Variants of the splitting method for unequal probability sampling

Lennart Bondesson[1]

[1] Umeå University, Sweden
e-mail: Lennart.Bondesson@math.umu.se

Abstract

This paper is mainly a review of the splitting method for unequal probability sampling. However, it also contains some significant novelties.

1 Introduction

Let $\mathcal{U} = \{1, 2, \ldots, N\}$ be a population of units for which information about the mean or total of some interesting y-variable is required. Often a sample survey is needed to get that information. Simple random sampling without replacement (SRS) is sometimes suitable but often it is more efficient to sample with unequal inclusion probabilities π_i, $i \in \mathcal{U}$, for the units . In Särndal *et al.* (1992) unequal probability sampling is called πps sampling. Many different designs for πps sampling have been suggested over the years. Comprehensive accounts are given by Brewer and Hanif (1983) and Chaudhuri and Vos (1988).

The *splitting method*, which was introduced by Deville and Tillé (1998) and is further described by Tillé (2006, Chapter 6), is a general method for obtaining a πps sample. It uses the idea that the inclusion probability vector $\boldsymbol{\pi}$ can be seen as a point in an N-dimensional unit cube and that a sample can be seen as corner of that cube. To get a sample, we may start at $\boldsymbol{\pi}$ and perform a very general random walk without drift within the cube and then on its faces and subfaces until a corner is reached. To get fixed sample size and different balancing conditions satisfied, restrictions should be put on the random walk.

In this paper the general method is first described. Then special cases are presented in examples. The paper also contains a brief description of the cube method for balanced sampling. Finally there are some concluding comments.

2 The splitting method, general description

Let π_i, $i \in \mathcal{U}$, be given inclusion probabilities. Often they sum to a desired fixed sample size n. The vector $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$ is a point in the cube

$$\mathcal{C} = \{\mathbf{x};\ 0 \leq x_i \leq 1,\ i = 1, 2, \ldots, N\}.$$

Each corner $\mathbf{x}$ of the cube can be seen as a sample: if $x_i = 1$, the population unit i belongs to the sample and if $x_i = 0$, it does not. The general idea of the splitting method is to perform in discrete time, $t = 0, 1, 2, \ldots$, a very general random walk without drift within the cube and on the faces and the subfaces of the cube. It starts at $\boldsymbol{\pi}(0) = \boldsymbol{\pi}$ and ends as soon as a corner has been reached. Since it has no drift, sampling is performed with the given inclusion probabilities. Mathematically the walk is described by a *martingale*

$$\boldsymbol{\pi}(t) = \boldsymbol{\pi}(t-1) + \mathbf{e}(t) \quad \text{with} \quad E(\mathbf{e}(t)|\mathcal{F}_{t-1}) = \mathbf{0}, \quad t = 1, 2, \ldots,$$

where $\mathcal{F}_{t-1}$ denotes the 'past'. Hence $E(\boldsymbol{\pi}(t)) = E(\boldsymbol{\pi}(t-1)) = \ldots = \boldsymbol{\pi}(0) = \boldsymbol{\pi}$.

There are many possible such random walks and thus the method is very general and yields many different sampling designs. If $\sum_{\mathcal{U}} \pi_i = n$ and the walk is restricted to be in the hyperplane $\{\mathbf{x}; \sum_{i=1}^{N} x_i = n\}$, which contains all corners of the cube with exactly n coordinates equal to 1, then a sample of size n is obtained.

The name of the method is better motivated by the following alternative and somewhat more general description. In the first step, $\boldsymbol{\pi}$ is split as $\boldsymbol{\pi} = \sum_{k=1}^{M} p_k \boldsymbol{\pi}^{(k)}$, where each $\boldsymbol{\pi}^{(k)}$ belongs to the cube and $0 < p_k < 1$ with $\sum_{k=1}^{M} p_k = 1$. A random choice of vector $\boldsymbol{\pi}^{(k)}$ according to the probabilities p_k, $k = 1, 2, \ldots, M$, is made. The new vector $\boldsymbol{\pi}(1)$ is then further split and so on. It can be arranged so that for some $\boldsymbol{\pi}(t)$ it may happen that there is a simple way of getting the sample and then the procedure can be finished rapidly. If $\sum_{\mathcal{U}} \pi_i = n$ and, in every step, each $\boldsymbol{\pi}^{(k)}$ has coordinate sum equal to n, the sample size will be n.

The case $M = 2$ is illuminating. An arbitrary vector $\mathbf{u} = (u_1, u_2, \ldots, u_N)$ is chosen. A random $\mathbf{u}$ is possible. Let

$$\boldsymbol{\pi}(1) = \begin{cases} \boldsymbol{\pi} + \lambda_1 \mathbf{u} & \text{with probability } p_1 \\ \boldsymbol{\pi} - \lambda_2 \mathbf{u} & \text{with probability } p_2, \end{cases}$$

where $p_1 = \lambda_2/(\lambda_1 + \lambda_2)$ and $p_2 = 1 - p_1$. Then $E(\boldsymbol{\pi}(1)) = \boldsymbol{\pi}$. The scalars λ_1 and λ_2 must be chosen such that all the coordinates of $\boldsymbol{\pi}(1)$ are in the interval $[0, 1]$. They can also be chosen such that $\max_i\{\pi_i + \lambda_1 u_i\} = 1$ and $\min_i\{\pi_i - \lambda_2 u_i\} = 0$, meaning that $\boldsymbol{\pi}(1)$ will be on one of the faces of the cube, i.e. one of the coordinates of $\boldsymbol{\pi}(1)$ is 0 or 1. The procedure then proceeds on that face with a new $\mathbf{u}$ and after a new step one further coordinate is 0 or 1, etc. A coordinate that once has become 0 or 1 cannot be further changed, i.e. the sample is gradually appearing during the procedure. A sample of fixed size n is obtained if $\sum_{\mathcal{U}} \pi_i = n$ and, in each step, $\sum_{i=1}^{N} u_i = 0$, i.e. if $\mathbf{u} \perp (1, 1, \ldots, 1)$.

3 Different applications of the splitting method

There are many different implementations of the splitting method, cf. Tillé (2006, Chapter 6). This section presents old and novel ones. As will be seen, not seldom well-established sampling procedures can be described in terms of splitting.

Often only the first step of the procedure needs to be described since in the subsequent steps everything is repeated on a smaller population, or, equivalently, on a cube with lower dimension, and with an updated inclusion probability vector. It is assumed that all π_i are strictly between 0 and 1 since for a unit with π_i equal to 0 or 1 the sampling outcome is already clear.

In the examples there are also given small comments on the properties of the sampling designs. These comments concern entropy and possible invariance with respect to the dual *eliminatory* design defined below. The entropy of a design with inclusion probabilities π_i and probability function $p(\mathbf{x}; \boldsymbol{\pi})$ is given by $\mathcal{E} = -\sum_{\mathbf{x}} p(\mathbf{x}; \boldsymbol{\pi}) \log(p(\mathbf{x}; \boldsymbol{\pi}))$, with summation over all possible samples $\mathbf{x}$. It is desirable to have a high entropy, i.e. randomness, for a sampling design. For each design there is an eliminatory design with the same inclusion probabilities and with probability function $p^*(\mathbf{x};\ \boldsymbol{\pi}) = p(\mathbf{1}-\mathbf{x}; \mathbf{1}-\boldsymbol{\pi})$. It derives from choosing the sample as the complement of a sample according to the inclusion probabilities $1-\pi_i$, $i \in \mathcal{U}$. Preferably a chosen design is invariant with respect to its eliminatory design, i.e. $p^*(\mathbf{x};\ \boldsymbol{\pi}) = p(\mathbf{x};\ \boldsymbol{\pi})$, since otherwise the entropy can be increased by making a random choice among the two designs.

Example 1 (Poisson sampling). Assume that the π_i:s are given with two decimals. With start at $\boldsymbol{\pi}$, perform a simple symmetric random walk with step-length 0.01, or a Brownian motion without drift, within the cube and then on its faces and subfaces until a corner is reached. It can be verified that this procedure is equivalent to Poisson sampling. Thus, if the random indicator I_i is 1 when unit i is sampled and otherwise 0, then $P(I_i = 1) = \pi_i$ and all the indicators are independent. The sample size is random. Of course, the random walk method is not the most efficient way of implementing Poisson sampling. The design has maximal entropy and it is invariant with respect to its eliminatory design.

Example 2 (The pivot method). Let $\sum_{\mathcal{U}} \pi_i = n$. The pivot method is a simple method to obtain a πps sample of fixed size n. Choose two units in the population, either randomly or according to some systematic scheme. It is no restriction to assume that those are units 1 and 2. Then the reduced vector $\boldsymbol{\pi} = (\pi_1, \pi_2)$ is split and updated as follows:

$$\text{if } \pi_1 + \pi_2 < 1: \qquad \boldsymbol{\pi}(1) = \begin{cases} (\pi_1 + \pi_2,\, 0) & \text{with probability } p_1 = \frac{\pi_1}{\pi_1+\pi_2} \\ (0,\, \pi_1 + \pi_2) & \text{with probability } p_2 = \frac{\pi_2}{\pi_1+\pi_2} \end{cases}$$

$$\text{if } \pi_2 + \pi_2 \geq 1: \qquad \boldsymbol{\pi}(1) = \begin{cases} (\pi_1 + \pi_2 - 1,\, 1) & \text{with probability } p_1 = \frac{1-\pi_1}{2-\pi_1-\pi_2} \\ (1,\, \pi_1 + \pi_2 - 1) & \text{with probability } p_2 = \frac{1-\pi_2}{2-\pi_1-\pi_2}. \end{cases}$$

Thus if $\pi_1 + \pi_2 < 1$, one of the two units will not be in the final sample whereas if $\pi_1 + \pi_2 \geq 1$, one of them will be there. This non-sampled/sampled unit is then excluded from further consideration and the procedure is repeated on the $N - 1$ remaining units and with an updated inclusion probability vector with all its coordinates strictly between 0 and 1. Finally a sample of size n is obtained.

The pivot design is invariant with respect to its eliminatory design but the entropy is usually a little lower than what is possible to achieve for a πps design of fixed size.

Example 3 (Generalized Sampford sampling). This is an extension of the pivot method to the case that an arbitrary number of units are picked out in each step. It is no restriction to assume that these units are units $1, 2, \ldots, M$ with corresponding inclusion probabilities $\pi_1, \pi_2, \ldots, \pi_M$. The sum of them is $m + a$, where m is an integer and $a \in [0, 1)$. We would like to give m units the new inclusion probabilities 1, one unit the new inclusion probability a, and all the remaining $M - m - 1$ units the new inclusion

probabilities 0. Thus the sampling outcome should be made definite for all the units except a single one. That unit can then be picked out again in some later step.

We proceed as follows. First one unit is drawn with replacement according to the probabilities π_i, $i = 1, 2, \ldots, M$, normalized to have sum 1. Then m further units are drawn with replacement according to probabilities $p_i' \propto \pi_i/(1-\pi_i)$ with sum 1. Provided that all these $m+1$ units are distinct, the outcome is accepted, otherwise the procedure is fully repeated until acceptance. Then the first drawn unit is given the new inclusion probability a whereas the other m units get the inclusion probabilities 1. All remaining units get the inclusion probabilities 0. With sampling indicators $I_i, i = 1, 2, \ldots, M$, we put $I_i = a$ for the first drawn unit, $I_i = 1$ for the m units, and $I_i = 0$ for all remaining units. Then indeed, as desired, $E(I_i) = \pi_i$. This result is an extension of Sampford's (1967) well-known result. Sampford's result covers the case $a = 1$. The verification of the extension is omitted since it is rather technical.

This method can also be used when the units of the population come in blocks to the sampler. The sampler decides the sampling outcome for all the units except one in the block and leave the sampling decision for the remaining unit to a later occasion.

Example 4 (Reduction to SRS). The goal is to obtain a πps sample of fixed size n by reducing after splitting the sampling to simple random sampling. In each step the splitting is into two components but one of these corresponds to SRS. In the first step we split as $\boldsymbol{\pi} = p_1\boldsymbol{\pi}^{(1)} + (1-p_1)\boldsymbol{\pi}^{(2)}$, where $\boldsymbol{\pi}^{(1)} = \frac{n}{N}[1, 1, \ldots, 1]$. Both $\boldsymbol{\pi}^{(1)}$ and $\boldsymbol{\pi}^{(2)}$ have coordinate sum equal to n. To get the coordinates of $\boldsymbol{\pi}^{(2)}$ to belong to [0,1], it is necessary and sufficient that $0 \le p_1 \le p_u = \min(\alpha, \beta)$, where

$$\alpha = \min_i\{\frac{N}{n}\pi_i\} \quad \text{and} \quad \beta = \min_i\{\frac{N}{N-n}(1-\pi_i)\}.$$

The upper bound p_u can be used as p_1-value and then $\boldsymbol{\pi}^{(2)}$ be calculated. If $\boldsymbol{\pi}^{(1)}$ is chosen, SRS is performed and the procedure ends. If $\boldsymbol{\pi}^{(2)}$ is chosen, first all coordinates that are 0 are removed, and then a new analogous splitting of the reduced $\boldsymbol{\pi}^{(2)}$ is performed, and so on. Thus sometimes the procedure rapidly yields a sample. However, the execution time has a large variation.

The design is invariant with respect to its eliminatory design and but it does not give high entropy.

Example 5 (Brewer's method). The goal is again to get a πps sample of fixed size n. Let $\boldsymbol{\pi}^{(k)}$, $k = 1, 2, \ldots, N$, be inclusion probability vectors with coordinates

$$\pi_i^{(k)} = \begin{cases} 1 & \text{if } i = k \\ \frac{n-1}{n-\pi_k}\pi_i & \text{if } i \neq k. \end{cases}$$

Each $\boldsymbol{\pi}^{(k)}$ has coordinate sum equal to n. Let p_k be *proportional* to $\frac{n-\pi_k}{1-\pi_k}\pi_k$ and such that $\sum_{k=1}^N p_k = 1$. Then, as can be verified, $\boldsymbol{\pi} = \sum_{k=1}^N p_k \boldsymbol{\pi}^{(k)}$. Choose $\boldsymbol{\pi}^{(k)}$ with probability p_k. This implies that unit k is sampled and it is not further considered. Then a new splitting with n changed to $n-1$ is performed, etc. The procedure stops after n splits with a sample of size n obtained.

Originally Brewer (1975) suggested this method as a draw by draw procedure. After unit k has been selected in the first draw, the inclusion probabilities for the remaining units are updated to $\pi_i^{(k)} = \beta_k \pi_i$, $i \neq k$. Since their sum should be $n-1$, necessarily $\beta_k = (n-1)/(n-\pi_k)$. The probabilities p_k for the first unit to draw must be as above.

Brewer's design is not invariant with respect to its eliminatory design but in spite of that it has a fairly high entropy.

Example 6 (Generalized Sunter method). Sunter's (1986) method is a list sequential procedure for obtaining a πps sample of fixed size n. The units of $\mathcal{U}$ are successively gone through in the order $1, 2, \ldots, N$. In the first step unit 1 is sampled with probability π_1 and the inclusion probabilities for the other units are updated as $\pi_i(1) = \frac{n-I_1}{n-\pi_1}\pi_i$, $i \geq 2$, where I_1 is the outcome (0 or 1) for unit 1. In the SRS case the procedure reduces to a procedure by Fan *et al.* (1962). The following scheme illustrates the updating:

$$\begin{array}{lccccc} t=0: & \pi_1 & \pi_2 & \pi_3 & \pi_4 & \ldots \\ t=1: & I_1 & \pi_2^{(1)} & \pi_3^{(1)} & \pi_4^{(1)} & \ldots \\ t=2: & I_1 & I_2 & \pi_3^{(2)} & \pi_4^{(2)} & \ldots \\ t=3: & I_1 & I_2 & I_3 & \pi_4^{(3)} & \ldots \end{array}$$

Since the updated inclusion probabilities may exceed 1, the updating must be modified sometimes. Sunter's widely applied practical procedure is not a strict πps sampling method. Tillé (2006, pp. 108-111) presented a remedy.

The vector $\boldsymbol{\pi}$ is split into two vectors $\boldsymbol{\pi}^{(1)}$ and $\boldsymbol{\pi}^{(2)}$, but $\boldsymbol{\pi}^{(2)}$ has two alternative forms:

$$\pi_i^{(1)} = \begin{cases} 1 & \text{if } i = 1 \\ \frac{n-1}{n-\pi_1}\pi_i & \text{else} \end{cases} \quad \text{and} \quad \pi_i^{(2a)} = \begin{cases} 0 & \text{if } i = 1 \\ \frac{n}{n-\pi_1}\pi_i & \text{else} \end{cases} \quad \text{or} \quad \pi_i^{(2b)} = \begin{cases} \alpha & \text{if } i = 1 \\ \beta\pi_i & \text{else,} \end{cases}$$

where $\beta = (\max_{i\geq 2} \pi_i)^{-1}$ and $\alpha = n - (n - \pi_1)\beta$ guaranteeing that $\sum_{\mathcal{U}} \pi_i^{(2b)} = n$. The alternative $\boldsymbol{\pi}^{(2b)}$ is relevant if $\pi_i^{(2a)} > 1$ for some $i \geq 2$. For alternative (a), the vector $\boldsymbol{\pi}^{(1)}$ is chosen with probability $p_1 = \pi_1$ and $\boldsymbol{\pi}^{(2a)}$ with probability $p_2 = 1 - \pi_1$. For alternative (b), the probabilities are $p_1 = (\pi_1 - \alpha)/(1 - \alpha)$ and $p_2 = 1 - p_1$. Since $0 < \alpha < \pi_1$ and $\boldsymbol{\pi} = p_1\boldsymbol{\pi}^{(1)} + p_2\boldsymbol{\pi}^{(2)}$ for both the alternatives, the procedure works.

If the alternative $\boldsymbol{\pi}^{(2a)}$ is suitable, after the first random choice the sampling outcome, 0 or 1, becomes known for unit 1. The outcome may also become known for some other unit since $\pi_i^{(2a)}$ may be 1 for some $i \geq 2$. However, if the alternative $\boldsymbol{\pi}^{(2b)}$ is the relevant one, the sampling outcome for unit 1 is still unknown if $\boldsymbol{\pi}^{(2b)}$ happens to be chosen. But in that case at least some other unit of the population gets the new inclusion probability 1. Thus in both cases the procedure can proceed on a smaller population.

Like Brewer's design, the generalized Sunter design is not invariant with respect to its eliminatory design but it has a fairly high entropy.

Example 7 (Bulldozer method). The bulldozer method is also a list sequential method. It is similar to Sunter's method but the updating is not done with factors. Given weights $w_2, w_3, \ldots, w_N$, usually non-negative, put

$$\pi_i^{(1)} = \begin{cases} 1 & \text{if} \quad i = 1 \\ \pi_i - (1 - \pi_i)w_i & \text{else} \end{cases} \quad \text{and} \quad \pi_i^{(2)} = \begin{cases} 0 & \text{if} \quad i = 1 \\ \pi_i + \pi_i w_i & \text{else} \end{cases}$$

and choose $\boldsymbol{\pi}^{(1)}$ with probability π_1. For $\boldsymbol{\pi}^{(1)}$ and $\boldsymbol{\pi}^{(2)}$ to have coordinates in $[0, 1]$, it is required that

$$-\min(\frac{1 - \pi_i}{1 - \pi_1}, \frac{\pi_i}{\pi_1}) \leq w_i \leq \min(\frac{\pi_i}{1 - \pi_1}, \frac{1 - \pi_i}{\pi_1}), \quad i = 2, 3, \ldots, N,$$

but otherwise the weights can be freely chosen which makes the method very flexible.

The full procedure can conveniently be described in terms of sampling indicators and updated inclusion probability vectors $\boldsymbol{\pi}(t)$, $t = 0, 1, 2, \ldots$, with $\boldsymbol{\pi}(0) = \boldsymbol{\pi}$ as follows:

$$\begin{cases} P(I_t = 1) = \pi_t(t - 1) \\ \pi_i(t) = \pi_i(t - 1) - (I_t - \pi_t(t - 1))w_i(t), \ i > t \end{cases}, \quad t = 1, 2, \ldots, N.$$

An inclusion probability that once has become 0 or 1 will not be changed further. The weights $w_i(t)$, $i > t$, which should satisfy restrictions of the type above, may depend

on past outcomes $I_1, I_2, \ldots, I_{t-1}$ but not on I_t. If $\sum_{\mathcal{U}} \pi_i = n$ and $\sum_{i=t+1}^{N} w_i(t) = 1$ for each t, a design of fixed size n is obtained.

The bulldozer method was originally suggested by Bondesson and Thorburn (2008) as a method for real time sampling, i.e. suitable for cases when the sampler successively visits the units in the population or vice versa. Suitable choices of the weights are discussed by them. With appropriate weights, every design without replacement can be reproduced by the method.

The method can be extended as follows. Let

$$\pi_i(1) = \begin{cases} \pi_1 + (J - \pi_1)W & \text{if } i = 1 \\ \pi_i - (J - \pi_1)w_i & \text{else} \end{cases}, \quad \text{where} \quad J = \begin{cases} 1 & \text{with probability } \pi_1 \\ 0 & \text{with probability } 1 - \pi_1. \end{cases}$$

The choice $W = 1$ (and $J = I_1$) yields the original procedure.

If $\sum_{\mathcal{U}} \pi_i = n$ and $W = \sum_{i \geq 2} w_i$, then $\sum_{\mathcal{U}} \pi_i(1) = n$. Given a set of preliminary weights w_i, $i \geq 2$, with sum 1, there may be some weights that do not satisfy their restrictions. These weights can then be put equal to their bounds, and the procedure can be used with W equal to the sum of the modified weights. If in the first step the preliminary weights w_i are chosen to be proportional to $\pi_i(1 - \pi_i)$ (which equals 0 if π_i is 0 or 1), and in the later steps analogous choices of all weights are made, a sampling design of fixed size is obtained. The number of steps needed to complete the procedure is random. However, the design has high entropy and it is also invariant with respect to its eliminatory design.

4 The cube method for balanced sampling

An important consequence of the splitting method is the cube method, a method for balanced sampling presented by Deville and Tillé (2004) and by Tillé (2006, Chapter 8). In survey sampling there are often auxiliary variables, $z_1, z_2, \ldots, z_M$, with known values for all the units of the population and with known totals $Z_1, Z_2, \ldots, Z_M$ in particular. Then it may be desirable to have a sample $\mathbf{x}$ such that the Horvitz-Thompson estimates of these totals equal the known totals, i.e. such that $\hat{Z}_m = \sum_{i=1}^{N} \frac{z_{mi}}{\pi_i} x_i = Z_m$, for $m = 1, 2, \ldots, M$. With $a_{mi} = z_{mi}/\pi_i$, we then have the balancing linear restrictions

$$\sum_{i=1}^{N} a_{mi} x_i = Z_m, \quad m = 1, 2, \ldots, M,$$

or, in matrix language, $\mathbf{Ax} = \mathbf{Z}$, where now $\mathbf{x}$ and $\mathbf{Z}$ are considered as column vectors. Notice that $\mathbf{A}\boldsymbol{\pi} = \mathbf{Z}$. We may assume that the first such restriction is the fixed sample size restriction $\sum_{i=1}^{N} x_i = n$. Other restrictions could originate from some stratification. Then $\sum_{i \in \mathcal{S}_m} x_i = n_m,\ m = 1, 2, \ldots, M$, where the strata $\mathcal{S}_m \subseteq \mathcal{U}$ are disjoint and the numbers n_m are prescribed stratum sample sizes.

When the splitting method is applied it is necessary to let the random walk proceed within the intersection of the cube $\mathcal{C}$ and the affine space $\{\mathbf{x};\ \mathbf{Ax} = \mathbf{Z}\}$, i.e. so that

$$\begin{cases} \boldsymbol{\pi}(t) = \boldsymbol{\pi}(t-1) + \mathbf{e}(t) \\ \mathbf{Ae}(t) = \mathbf{0} \text{ and } E(\mathbf{e}(t)|\mathcal{F}_{t-1}) = \mathbf{0}. \end{cases}$$

Since $\mathbf{A}\boldsymbol{\pi}(t) = \mathbf{Z}$ for each t, necessarily $\mathbf{Ae}(t) = \mathbf{0}$ as required above, i.e. $\mathbf{e}(t)$ must belong to the nullspace of the matrix $\mathbf{A}$.

Due to the restrictions, it may happen that the procedure - *the flight phase* - is not capable of ending at a corner of a cube, i.e. at a sample. Instead it ends at a point with some coordinates strictly between 0 and 1. A special procedure - *the landing phase* - is then needed to find a close corner. As a consequence, the balancing restrictions are only approximately satisfied.

The cube method has found important practical applications, see Tillé (2006). To increase the entropy of the sampling design, the order of the units of the population can be randomized.

5 Final comments

The splitting method is a unifying method and its geometric approach casts new light on πps sampling. The method leads to new sampling procedures. Several well-known sampling procedures can also be presented in terms of splitting. Algorithms for many splitting methods can be found in Tillé's (2006) book. R language programs by Tillé and Matei (2005) are available for users of R.

For practical sampling, there are other efficient designs with high entropy. See e.g. Bondesson (2008b) or Bondesson *et al.* (2006). These articles treat conditional Poisson, Sampford, and Pareto sampling. The present paper is a modified version of the article Bondesson (2008a) in Encyclopedia of Statistical Sciences, Second edition.

References

Bondesson, L. (2008a). The splitting method for unequal probability sampling. In *Encyclopedia of Statistical Sciences*, Second edition (ed. N. Balakrishnan). Wiley, New York.

Bondesson, L. (2008b). Unequal probability sampling designs with high entropy. In *Encyclopedia of Statistical Sciences*, Second edition (ed. N. Balakrishnan). Wiley, New York.

Bondesson, L. and Thorburn, D. (2008). A list sequential sampling method suitable for real time sampling. *Scand. J. Statist.* (to appear).

Bondesson, L., Traat, I. & Lundqvist, A. (2006). Pareto sampling versus Sampford and conditional Poisson sampling. *Scand. J. Statist.* **33**, 699-720.

Brewer, K.R.W. (1975). A simple procedure for πpswor. *Austral. J. Statist.*, **17**, 166-172.

Brewer, K.R.W. and Hanif, M. (1983). *Sampling with Unequal Probabilities.* Lecture Notes in Statistics, No. 15. Springer-Verlag, New York.

Chaudhuri, A. and Vos, J.W.E. (1988). *Unified Theory and Strategies of Survey Sampling.* North-Holland, Amsterdam.

Deville, J-C. and Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika*, **85**, 89-101.

Deville, J-C. and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, **91**, 893-912.

Fan, C.T., Muller, M.E. and Rezucha, I. (1962). Development of sampling plans by using sequential (item by item) selection techniques and digital computers. *J. Amer. Statist. Assoc.*, **57**, 387-402.

Särndal, C-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling.* Springer-Verlag, New York.

Sunter, A.B. (1986). Solutions to the problem of unequal probability sampling without replacement. *Internat. Statist. Rev.*, **54**, 33-50.

Tillé, Y. (2006). *Sampling Algorithms.* Springer series in statistics. Springer science + Business media, Inc., New York.

Tillé, Y. and Matei, A. (2005). *The R package sampling.* The comprehensive R archive network, http://cran.r-project.org/, Manual of the contributed packages.