$ be all available auxiliary variables, where $\{1,\cdots,p\}$ correspond to the quantitative variables and $\{p+1,\cdots,q\}$ to the qualitative variables. To measure the distance between unit i and j in this q-dimensional space, [11] propose the following definition of distance

$$d(\boldsymbol{x}_i,\boldsymbol{x}_j) = \sqrt{\sum_{k=1}^{p}\left(x'_{ik}-x'_{jk}\right)^2} + \sum_{k=p+1}^{q}1_{x_{ik}\neq x_{jk}},$$

where x'_k is the standardized version of x_k. By standardizing, the auxiliary variables are approximately of equal importance. However, the above distance function is just an example and in a particular situation some other distance function may be more appropriate. Given the distace measure, the design should create a negative correlation of the inclusion indicators for close units, so that two close units seldom appear in the sample together. Such a design is not necessarily complicated. For instance, the local pivotal method (LPM) introduced in [5] is quite simple. The LPM is based on the pivotal method [4]. The main idea in LPM is to make similar units (*i.e.* nearby units) compete with each other for inclusion in the sample. The LPM successively updates the prescribed vector of inclusion probabilities $(\pi_1,\pi_2,\cdots,\pi_N)$ to become a vector with zeros and ones, where the ones indicate inclusion in the sample. In one step of LPM, two close units i and j with $0<\pi_i<1$ and $0<\pi_j<1$ are chosen to compete. The winner takes as much probability mass as possible from the other unit. Hence, the winner receives the new probability $a=\min(1,\pi_i+\pi_j)$ and the looser gets the new probability $b=\pi_i+\pi_j-a$. Thus, if $\pi_i+\pi_j\geq 1$, then a = 1 and the winning unit will definitely be in the sample. If $\pi_i+\pi_j<1$, then the looser will definitely not be in the sample (since b = 0). The reduced probability vector (π_i,π_j) is updated as

$$\left(\pi_i', \pi_j'\right) = \begin{cases} (a,b), \text{with probability } \left(a-\pi_j\right)/\left(a-b\right) \\ (b,a), \text{with probability } \left(a-\pi_i\right)/\left(a-b\right) \end{cases}$$

Now, replace $\left(\pi_i, \pi_j\right)$ with $\left(\pi_i', \pi_j'\right)$. The final outcome is decided for at least one unit each update, and thus the procedure has at most N steps. In each update, unit i is chosen randomly (with equal probabilities among the units with $0<\pi_i<1$) and then its nearest neighbor $j \neq i$ (among the units with $0<\pi_j<1$) is chosen.

Another method, spatially correlated Poisson sampling (SCPS) was first described in [6] and it is a special case of the method introduced in [7]. The SCPS algorithm is a bit more complicated than LPM, but is based on the same idea. Weights are used to create a negative correlation between the inclusion indicators of nearby units, forcing the sample to be well spread. For more on the above discussed methods, for selecting well spread samples, we refer the reader to the previously mentioned papers.

The two designs, LPM and SCPS were used in [11] to obtain well spread probability samples. The fact that LPM and SCPS produce well spread samples has been justified by both theoretical results and simulation results in the previously mentioned papers. Variance estimators for the HT-estimator under well spread samples was suggested in the papers [11,12]. To our knowledge LPM and SCPS are the designs that in general produce the lowest mean value of the balance measure (1) in general auxiliary space with prescribed inclusion probabilities.

When it comes to efficiency of the HT-estimator for well spread samples, we can also make heuristic arguments that such samples produce a low variance of the HT-estimator. When the sample size is fixed, the variance of $\hat{X} = \sum_{i\in s} x_i/\pi_i$ can be written as

$$V\left(\hat{X}\right) = -\frac{1}{2}\sum_{(i,j)\in U}\left(\pi_{ij}-\pi_i\pi_j\right)\left(\frac{x_i}{\pi_i}-\frac{x_j}{\pi_j}\right)^2. \qquad (11)$$

A property that e.g. the LPM and SCPS design have is that π_{ij} is small (minimum or close to minimum) when $d\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)$ is small and π_{ij} is large (close to $\pi_i\pi_j$) when $d\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)$ is large. If π_{ij} is small when $d\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)$ is small, then $\left(x_i/\pi_i - x_j/\pi_j\right)^2$ is small when π_{ij} is small, *i.e.* $\left(x_i/\pi_i - x_j/\pi_j\right)^2$ is small when $-\left(\pi_{ij}-\pi_i\pi_j\right)$ is large. Also, if $\left(x_i/\pi_i - x_j/\pi_j\right)^2$ is large (*i.e.* $d\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)$ is large), then $-\left(\pi_{ij}-\pi_i\pi_j\right)$ is small since $\pi_{ij} \approx \pi_i\pi_j$. As a result the variance (11) becomes small.

For well spread samples, the balancing property can only be shown to hold exactly in very specific situations, *i.e.* under assumptions (*A*.0)-(*A*.2), see (2). For a categorical auxiliary variable, the sample will be balanced if the design produces stratification with fixed sample size for each category. A simple example follows.

Example 1. Let U be a population of males U_m and females U_f. Let x be the only auxiliary variable and let $x_i = 0$ if male and $x_i = 1$ if female. Also, let $\pi_i^m = n_m/N_m$ and $\pi_i^f = n_f/N_f$ be the inclusion probabilities, where n_m and n_f are integers. In this special case, we have that e.g. the LPM and SCPS automatically produces stratification with fixed sample sizes. Hence we have

$$\sum_{i\in s}\frac{x_i}{\pi_i} = \sum_{i\in s_m}\frac{0}{\pi_i^m} + \sum_{i\in s_f}\frac{1}{\pi_i^f} = N_f = \sum_{i\in U} x_i,$$

where s_m and s_f are the sampled males and females respectively.

Example 2. We compare the different sampling methods LPM, SCPS, GRTS and simple random sampling (SRS) using a model satisfying (6). In particular, the population is generated from $y_i = f\left(\boldsymbol{x}_i\right) + \epsilon_i$ with

$$f\left(\boldsymbol{x}_i\right) = 3\left(x_{i1}+x_{i2}\right) + \sin\left\{6\left(x_{i1}+x_{i2}\right)\right\}.$$

The population size is N = 200 and the $\boldsymbol{x}$-values are generated from a uniform distribution on the unit square. Using Euclidean distance, the covariance function for $\boldsymbol{\epsilon}$ is defined as $\mathrm{cov}_M\left(\epsilon_i, \epsilon_j\right) = \sigma^2 \cdot \exp\left\{-d\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)/r\right\}$, which is a simple covariance function used for stationary fields [13]. The $\boldsymbol{\epsilon}$-values are generated in two steps. First random independent and identically distributed data $\boldsymbol{e} = \left(e_1, \cdots, e_N\right)'$ is generated from $N(0,1)$. Then, the $\boldsymbol{\epsilon}$-values are constructed as $\boldsymbol{\epsilon} = \Sigma^{1/2}\cdot\boldsymbol{e}$ using the covariance matrix $\Sigma = \left\{\mathrm{cov}_M\left(\epsilon_i, \epsilon_j\right)\right\}$. In this example $\sigma = 0.5$ and $r = 0.1$. The units are sampled with equal inclusion probabilities and $n = 50$ units are sampled. The target parameter is $Y = \sum_{i\in U} y_i$. In our particular realization the true value is $Y \approx 607$. The result is presented in **Table 1**. A clear connection between well spread samples and variance can be observed. A design with a small expected value of *B*, see (1), gives better estimates. Concerning anticipated variance we get similar results if we average over repeated realizations from the model.

Table 1. Results for Example 2. Empirical variance $\hat{V}$ of the HT-estimator and the mean of the measure *B* for 1000 samples of size 50.

Design	$\hat{V}$	Mean (*B*)
SCPS	166	0.137
LPM	196	0.138
GRTS	228	0.176
SRS	1305	0.359

5. Final Comments

It has been shown that in general there is a significant balancing effect for well spread samples. Usually, well spread samples are not as balanced on the auxiliary $\boldsymbol{x}$-variables as samples selected by the cube method, but nearly so if the sample size is not too small. However, for target variables that are non-linear in $\boldsymbol{x}$, well spread samples are likely to be more balanced on the target variables than samples selected by the cube method. In that way, well spread samples are good for more general situations. Hopefully, the fact that a significant balancing effect has been shown will increase the interest of using well spread probability samples when auxiliary $\boldsymbol{x}$-variables are available.

There also exists a possibility to combine the cube method with a similar idea as used in the LPM, to have a local cube method. Then samples that are both well spread (spatially balanced) and balanced on the auxiliary variables can be selected. Such a method was developed in [9].

In [1,14], properties of spatial total estimators are studied under a tessellation stratified design in a continues universe. With similar assumptions on the target function, as used in this paper, they show that the convergence rate of the variance of the total estimator is $O\left(n^{-2}\right)$ for such a design. Even though our setting is different and does not imply a strict stratification, this indicates that spreading the sample locations well probably gives a small variance when there are spatial trends.

In the setting of Voronoi polytopes used in this paper, we may consider the nearest neighbor estimator (NN-estimator) in place of the HT-estimator. The NN-estimator of Y is, if n_i is the number of units in polytope p_i,

$$\sum_{i\in s} y_i n_i.$$

Under the assumptions (A.0) and (A.1), we have $n_i = 1/\pi_i$ and the NN-estimator is equal to the HT-estimator. This implies that the NN-estimator will be approximately design unbiased for well spread samples. Moreover, the NN-estimator can probably adjust for some minor spatial imbalance in the sample by using the realized polytope sizes n_i instead of $1/\pi_i$, which can be viewed as the estimated polytope sizes. The possible benefit of using the NN-estimator in place of the HT-estimator will be investigated in a future paper.

6. Acknowledgements

Thanks to Lennart Bondesson and an anonymous reviewer for helpful comments that improved this manuscript.

REFERENCES

[1] L. Barabesi and S. Franceschi, "Sampling Properties of Spatial Total Estimators under Tessellation Stratified Designs," *Environmetrics*, Vol. 22, No. 3, 2011, pp. 271-278. doi:10.1002/env.1046

[2] D. L. Stevens Jr. and A. R. Olsen, "Spatially Balanced Sampling of Natural Resources," *Journal of the American Statistical Association*, Vol. 99, No. 465, 2004, pp. 262-278. doi:10.1198/016214504000000250

[3] J.-C. Deville and Y. Tillé, "Efficient Balanced Sampling: the Cube Method," *Biometrika*, Vol. 91, No. 4, 2004, pp. 893-912. doi:10.1093/biomet/91.4.893

[4] J.-C. Deville and Y. Tillé, "Unequal Probability Sampling without Replacement through a Splitting Method," *Biometrika*, Vol. 85, No. 1, 1998, pp. 89-101. doi:10.1093/biomet/85.1.89

[5] A. Grafström, N. L. P. Lundström and L. Schelin, "Spatially Balanced Sampling through the Pivotal Method," *Biometrics*, Vol. 68, No. 2, 2012, pp. 514-520. doi:10.1111/j.1541-0420.2011.01699.x

[6] A. Grafström, "Spatially Correlated Poisson Sampling," *Journal of Statistical Planning and Inference*, Vol. 142, No. 1, 2012, pp. 139-147. doi:10.1016/j.jspi.2011.07.003

[7] L. Bondesson and D. Thorburn, "A List Sequential Sampling Method Suitable for Real-Time Sampling," *Scandinavian Journal of Statistics*, Vol. 35, No. 3, 2008, pp. 466-483. doi:10.1111/j.1467-9469.2008.00596.x

[8] D. G. Horvitz and D. J. Thompson, "A Generalization of Sampling without Replacement from a Finite Universe," *Journal of the American Statistical Association*, Vol. 47, No. 260, 1952, pp. 663-685. doi:10.1080/01621459.1952.10483446

[9] A. Grafström and Y. Tillé, "Doubly Balanced Spatial Sampling with Spreading and Restitution of Auxiliary Totals," *Environmetrics*, in Press, 2012. doi:10.1002/env.2194

[10] A. J. Lister and C. T. Scott, "Use of Space-Filling Curves to Select Sample Locations in Natural Resource Monitoring Studies," *Environmental Monitoring and Assessment*, Vol. 149, No. 1-4, 2009, pp. 71-80. doi:10.1007/s10661-008-0184-y

[11] A. Grafström and L. Schelin, "How to Select Representative Samples," *Scandinavian Journal of Statistics*, 2013.

[12] D. L. Stevens Jr. and A. R. Olsen, "Variance Estimation for Spatially Balanced Samples of Environmental Resources," *Environmetrics*, Vol. 14, No. 6, 2003, pp. 593-610. doi:10.1002/env.606

[13] N. A. C. Cressie, "Statistics for spatial data," Wiley, New York, 1993.

[14] L. Barabesi and M. Marcheselli, "A Modified Monte Carlo Integration," *International Mathematical Journal*, Vol. 3, No. 5, 2003, pp. 555-565.

