DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 4, 2015, pp. 649–672, http://dx.doi.org/10.1515/JOS-2015-0039

# Coordination of Conditional Poisson Samples

*Anton Grafström*[1] *and Alina Matei*[2]

Sample coordination seeks to maximize or to minimize the overlap of two or more samples. The former is known as positive coordination, and the latter as negative coordination. Positive coordination is mainly used for estimation purposes and to reduce data collection costs. Negative coordination is mainly performed to diminish the response burden of the sampled units. Poisson sampling design with permanent random numbers provides an optimum coordination degree of two or more samples. The size of a Poisson sample is, however, random. Conditional Poisson (CP) sampling is a modification of the classical Poisson sampling that produces a fixed-size $\pi ps$ sample. We introduce two methods to coordinate Conditional Poisson samples over time or simultaneously. The first one uses permanent random numbers and the list-sequential implementation of CP sampling. The second method uses a CP sample in the first selection and provides an approximate one in the second selection because the prescribed inclusion probabilities are not respected exactly. The methods are evaluated using the size of the expected sample overlap, and are compared with their competitors using Monte Carlo simulation. The new methods provide a good coordination degree of two samples, close to the performance of Poisson sampling with permanent random numbers.

*Key words:* Sample coordination; expected overlap; permanent random numbers; unequal probability sampling designs.

## 1. Introduction

Consider samples drawn successively or simultaneously from finite overlapping populations. Sample coordination seeks to create a dependence between these samples. This dependence leads to a maximization or to a minimization of the sample overlap. The former is known as positive coordination, and the latter as negative coordination. Positive coordination is mainly used for estimation purposes and to reduce data collection costs. Negative coordination is mainly performed to diminish the response burden of the sampled units. Changes in population definition pose a significant challenge in sample coordination. Thus, births, deaths, or splits of units frequently occur. To overcome this problem, an overall population defined as the union of the overlapping populations is generally used.

Sample coordination methods can be roughly divided into two categories: methods based on so-called Permanent Random Numbers (PRN) and non-PRN methods (for an overview, see for example Ernst 1999; Mach et al. 2006, and the references therein).

[1] Department of Forest Resource Management, Swedish University of Agricultural Sciences, SE-90183, Umeå, Sweden. Email: anton.grafstrom@slu.se
[2] Institute of Statistics, University of Neuchâtel, Rue Bellevaux 51, 2000, Neuchâtel and Institute of Pedagogical Research and Documentation (IRDP) Neuchâtel, Switzerland. Email: alina.matei@unine.ch
**Aknowledgments:** The authors wish to thank the Associate Editor and two referees for their valuable comments and suggestions that helped to improve the quality of the article significantly.

PRN methods are based on the following principle: assign to each unit in the overall population a uniform random number and use this number in all sample selections. Sample dependence is created based on the use of the same permanent random number of a unit over different surveys. For an overview of the main PRN methods, see Ohlsson (1995, 2000).

In the category of non-PRN schemes, methods following to Keytz (1951), Kish and Scott (1971) or Matei and Tillé (2005b) and methods based on mathematical programming (e.g., Raj 1968; Arthnari and Dodge 1981; Causey et al. 1985; Ernst and Ikeda 1995; Ernst 1996, 1998; Ernst and Paben 2002; Mach et al. 2006; Matei and Skinner 2009; Schiopu-Kratina et al. 2014) may be included.

The main difference between PRN methods and non-PRN ones lies in implementation difficulties. The PRN methods allow a good positive and negative coordination degree, and, in general, are easy to implement. However, they can provide a random sample size (e.g., Poisson sampling with PRN), and a lack of optimality (for example, the inclusion probabilities are not respected in each design) as shown in Mach et al. (2006) and Nedyalkova et al. (2008). For the non-PRN methods, and most specifically for approaches based on linear programming, Ernst (1999, 295) noted the following advantages (which can be generalized for all mathematical programming methods): "easy formulation, optimality, and flexibility in what to optimize". The main inconvenience of most of the mathematical programming methods is their implementation. In general, they can be employed only for small-size problems.

There is no perfect method of sample coordination that can be applied in all circumstances. As noted by Nedyalkova et al. (2009, 270), these methods "give partial but important solutions to real-life problems. However, one drawback of these methods is that they do not allow the important advances made in the domain of one-sample selection over the last decades to be integrated. For example, none of these allow one to use maximum fixed-size entropy sampling (see, e.g., Chen et al. 1994) or balanced sampling (Deville and Tillé 2004) as a cross-sectional sampling design."

We consider here the coordination of Conditional Poisson (CP) samples (or maximum fixed-size entropy samples) over time or simultaneously. As mentioned earlier, methods to coordinate CP samples have not yet been introduced in the literature. CP sampling has the maximum-entropy property in the class of fixed-size $\pi ps$ sampling designs with the same first-order inclusion probabilities, among other desirable properties, and has recently received considerable attention. We propose two methods here: the first one is a PRN method that uses the list-sequential implementation of two CP samples; the second one is a non-PRN method. The first method provides the coordination of two CP samples. The second method uses a first CP sample, and provides an approximate one in the second selection because the prescribed inclusion probabilities are not respected exactly.

The methods are evaluated using the size of the expected sample overlap, and are compared with their competitors. We focus on positive coordination, but negative coordination is also possible using the proposed methods. The article is organized as follows: Section 2 introduces the general framework and the notation; Section 3 provides a reminder of CP sampling and its main features. Sections 4 and 5 present the two proposed

methods. Section 6 shows the performance of the proposed methods compared with their competitors using Monte Carlo simulation. Finally, Section 7 concludes the article.

## 2. General Framework

Consider two finite overlapping populations. Let $U_1$ and $U_2$ denote the sets of labels of units in the two populations. Two sampling designs $p_1$ and $p_2$ of fixed size $n_1$ and $n_2$ are defined on $U_1$ and $U_2$, respectively. Let $S_1$ and $S_2$ be the sets of all possible samples defined by $p_1$ and $p_2$ on $U_1$ and $U_2$, respectively. Samples defined on $S_1$ are denoted $s_{1i}$, $i = 1, 2, \ldots, m$, while samples defined on $S_2$ are denoted $s_{2j}$, $j = 1, 2, \ldots, q$. Our general notation for samples is $s_1 \in S_1$ and $s_2 \in S_2$. We note that $\pi_{1k} = \sum_{s_1 \ni k, s_1 \in S_1} p_1(s_1), k \in U_1$ and $\pi_{2k} = \sum_{s_2 \ni k, s_2 \in S_2} p_2(s_2), k \in U_2$ are the first-order inclusion probabilities of unit $k$ in the two samples, respectively. For simplicity, let $U = \{1, \ldots, k, \ldots, N\}$ be the union of $U_1$ and $U_2$. Thus, for units $k \in U \backslash U_1$, we set $\pi_{1k} = 0$, while for $k \in U \backslash U_2$, we set $\pi_{2k} = 0$. An overall sampling design $p$ is defined on $S_1 \times S_2$, with marginal designs $p_1$ and $p_2$. The overall sampling design is said to be coordinated (see Cotton and Hesse 1992; Mach et al. 2006) if

$$p(s_{1i}, s_{2j}) = p_{ij} \neq p_1(s_{1i}) p_2(s_{2j}),$$

that is, if the two samples are not selected independently. The joint inclusion probability of unit $k$ in $s_1$ and $s_2$ is denoted

$$\pi_k^{1,2} = P(k \in s_1, k \in s_2) = \sum_{\substack{s_{1i} \cap s_{2j} \ni k \\ s_{1i} \in S_1, s_{2j} \in S_2}} p_{ij}.$$

Let $c_{ij}$ be the overlap size of samples $s_{1i}$ and $s_{2j}$

$$c_{ij} = |s_{1i} \cap s_{2j}|,$$

where $|A|$ denotes the cardinality of a set $A$. In general, the overlap size $c_{ij}$ is random. Let $c$ denote the random variable called 'overlap size'. A measure of the coordination degree between two samples is given by the expected value of $c$

$$E(c) = \sum_{i=1}^{m} \sum_{j=1}^{q} c_{ij} p_{ij} = \sum_{k \in U} \pi_k^{1,2}.$$

In positive coordination, the goal is to maximize $E(c)$, while in negative coordination, we want to minimize it. Bounds for $E(c)$ exist. They are determined by the Fréchet bounds of the joint inclusion probabilities $\pi_k^{1,2}$

$$\sum_{k \in U} \max(0, \pi_{1k} + \pi_{2k} - 1) \leq E(c) \leq \sum_{k \in U} \min(\pi_{1k}, \pi_{2k}). \tag{1}$$

Matei and Tillé (2005b) called the left-hand-part in (1) the Absolute Lower Bound (ALB) and the right-hand-part in (1) the Absolute Upper Bound (AUB). Ideally, in positive coordination we want to achieve the AUB, and in negative coordination the ALB.

Few methods achieve these bounds. In positive coordination, Poisson sampling with PRN (Brewer et al. 1972) applied in both selections provides an important property: $\pi_k^{1,2} = \min(\pi_{1k}, \pi_{2k})$, and thus the AUB is reached. The sample sizes are, however, random for $s_1, s_2$ and $s_1 \cap s_2$.

While all sample coordination methods seek to increase or decrease the sample overlap, there are different ways to measure the effectiveness of the positive or negative coordination (e.g., the size of the expected overlap or the expected load of a unit which is defined as the sum of its selection probabilities in the surveys). Consequently, there is no unique definition of optimality in sample coordination. We focus here on methods which try to reach the AUB.

## 3. Conditional Poisson Sampling

### 3.1. General Properties of CP Sampling

Let $0 \leq p_k \leq 1,\ k = 1, 2, \ldots, N$ be given parameters. In Poisson sampling an independent Bernoulli trial is performed for each unit $k$, so that unit $k$ is selected in the sample with probability $p_k$. Hence, Poisson sampling provides random sample size.

Conditional Poisson sampling is a fixed-size $\pi ps$ sampling design. It was introduced by Hájek (1964) as a modification of the classical Poisson sampling. Different implementations of CP sampling are available (see e.g., Tillé 2006; Bondesson et al. 2006). The initial implementation of CP sampling given by Hájek (1964, 1981) uses a rejective algorithm to obtain a sample of size $n$ as follows: draw Poisson samples (with parameters $p_k$) until we get a sample of size $n$, that is, we condition the Poisson design on the fixed sample size $n$. Usually, it is assumed that $\sum_{k=1}^N p_k = n$ because it maximizes the probability of obtaining samples of size $n$. The assumption $\sum_{k=1}^N p_k = n$ is, however, not restrictive. If it is not satisfied, the $p_k$s can be transformed to satisfy that condition (see eg., Broström and Nilsson 2000 or Tillé 2006, 89). Assume that $\sum_{k=1}^N p_k \neq n$, then transformed parameters $p'_k,\ k = 1, 2, \ldots, N$, with sum $n$ can be calculated. As long as

$$\frac{p'_k}{1 - p'_k} \propto \frac{p_k}{1 - p_k},$$

the design remains unchanged. We can let $p'_k/(1 - p'_k) = dp_k/(1 - p_k)$, which implies that

$$p'_k = \frac{dp_k}{1 - p_k + dp_k}, \tag{2}$$

and then find $d$ such that $\sum_{k=1}^N p'_k = n$. Practically, we can find $d$ by applying the Newton-Raphson method. If we start the Newton-Raphson method with $d = 1$, then usually only a few iterations are needed before convergence.

When implementing CP sampling of size $n$ with parameters $p_k$, $\sum_{k=1}^N p_k = n$, the true inclusion probabilities will only approximately equal the $p_k$s. Let $\pi_k^{CP(n)}$ denote the achieved inclusion probabilities for CP sampling of size $n$. These probabilities can

rapidly be calculated recursively with a formula proposed by Chen et al. (1994; see also Tillé 2006). The formula is

$$\pi_k^{CP(n)} = n \frac{p_k/(1-p_k)\cdot(1-\pi_k^{CP(n-1)})}{\sum_{\ell=1}^{N} p_\ell/(1-p_\ell)\cdot(1-\pi_\ell^{CP(n-1)})}, \tag{3}$$

and the start is given by $\pi_k^{CP(0)} = 0$, $k = 1, 2, \ldots, N$. Similarly, the second-order inclusion probabilities for CP sampling can be calculated recursively.

It is also possible to adjust the $p_k$s to obtain desired inclusion probabilities (Dupacová 1979; Chen et al. 1994; Deville 2000; Aires 2000b; Tillé 2006). Algorithms to obtain $p_k$ from given inclusion probabilities $\pi_k$ are given by Aires (2000b) and Tillé (2006, 83). Following Aires (2000b), an iterative algorithm can be applied. Let $\pi_k^{CP(n,t)}$ be the achieved inclusion probabilities derived by (3) with the parameters $p_k^t$, where $t$ denotes the current iteration of the algorithm, and let $p_k^0 = \pi_k$. Then, practically, only a few iterations of

$$p_k^t = p_k^{t-1} + (\pi_k - \pi_k^{CP(n,t-1)}), \tag{4}$$

are enough to find parameters $p_k^t$ that yield inclusion probabilities $\pi_k$.

CP sampling has an important property: it maximizes the entropy in the class of fixed-size *πps* designs with the same first-order inclusion probabilities. We recall that the entropy of a sampling design $\tilde{p}$ is defined as

$$I(\tilde{p}) = -\sum_{s\in S} \tilde{p}(s) \log\,(\tilde{p}(s)),$$

where $S = \{s \subset U | \tilde{p}(s) > 0\}$. There are at least three reasons for choosing a high-entropy sampling design (for a general discussion about the entropy of sampling designs, see also Grafström 2010):

1. The entropy is a measure of sample randomness: a higher entropy of the sampling design implies more randomness in sample selection.
2. High entropy is important for variance estimation. Tillé and Haziza (2010, 229) noted that: "The concept of entropy is useful in the context of variance estimation. When a sampling design has a high entropy, it is possible to obtain approximation of the second-order inclusion probabilities in terms of the first-order inclusion probabilities, which simplifies considerably the problem of variance estimation in the context of unequal probability sampling."
3. A higher entropy of a design results in a faster convergence to normal distribution of the Horvitz-Thompson estimator (Berger 1998).

All of these features make CP sampling a very attractive sampling design. Moreover, there are benefits to be gained from developing methods to coordinate CP samples.

### *3.2. List-Sequential Implementation of CP Sampling*

A CP sample can also be drawn using a list-sequential implementation. This method is recalled here since it is used afterwards in sample coordination.

List-sequential implementations of CP sampling can be found in for example, Chen and Liu (1997); Traat et al. (2004) and Tillé (2006). The units are sampled list sequentially starting from unit 1. Unit $k$ should be selected in the sample with an updated probability, here denoted by $\pi_k^{(k-1)}$. Thus, we select the unit $k$ in the sample if $r_k \leq \pi_k^{(k-1)}$, where $r_k$ is a random number from $U(0,1)$. The random number $r_k$ may be a permanent random number for unit $k$ (and it will be used in all coordination process). We assume that $r_1, \ldots, r_k, \ldots, r_N$ are independent.

Let $I_k \sim Bin(1, p_k)$, $k = 1, 2, \ldots, N$ be independent random variables, where $p_k$s are the Poisson parameters and $\sum_{k \in U} p_k = n$. The updated probabilities can be calculated as follows

$$\pi_k^{(k-1)} = P(I_k = 1 | S_k = n - n_{k-1}),$$

where $S_k = \sum_{\ell=k}^{N} I_\ell$, $n_k = \sum_{\ell=1}^{k} I_\ell$, and $n_0 = 0$. Note that $n_k$ is the realization of the random variable $\sum_{\ell=1}^{k} I_\ell$. The updated probabilities can be rewritten as

$$\pi_k^{(k-1)} = p_k \cdot \frac{P(S_{k+1} = n - n_{k-1} - 1)}{P(S_k = n - n_{k-1})},$$

where $S_{N+1} = 0$. The probabilities $P(S_k = a)$ for given $k$ and $a$ can easily be calculated recursively. The start is given by $P(S_N = 0) = 1 - p_N$ and $P(S_N = 1) = p_N$. Then, for $k = N-1, N-2, \ldots, 1$ and $a = 0, 1, \ldots, N-k+1$, we have

$$P(S_k = a) = p_k P(S_{k+1} = a - 1) + (1 - p_k) P(S_{k+1} = a), \text{ if } a > 0,$$

and

$$P(S_k = a) = (1 - p_k) P(S_{k+1} = a), \text{ if } a = 0.$$

If the population is very large, the recursions may take some time. Using this method, we can calculate the updated probabilities $\pi_k^{(k-1)}$, for $k = 1, 2, \ldots, N$, and directly get a sample.

## 4. Coordination of CP Samples Using the List-Sequential Implementation

A first method coordinating CP samples is based on the list-sequential implementation presented in Subsection 3.2. To coordinate two CP samples with inclusion probabilities $\pi_{1k}$ and $\pi_{2k}$, $k = 1, 2, \ldots, N$, we use the algorithm given by Expression (4) to find the corresponding Poisson parameters $p_{1k}$ and $p_{2k}$, respectively. We then apply the list-sequential method with the permanent random numbers $r_k$ in each selection. Even though it is logical to try to coordinate CP samples in this manner, the approach seems to be new. In fact, any design with a list-sequential implementation can easily be coordinated by the use of PRN.

**Remark 1** Negative coordination can be achieved using the list-sequential method. For negative coordination of two samples, antithetic random numbers $r_k^* = 1 - r_k$ can be used in the second selection. For $\beta > 2$ samples, new random numbers can be constructed by shifting the PRN by an amount $\alpha$ to the right before the selection of each sample different from the first one: $r_k + \alpha$. A possible choice of $\alpha$ is the inverse of the number of samples to

coordinate (see Ohlsson 2000). Ohlsson (2000, 257) comments for Poisson sampling with PRN "if $\beta$ samples are to be negatively coordinated, the choice $\alpha = 1/\beta$ should give a small sample overlap. In particular, if the target inclusion probabilities . . . are less than $1/\beta$ for all units $i$ in all $\beta$ designs, the expected overlap is 0." The same idea can be used to negatively coordinate $\beta > 2$ CP samples using the list-sequential method. If $r_k + \alpha$ is larger than 1, we can replace it by $(r_k + \alpha)$ mod 1, where mod is the modulo operator.

**Remark 2** Consider the case of positive coordination of two samples, and denote the selection probabilities in the list-sequential method by $\pi_{1k}^{(k-1)}$ and $\pi_{2k}^{(k-1)}$, respectively. It is interesting to quantify the size of the expected overlap of two samples drawn using the list-sequential method, and compare it to the AUB. The expected overlap of two samples drawn using the list-sequential method with PRN depends on the random variables $\pi_{1k}^{(k-1)}$ and $\pi_{2k}^{(k-1)}$. It is given by

$$\sum_{k\in U}\pi_k^{1,2} = \sum_{k\in U} P\left(r_k < \pi_{1k}^{(k-1)}, r_k < \pi_{2k}^{(k-1)}\right) = \sum_{k\in U} P\left(r_k < \min\left(\pi_{1k}^{(k-1)}, \pi_{2k}^{(k-1)}\right)\right). \quad (5)$$

However, it is difficult to quantify it exactly because the same permanent random numbers are used in the selection of the two samples. Consequently, $\pi_{1k}^{(k-1)}$ and $\pi_{2k}^{(k-1)}$ are dependent random variables. Thus, the method performance is studied only empirically in Section 6.

## 5. An Approximate Method to Coordinate CP Samples

### *5.1. Description of the Method*

We suggest a new method coordinating two samples that does not use permanent random numbers, but instead uses updated parameters for the second selection. In the second selection the proposed method provides a new sampling design which is approximately a CP sampling. Recall that $U = U_1 \bigcup U_2$, $\pi_{1k} = 0$ if $k \in U \backslash U_1$ and $\pi_{2k} = 0$ if $k \in U \backslash U_2$, for all $k = 1, \ldots, N$.

In the first selection, we select a CP sample $s_1$ of size $n_1$ with inclusion probabilities $\pi_{1k}$, $k = 1, 2, \ldots, N$, using any suitable method to obtain a CP sample. Given $s_1$, we obtain the following conditional probabilities

$$P(k \in s_2 | k \notin s_1) = \left(\pi_{2k} - \pi_k^{1,2}\right)/(1 - \pi_{1k}) \, , P(k \in s_2 | k \in s_1) = \pi_k^{1,2}/\pi_{1k}, \quad (6)$$

assuming that $0 < \pi_{1k} < 1$. By letting $\pi_k^{1,2} = \min(\pi_{1k}, \pi_{2k})$ in (6), we compute the following updated parameters $p_{2k|s_1}$, $k = 1, 2, \ldots, N$:

- if $\pi_{1k} \leq \pi_{2k}$, then

$$p_{2k|s_1} = \begin{cases} 1 & \text{if } k \in s_1 \\ (\pi_{2k} - \pi_{1k})/(1 - \pi_{1k}) & \text{if } k \notin s_1 \end{cases},$$

- if $\pi_{1k} > \pi_{2k}$, then

$$p_{2k|s_1} = \begin{cases} \pi_{2k}/\pi_{1k} & \text{if } k \in s_1 \\ 0 & \text{if } k \notin s_1 \end{cases}.$$

In the second selection we select a CP sample $s_2$ of size $n_2$ with the parameters $p_{2k|s_1}$. The updated parameters are only used for units $k \in U_1$; for new units $k \in U \backslash U_1$, we let $p_{2k|s_1} = \pi_{2k}$. If we achieve conditional inclusion probabilities equal to these parameters, we obtain the prescribed inclusion probabilities $\pi_{2k}$. Moreover, $\pi_k^{1,2} = \min(\pi_{1k}, \pi_{2k})$ and the expected overlap is maximized (the AUB is achieved).

The parameters $p_{2k|s_1}$ cannot be used as inclusion probabilities for a fixed-size design because they do not in general sum to $n_2$ for a given $s_1$. Only the sum of the expected value of the $p_{2k|s_1}$ equals $n_2$ because

$$\begin{aligned} E(p_{2k|s_1}) &= \min\left(1, \frac{\pi_{2k}}{\pi_{1k}}\right) E(I_{1k}) + \max\left(0, \frac{\pi_{2k} - \pi_{1k}}{1 - \pi_{1k}}\right)(1 - E(I_{1k})) \\ &= \min\left(1, \frac{\pi_{2k}}{\pi_{1k}}\right) \pi_{1k} + \max\left(0, \frac{\pi_{2k} - \pi_{1k}}{1 - \pi_{1k}}\right)(1 - \pi_{1k}) \\ &= \pi_{2k}, \end{aligned}$$

where $E(.)$ is the expectation operator, and $I_{1k}$ is the indicator variable of unit $k$ for sample $s_1$ ($I_{1k} = 1$ if $k \in s_1$ and 0 otherwise). Thus it is impossible to achieve inclusion probabilities equal to these parameters for a given $s_1$ if only samples of size $n_2$ are accepted.

If $p_{2k|s_1}$ are used as parameters in the rejective implementation of CP sampling, we can maximize the probability of obtaining a sample of size $n_2$ by using transformed parameters with sum $n_2$ (see Subsection 3.1). However, Expression (2) can provide unchanged parameters if their values are 0 or 1.

Some situations may arise where it is impossible to draw a sample $s_2$ using the parameters $p_{2k|s_1}$. Consider, for example, the case where $N = 6, n_1 = n_2 = 3$, $\boldsymbol{\pi}_1 = (0.3, 0.3, 0.3, 0.7, 0.7, 0.7)'$, $\boldsymbol{\pi}_2 = (0.4, 0.4, 0.4, 0.4, 0.4, 1)'$ and the sample $s_1$ is $\{1, 2, 3\}$. The parameters $p_{2k|s_1}$ are $(1, 1, 1, 0, 0, 1)'$, and they do not allow the selection of a sample $s_2$ of size 3.

In these unlikely situations, it will be impossible to achieve a sample of size $n_2$ using the parameters $p_{2k|s_1}$ because either more than $n_2$ of the parameters equal 1 or more than $N - n_2$ equal 0. In such cases, we suggest the following modification to the parameters. If there are more than $n_2$ of the $p_{2k|s_1}$ that equal 1, we use

$$p^*_{2k|s_1} = \begin{cases} 0 & \text{if } p_{2k|s_1} < 1 \\ 1 & \text{if } \pi_{2k} = 1 \\ \dfrac{n_2 - |\{j:\ \pi_{2j} = 1\}|}{|\{j:\ p_{2j|s_1} = 1,\ \pi_{2j} < 1\}|} & \text{otherwise} \end{cases}.$$

If there are more than $N - n_2$ of the $p_{2k|s_1}$ that equal 0, we use

$$p^*_{2k|s_1} = \begin{cases} 1 & \text{if } p_{2k|s_1} > 0 \\ 0 & \text{if } \pi_{2k} = 0 \\ \dfrac{n_2 - |\{j:\, p_{2j|s_1} > 0\}|}{|\{j:\, p_{2j|s_1} = 0,\ \pi_{2j} > 0\}|} & \text{otherwise} \end{cases},$$

where $|\{\cdot\}|$ is the size of $\{\cdot\}$. The new parameters $p^*_{2k|s_1}$ sum to $n_2$.

**Example:** To exemplify the proposed method, we consider the same example as before ($N = 6, n_1 = n_2 = 3$). Consider that $s_1$ is $\{4, 5, 6\}$. The parameters $p_{2k|s_1}$ are 0.1428571, 0.1428571, 0.1428571, 0.5714286, 0.5714286, 1. Their sum is $2.571429 \neq 3$. The parameters $p_{2k|s_1}$ are transformed because they do not sum to $n_2 = 3$. The transformation is given by Expression (2). The transformed parameters are 0.2117249, 0.2117249, 0.2117249, 0.6824127, 0.6824127, 1, which sum to 3. These parameters are finally used to draw $s_2$. Here it is not necessary to adjust the parameters to obtain $p^*_{2k|s_1}$ because there are no more than $n_2$ parameters that equal 1 and no more than $N - n_2$ that equal 0.

If $s_1$ is $\{1, 2, 3\}$, one obtains the parameters $\boldsymbol{p}_{2|s_1} = (1, 1, 1, 0, 0, 1)'$. The transformation given by Expression (2) provides the same parameters. They are modified to obtain $\boldsymbol{p}^*_{2|s_1} = (0.6666667, 0.6666667, 0.6666667, 0, 0, 1)'$, with $\sum_{k \in U} p^*_{2k|s_1} = 3$. These parameters are finally used to draw $s_2$ using the rejective method.

**Remark 3** The use of the parameters $p_{2k|s_1}$ corresponds to the following rejective algorithm for the second sample (positive coordination). Perform Poisson sampling until the sample size is $n_2$ as follows: if unit $k \in s_1$, draw a uniform random number $u_k \sim U(0, \pi_{1k})$; otherwise, draw a uniform random number $u_k \sim U(\pi_{1k}, 1)$. In both cases, select the unit $k$ in $s_2$ if $u_k < \pi_{2k}$. Similarly, for the negative coordination of two samples, perform Poisson sampling until the sample size is $n_2$ as follows: if unit $k \in s_1$, draw a uniform random number $u_k \sim U(1 - \pi_{1k}, 1)$; otherwise, draw a uniform random number $u_k \sim U(0, 1 - \pi_{1k})$. In both cases, select the unit $k$ in $s_2$ if $u_k < \pi_{2k}$. The algorithm for the negative coordination results from using $\pi_k^{1,2} = \max(0, \pi_{1k} + \pi_{2k} - 1)$ in the conditional probabilities in (6).

**Remark 4** Any design that achieves the prescribed inclusion probabilities for second selection and reaches the AUB must respect the following conditional inclusion probabilities for all $k \in U$

$$P(k \in s_2 | k \notin s_1) = \begin{cases} (\pi_{2k} - \pi_{1k})/(1 - \pi_{1k}) & \text{if } \pi_{1k} \leq \pi_{2k} \\ 0 & \text{if } \pi_{1k} > \pi_{2k} \end{cases}$$

and

$$P(k \in s_2 | k \in s_1) = \begin{cases} 1 & \text{if } \pi_{1k} \leq \pi_{2k} \\ \pi_{2k}/\pi_{1k} & \text{if } \pi_{1k} > \pi_{2k} \end{cases}.$$

In our approximate method we apply the rejective algorithm with these probabilities as parameters. We have several situations where we get the exact marginal CP design at

second selection. The most trivial example is when $\pi_{2k} = \pi_{1k}$ for all $k$. Then $s_2$ is always equal to $s_1$, and since $s_1$ is a CP sample, so is $s_2$. Another example is when $\pi_{1k}$ is 0 or 1 for all $k$. Then the marginal design for the second sample is CP with parameters $\pi_{2k}$. Finally, we get the marginal CP design for the second selection if $\pi_{2k}$ is equal to $\pi_{1k}$, 0 or 1, for all $k$. These results indicate that if $\boldsymbol{\pi}_2$ is close to $\boldsymbol{\pi}_1$, the marginal design for the second selection must be very close to CP. Also if $\boldsymbol{\pi}_2$ is very far from $\boldsymbol{\pi}_1$ (i.e., the elements of $\boldsymbol{\pi}_2$ are close to 0 or 1), then we also get close to CP design.

### 5.2. *Examples*

The examples below are used to compare the first- and second-order inclusion probabilities of the proposed design (used to draw $s_2$) to those of the CP design. The first and second-order inclusion probabilities of the proposed design are denoted by $\tilde{\pi}_{2k}$ and $\tilde{\pi}_{2k\ell}$, respectively. However, they cannot be directly computed. In what follows, the probabilities $\tilde{\pi}_{2k}$ and $\tilde{\pi}_{2k\ell}$ have been estimated using $10^6$ simulated samples (see Thompson and Wu 2008 for a discussion about the number of simulated samples used to obtain estimated inclusion probabilities) and are rounded to four decimal places. The estimated inclusion probabilities are denoted by $\hat{\tilde{\pi}}_{2k}$ and $\hat{\tilde{\pi}}_{2k\ell}$, respectively, while the prescribed first- and second-order inclusion probabilities are denoted by $\pi_{2k}$ and $\pi_{2k\ell}$, respectively. The R software (version 3.0.1) was used to compute $\hat{\tilde{\pi}}_{2k}$ and $\hat{\tilde{\pi}}_{2k\ell}$. The same number of $10^6$ simulated samples has been used to compute $\hat{\tilde{\pi}}_{2k}$ and $\hat{\tilde{\pi}}_{2k\ell}$ regardless of the population size or characteristics. Another method is to use an adaptive algorithm for estimating the inclusion probabilities, where the number of replications is determined on the basis of the stability of the Horvitz-Thompson estimates and their precision (see Fattorini 2009).

**Example 1**: A population of size $N = 5$ is used to show that the proposed design gives inclusion probabilities close to those of the CP design even for small populations. Let $\boldsymbol{\pi}_1 = (0.1, 0.2, 0.3, 0.5, 0.9)'$, $\boldsymbol{\pi}_2 = (0.7, 0.2, 0.2, 0.4, 0.5)'$, $n_1 = n_2 = 2$. The corresponding Poisson parameters that give these inclusion probabilities (rounded to four decimal places) are

$$\boldsymbol{p}_1 = (0.1328, 0.2387, 0.3253, 0.4595, 0.8438)'$$
$$\boldsymbol{p}_2 = (0.6430, 0.2354, 0.2354, 0.4066, 0.4797)'.$$

The prescribed and the estimated inclusion probabilities are given in Tables 1, 2 and 3. The highest absolute difference between $\hat{\tilde{\pi}}_{2k}$ and $\pi_{2k}$ is about 0.02, while for the second-order inclusion probabilities $\hat{\tilde{\pi}}_{2k\ell}$ and $\pi_{2k\ell}$ it is about 0.07.

*Table 1. Prescribed $\pi_{2k}$ and estimated $\hat{\tilde{\pi}}_{2k}$ in Example 1; $\hat{\tilde{\pi}}_{2k}$ are computed using $10^6$ simulated samples*

| $k$ | $\pi_{2k}$ | $\hat{\tilde{\pi}}_{2k}$ |
|---|---|---|
| 1 | 0.7 | 0.6971 |
| 2 | 0.2 | 0.1992 |
| 3 | 0.2 | 0.2114 |
| 4 | 0.4 | 0.4165 |
| 5 | 0.5 | 0.4759 |

*Table 2. Matrix of prescribed $\pi_{2k\ell}$ in Example 1*

| $k\backslash\ell$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0.7000 | 0.0969 | 0.0969 | 0.2158 | 0.2903 |
| 2 | 0.0969 | 0.2000 | 0.0166 | 0.0369 | 0.0496 |
| 3 | 0.0969 | 0.0166 | 0.2000 | 0.0369 | 0.0496 |
| 4 | 0.2158 | 0.0369 | 0.0369 | 0.4000 | 0.1104 |
| 5 | 0.2903 | 0.0496 | 0.0496 | 0.1104 | 0.5000 |

**Example 2**: **MU284 static population** We consider the so-called 'MU284 population' of 284 Swedish municipalities presented in Appendix B of Särndal et al. (1992). The regional REG variable, giving eight strata of sizes between 15 and 56, is used. We consider samples of sizes $n_1 = 10, n_2 = 6$ drawn from the stratum number 2 (REG $= 2$), with $N = 48$. No births or deaths are assumed. The inclusion probabilities $\pi_{1k}$ are computed using the variable P75 (population in 1975 in thousands), and $\pi_{2k}$ using the variable P85 (population in 1985 in thousands). Due to lack of space, the probabilities $\hat{\tilde{\pi}}_{2k}$, $\pi_{2k}$ and the matrices of $\hat{\tilde{\pi}}_{2k\ell}$ and $\pi_{2k\ell}$ are not shown here. For this example, the estimated $\hat{\tilde{\pi}}_{2k}$s are very close to those of the CP design (the largest absolute difference between them is about 0.006). The highest absolute difference between $\hat{\tilde{\pi}}_{2k\ell}$ and $\pi_{2k\ell}$ is about 0.006.

**Example 3**: **MU284 dynamic population** We consider the same stratum 2 from the MU284 population as before, but 50% of the units are new on the second occasion (births), and 50% of the units change stratum (deaths). We have considered that the births were initially in the third stratum. Thus, 24 units have been randomly drawn from the third stratum using simple random sampling without replacement; these units represent the births for the second stratum. Similarly, 24 units have been randomly drawn from the second stratum using simple random sampling without replacement; these units represent the deaths for the second stratum. The number of persistent units in the two occasions is 24. The overall population (called 'MU284 dynamic population') is formed by the persistents, births and deaths; its size is 72. Samples of sizes $n_1 = 10, n_2 = 6$ respectively are drawn from this population. The inclusion probabilities $\pi_{1k}$ and $\pi_{2k}$ are computed using the same variables as in Example 2. The highest absolute difference between $\hat{\tilde{\pi}}_{2k}$ and $\pi_{2k}$ is about 0.02; the same value is obtained as the highest absolute difference between $\hat{\tilde{\pi}}_{2k\ell}$ and $\pi_{2k\ell}$. For the same example, but where $n_1 = 15$ and $n_2 = 20$, the highest absolute difference between $\hat{\tilde{\pi}}_{2k}$ and $\pi_{2k}$ is about 0.011; the highest absolute difference between $\hat{\tilde{\pi}}_{2k\ell}$ and $\pi_{2k\ell}$ is about 0.015. Finally, for $n_1 = 30$ and $n_2 = 25$ we obtained the highest absolute difference between $\hat{\tilde{\pi}}_{2k}$ and $\pi_{2k}$ equal to 0.008, while for $\hat{\tilde{\pi}}_{2k\ell}$ and $\pi_{2k\ell}$ it equals to 0.010.

*Table 3. Matrix of estimated $\hat{\tilde{\pi}}_{2k\ell}$ in Example 1 computed on $10^6$ simulated samples*

| $k\backslash\ell$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0.6971 | 0.1159 | 0.1220 | 0.2475 | 0.2116 |
| 2 | 0.1159 | 0.1992 | 0.0069 | 0.0164 | 0.0600 |
| 3 | 0.1220 | 0.0069 | 0.2114 | 0.0154 | 0.0671 |
| 4 | 0.2475 | 0.0164 | 0.0154 | 0.4165 | 0.1372 |
| 5 | 0.2116 | 0.0600 | 0.0671 | 0.1372 | 0.4759 |

**Example 4**: We use an artificial population (called 'artificial population I') of size $N = 1,000$. No births and no deaths are assumed. Samples of sizes $n_1 = 100$ and $n_2 = 250$, respectively, are drawn. The inclusion probabilities $\pi_{1k}$ and $\pi_{2k}$ are randomly generated using the $U(0,1)$ distribution and are normalized to sum to 100 and 250, respectively. The highest absolute difference between $\hat{\tilde{\pi}}_{2k}$ and $\pi_{2k}$ is about 0.0015; the highest absolute difference between $\hat{\tilde{\pi}}_{2k\ell}$ and $\pi_{2k\ell}$ is about 0.0016. For a similar example, but where $N = 2000$, $n_1 = 200$ and $n_2 = 500$, the highest absolute difference between the estimated and the prescribed first-order inclusion probabilities is about 0.0017; the same value is obtained as the highest absolute difference between $\hat{\tilde{\pi}}_{2k\ell}$ and $\pi_{2k\ell}$.

Extensive simulations (not shown here) performed on large-size populations and different vectors $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ suggest that the differences between the inclusion probabilities of the two designs vanish when population and sample sizes are growing.

### 5.3. *Estimation Using the Proposed Sampling Design*

As noted in Subsection 5.1, the proposed sampling design applied in the second occasion is only approximately a CP sampling. This section contains different simulation studies used to compare the performance of the proposed design and CP sampling design in estimations.

Let $y$ be the variable of interest and $Y = \sum_{k \in U_2} y_k$, where $y_k$ is the value of the variable of interest taken on unit $k$. To check the impact of the proposed sampling design on estimations, we focus on the Horvitz-Thompson (HT) estimator of $Y$

$$\hat{Y}_{HT} = \sum_{k \in s_2} y_k / \tilde{\pi}_{2k}. \tag{7}$$

Since the inclusion probabilities $\tilde{\pi}_{2k}$ cannot be directly computed, this estimator cannot be used. Following Fattorini (2006, 270), a "natural modification" of (7) is

$$\hat{Y}_m = \sum_{k \in s_2} y_k / \hat{\tilde{\pi}}_{2k} \tag{8}$$

where $\hat{\tilde{\pi}}_{2k}$, the estimated value of $\tilde{\pi}_{2k}$, was computed using Monte Carlo simulation and $m = 10^6$ runs. The estimator $\hat{\tilde{Y}}_m$ converges almost surely to $\hat{Y}_{HT}$ as $m$ increases. A first option is to use Estimator (8), drawing $s_2$ using the proposed design. A second option is to draw $s_2$ using the proposed design but to compute the HT estimator using $\pi_{2k}$ instead of $\tilde{\pi}_{2k}$. The second option is determined by the closenesses between $\tilde{\pi}_{2k}$ and $\pi_{2k}$, as shown in Subsection 5.2, for large $N$ and $n_2$.

To assess the performance of the previous estimators under the proposed sampling design, one artificial population and one real population were used. For both populations, we selected $m = 10^5$ samples according to the CP sampling and the proposed sampling, respectively.

In the artificial population of size $N = 200$, no births or deaths are assumed. Let $x_1$ and $x_2$ be the size variables in the first and second design, respectively. To underline a potential change of the size variables over time, we used $x_{1k} = k$ and $x_{2k} = x_{1k} + u_k$, where $x_{1k}, x_{2k}$ are the values of $x_1, x_2$ taken on unit $k$, respectively, and $u_k \sim U(0,1)$ are independent random variables, $k = 1, \ldots, N$. The model used to generate the variable of interest was $y_k = 5x_{2k}(1 + \varepsilon_k)$, where $\varepsilon_k \sim N(0, \sigma^2 = 0.4^2)$ are independent random variables.

The correlation between $y$ and $x_2$ was about 0.8. The first-order inclusion probabilities $\pi_{1k}$ and $\pi_{2k}$ were proportional to $x_{1k}$ and $x_{2k}$, respectively.

Two different sample size settings were used. In the first setting, $n_1 = 15$ and $n_2 = 20$, while in the second one $n_1 = 50$ and $n_2 = 60$. For each setting and in each simulation run, CP samples of sizes $n_1$ and $n_2$ were drawn using the rejective method. Additionally, samples of size $n_2$ were also drawn using the proposed design. On the samples drawn using the proposed design, two estimators were computed: the HT estimator using $\pi_{2k}$ instead of $\tilde{\pi}_{2k}$ and Estimator (8). On the CP samples of size $n_2$, the HT estimator (using $\pi_{2k}$) was computed. In all tables of this section, the column 'Prob.' denotes the inclusion probabilities used in estimations, while the column 'Design' gives the design.

The performance of the HT-type estimators were compared using the percent absolute relative bias (ARB)

$$ARB_{\hat{Y}} = 100 \times \frac{|E_{sim}(\hat{Y}) - Y|}{Y},$$

and the empirical root mean square error (RMSE)

$$RMSE_{\hat{Y}} = \sqrt{(E_{sim}(\hat{Y}) - Y)^2 + var_{sim}(\hat{Y})},$$

where $\hat{Y}$ is a generic estimator, $\hat{Y}_{\tilde{m}}$ is the estimator computed on the $\tilde{m}$th simulated sample, $E_{sim}(\hat{Y}) = \frac{1}{m}\sum_{\tilde{m}=1}^{m} \hat{Y}_{\tilde{m}}$ and $var_{sim}(\hat{Y})$ is the variance of $\hat{Y}_{\tilde{m}}$ computed as $\sum_{\tilde{m}=1}^{m} \left(\hat{Y}_{\tilde{m}} - E_{sim}(\hat{Y})\right)^2/(m-1)$.

The results for the artificial population are provided in Table 4. They show small values of the ARB (less than or equal to 0.01%) for the estimators using $\pi_{2k}$ and $\hat{\tilde{\pi}}_{2k}$ and computed on the samples drawn using the proposed design. In both settings, the RMSE values reported in Table 4 are similar for all estimators. It is worth noting that the proposed design using $\pi_{2k}$ and $\pi_{2k\ell}$ provides similar ARB and RMSE values as compared to those obtained by CP sampling.

*Table 4. Percent absolute relative bias and root mean square error of the HT type estimators computed using CP sampling and the proposed design, $m = 10^5$ simulation runs; artificial population, $N = 200$*

| Setting | $n_1 = 15$, | $n_2 = 20$ | |
|---|---|---|---|
| Design | Prob. | $ARB_{\hat{Y}}$ (%) | $RMSE_{\hat{Y}}$ |
| CP sampling | $\pi_{2k}$ | 0.01 | 7875.13 |
| Proposed design | $\pi_{2k}$ | $<0.01$ | 7912.50 |
| Proposed design | $\hat{\tilde{\pi}}_{2k}$ | $<0.01$ | 7906.24 |
| Setting | $n_1 = 50$, | $n_2 = 60$ (%) | |
| CP sampling | $\pi_{2k}$ | $<0.01$ | 2932.30 |
| Proposed design | $\pi_{2k}$ | $<0.01$ | 2932.52 |
| Proposed design | $\hat{\tilde{\pi}}_{2k}$ | $<0.01$ | 2931.15 |

The real population is the 'MU284 dynamic population' described in Example 3. Births and deaths are assumed. The overall population size is $N = 72$. Two different sample size settings were used. In the first setting, $n_1 = 10$ and $n_2 = 6$, while in the second one $n_1 = 30$ and $n_2 = 25$. The same estimators as described for the artificial population were used. The results for the real population are provided in Table 5. In Table 5, the largest value of the ARB (about 0.17%) is shown by the HT-type estimator using $\pi_{2k}$ and the proposed design. This value can be explained by a larger difference between $\tilde{\pi}_{2k}$ and $\pi_{2k}$ for small samples (here $n_2 = 6$). The other values of the ARB are similar (about 0.01%) for all estimators in both settings. The values of the RMSE are very close for all estimators. An exception is Estimator (2) in the first setting, which turned out to be more efficient than the HT estimator computed using the CP sampling in terms of RMSE.

A second simulation study focuses on variance estimation. A variance estimator of the Horvitz-Thompson estimator is the Sen-Yates-Grundy (SYG) estimator

$$\widehat{var}(\hat{Y}_{HT})_{SYG} = \sum_{k \in s_2} \sum_{\ell \in s_2, \ell \neq k} \frac{\tilde{\pi}_{2k}\tilde{\pi}_{2\ell} - \tilde{\pi}_{2k\ell}}{\tilde{\pi}_{2k\ell}} \left( \frac{y_k}{\tilde{\pi}_{2k}} - \frac{y_\ell}{\tilde{\pi}_{2\ell}} \right)^2. \tag{9}$$

This estimator is unbiased when the sample size is fixed, provided that all the second-order inclusion probabilities are strictly positive. A disadvantage of the SYG variance estimator is the use of the second-order inclusion probabilities. It can also be very unstable because of the term $\tilde{\pi}_{2k\ell}^{-1}$ in (9) (see Haziza et al. 2008, 93). Another variance estimator using only the first-order inclusion probabilities proposed by Rosén (1991) and found to perform well under a high-entropy sampling design was used in our simulations

$$\widehat{var}(\hat{Y}_{HT})_{Ros} = \frac{n_2}{n_2 - 1} \sum_{k \in s_2} (1 - \tilde{\pi}_{2k}) \left( \frac{y_k}{\tilde{\pi}_{2k}} - a \right)^2, \tag{10}$$

*Table 5. Percent absolute relative bias and root mean square error of the HT type estimators computed using CP sampling and the proposed design, $m = 10^5$ simulation runs; MU284 dynamic population, $N = 72$*

| Setting | $n_1 = 10$, | $n_2 = 6$ | |
|---|---|---|---|
| Design | Prob. | $ARB_{\hat{Y}}$ (%) | $RMSE_{\hat{Y}}$ |
| CP sampling | $\pi_{2k}$ | 0.01 | 340.53 |
| Proposed design | $\pi_{2k}$ | 0.17 | 334.33 |
| Proposed design | $\tilde{\pi}_{2k}$ | 0.01 | 285.98 |
| Setting | $n_1 = 30$, | $n_2 = 25$ | |
| Design | Prob. | $ARB_{\hat{Y}}$ (%) | $RMSE_{\hat{Y}}$ |
| CP sampling | $\pi_{2k}$ | $<0.01$ | 69.80 |
| Proposed design | $\pi_{2k}$ | $<0.01$ | 69.14 |
| Proposed design | $\tilde{\pi}_{2k}$ | $<0.01$ | 69.00 |

where

$$a = \frac{\sum_{\ell \in s_2} y_\ell \dfrac{1 - \tilde{\pi}_{2\ell}}{\tilde{\pi}_{2\ell}^2} log\,(1 - \tilde{\pi}_{2\ell})}{\sum_{\ell \in s_2} \dfrac{1 - \tilde{\pi}_{2\ell}}{\tilde{\pi}_{2\ell}} log\,(1 - \tilde{\pi}_{2\ell})}.$$

As in the case of Estimator (7), we modify Estimators (9) and (10) using $\hat{\tilde{\pi}}_{2k}$ and $\hat{\tilde{\pi}}_{2k\ell}$ instead of $\tilde{\pi}_{2k}$ and $\tilde{\pi}_{2k\ell}$, respectively.

The variance estimators were computed on the same simulated samples as the HT-type estimators. On the CP samples of size $n_2$, they were computed using $\pi_{2k}$ and $\pi_{2k\ell}$. The variance estimators are compared using the percent absolute ratio of bias and the empirical root mean square. The ARB is now defined as

$$ARB_{\widehat{var}} = 100 \times \frac{|E_{sim}(\widehat{var}) - V|}{V},$$

where $\widehat{var}$ is a generic variance estimator, $\widehat{var}_{\tilde{m}}$ is the estimator computed on the $\tilde{m}$th simulated sample, $E_{sim}(\widehat{var}) = \frac{1}{m}\sum_{\tilde{m}=1}^{m} \widehat{var}_{\tilde{m}}$ and $V$ is the true variance of the HT-type estimator. The values of $V$ were computed through simulation using another set of $10^5$ simulated samples. Similarly, the RMSE is now defined as

$$RMSE_{\widehat{var}} = \sqrt{(E_{sim}(\widehat{var}) - V)^2 + var_{sim}(\widehat{var})},$$

where $var_{sim}(\widehat{var}) = \sum_{\tilde{m}=1}^{m} (\widehat{var}_{\tilde{m}} - E_{sim}(\widehat{var}))^2/(m-1)$.

Table 6 shows the values of the ARB and RMSE of the SYG and Rosén variance estimators for the artificial population. In terms of absolute relative bias, all three SYG variance estimators perform equally in both settings. The largest bias is displayed by the SYG estimators using the proposed design. The three Rosén estimators show larger but similar values of the ARB. The magnitude of the values of the ARB for Rosén estimators

*Table 6. Percent absolute relative bias and root mean square error of the variance estimators using CP sampling and the proposed design, $m = 10^5$ simulation runs; artificial population, $N = 200$*

| Setting | $n_1 = 15$, | $n_2 = 20$ | | | |
|---|---|---|---|---|---|
| Design | Prob. | $ARB_{SYG}$ (%) | $RMSE_{SYG}$ | $ARB_{Ros}$ (%) | $RMSE_{Ros}$ |
| CP sampling | $\pi_{2k}, \pi_{2k\ell}$ | 0.08 | 20731799 | 9.72 | 19541730 |
| Proposed design | $\pi_{2k}, \pi_{2k\ell}$ | 0.10 | 20624478 | 9.88 | 19483339 |
| Proposed design | $\hat{\tilde{\pi}}_{2k}, \hat{\tilde{\pi}}_{2k\ell}$ | 0.12 | 20701686 | 10.03 | 19554825 |
| Setting | $n_1 = 50$, | $n_2 = 60$ | | | |
| Design | Prob. | $ARB_{SYG}$ (%) | $RMSE_{SYG}$ | $ARB_{Ros}$ (%) | $RMSE_{Ros}$ |
| CP sampling | $\pi_{2k}, \pi_{2k\ell}$ | $<0.01$ | 1438744 | 2.28 | 1353542 |
| Proposed design | $\pi_{2k}, \pi_{2k\ell}$ | 0.07 | 1433427 | 3.30 | 1364083 |
| Proposed design | $\hat{\tilde{\pi}}_{2k}, \hat{\tilde{\pi}}_{2k\ell}$ | 0.06 | 1448275 | 3.33 | 1366047 |

are also similar to those reported by Matei and Tillé (2005a) and Haziza et al. (2008). The values of the RMSE for SYG estimators are comparable for the three sampling designs in both settings. As noted by Haziza et al. (2008), the Rosén estimators tend to underestimate the true variance in most of the scenarios, but their RMSE values are comparable for the three sampling designs.

The ARB and RMSE values of the SYG and Rosén variance estimators for the 'MU284 dynamic population' are given in Table 7 for both settings. The ARB values of the SYG estimators are larger than in the case of the artificial population. These results can be explained by the underlying model linking the variable of interest $y$ and the size variable $x_2$ in the real population. This model is different from a ratio model used to generate $y$ in the artificial population. The ARB values for the three Rosén estimators are similar in the second setting, but show slight differences in the first setting. In both settings, the RMSE values of the SYG and Rosén estimators agree closely for the CP sampling and the proposed design using the prescribed inclusion probabilities.

As shown by our simulation studies, the estimator based on the proposed design and computed using $\pi_{2k}$ and $\pi_{2k\ell}$ does not suffer from much larger variances than the Horvitz-Thompson estimator under CP sampling. A slight bias is shown for small sample sizes. With respect to the ARB and RMSE measures, the proposed design agrees approximately with the CP sampling design when $\pi_{2k}$ and $\pi_{2k\ell}$ are used in estimations. In our simulation studies, the proposed design using $\pi_{2k}$ and $\pi_{2k\ell}$ performs relatively well in estimations for large population and sample sizes.

## 6. Numerical Comparisons

To check the coordination performance of the two proposed methods, we also consider Poisson sampling with PRN (Brewer et al. 1972) and Pareto sampling with PRN (Rosén 1997a,b). Some details about their implementation are given below.

*Table 7. Percent absolute relative bias and root mean square error of the variance estimators using CP sampling and the proposed design, $m = 10^5$ simulation runs; MU284 dynamic population, $N = 72$*

| Setting | $n_1 = 10$, | $n_2 = 6$ | | | |
|---|---|---|---|---|---|
| Design | Prob. | $ARB_{SYG}$ (%) | $RMSE_{SYG}$ | $ARB_{Ros}$ (%) | $RMSE_{Ros}$ |
| CP sampling | $\pi_{2k}, \pi_{2k\ell}$ | 0.57 | 77902.13 | 32.58 | 61920.10 |
| Proposed design | $\pi_{2k}, \pi_{2k\ell}$ | 0.33 | 76168.41 | 29.90 | 59398.93 |
| Proposed design | $\hat{\tilde{\pi}}_{2k}, \hat{\tilde{\pi}}_{2k\ell}$ | 0.23 | 84302.57 | 23.06 | 58673.89 |
| Setting | $n_1 = 30$, | $n_2 = 25$ | | | |
| Design | Prob. | $ARB_{SYG}$ (%) | $RMSE_{SYG}$ | $ARB_{Ros}$ (%) | $RMSE_{Ros}$ |
| CP sampling | $\pi_{2k}, \pi_{2k\ell}$ | 0.38 | 2154.52 | 13.74 | 1872.53 |
| Proposed design | $\pi_{2k}, \pi_{2k\ell}$ | 0.53 | 2146.17 | 12.38 | 1849.40 |
| Proposed design | $\hat{\tilde{\pi}}_{2k}, \hat{\tilde{\pi}}_{2k\ell}$ | 0.29 | 2087.59 | 12.03 | 1849.61 |

Poisson sampling with PRN is very easy to implement: independent random numbers $(r_k)$ from the $U(0,1)$ distribution are permanently assigned to all units in the population. On the first occasion, one selects a unit $k$ in $s_1$ if $r_k < \pi_{1k}$; on the second occasion, $k$ is selected in $s_2$ if $r_k < \pi_{2k}$. It follows that $\pi_k^{1,2} = \min\left(\pi_k^1, \pi_k^2\right)$, and the AUB is reached.

As already pointed out in Subsection 3.1, Poisson sampling, whether employing PRN or not, provides random sample size. To overcome this problem, in the class of unequal sampling designs, several PRN schemes have been proposed in the literature. One of them is order $\pi ps$ sampling (Rosén 1997a,b). Order $\pi ps$ sampling with PRN and a fixed distribution shape $H$ is based on the following idea: independent random numbers $(r_k)$ from the $U(0,1)$ distribution are permanently assigned to all units in the population. On the first occasion, the $n_1$ units having the smallest values of $H(r_k)/H(\pi_{1k})$ are selected as a sample of size $n_1$. Similarly, on the second occasion, the $n_2$ units having the smallest values of $H(r_k)/H(\pi_{2k})$ are selected as a sample of size $n_2$. The sample coordination is assured by the use of the same $r_k$ in both occasions. Different distribution shapes result in various types of order $\pi ps$ sampling. In particular we have: uniform-order sampling or sequential Poisson sampling (Ohlsson 1995, 1998), exponential-order sampling or successive sampling (Hájek 1964), and Pareto-order sampling (Rosén 1997a,b; Saavedra 1995).

Simulations showed that the three order $\pi ps$ sampling designs perform equally in sample coordination. However, Pareto sampling is generally preferred because it is the most efficient design in the class of order $\pi ps$ sampling designs with the same prescribed inclusion probabilities (see Rosén 1997a, b, 2000). Pareto sampling uses the shape $H(x) = x/(1-x)$. While Pareto sampling with PRN performs well in sample coordination, it is, however, an approximate $\pi ps$ sampling design (the inclusion probabilities only agree with the prescribed inclusion probabilities approximately), and does not possess the maximum-entropy property.

All coordination methods are compared using the expected sample overlap computed using Monte Carlo simulation. The expected sample overlap given by the independent selection of samples is also reported in simulations. As a benchmark, we use the AUB given in (1).

Three simulation studies are shown below using the following five different sampling schemes:

a) two CP samples are drawn independently (IND) (using the rejective method for both),
b) two Poisson samples are drawn using Poisson sampling (POI) with PRN,
c) two Pareto samples are drawn using Pareto sampling (PAR) with PRN,
d) two CP samples are drawn using the list-sequential method (SEQ) with PRN,
e) the first sample is a CP one drawn using the rejective method; the second one is selected using the rejective method with updated parameters as described in Section 5.1. We call this method the mixed one (MIX).

For the methods a), d), and e) (only for the first sample) the parameters $p_1$ and $p_2$ were computed from $\pi_{1k}$ and $\pi_{2k}$ respectively (see Equation 4) and used in sample selection. A number of $10^5$ simulation runs was used to compute the expected overlap of two samples drawn using the five methods. $N$ random numbers from $U\left(0,1\right)$ distribution were generated in each simulation and used as PRN in each method. The expected overlap for

each method was computed using the formula

$$E_{sim}(c) = \frac{1}{m}\sum_{\tilde{m}=1}^{m} c_{\tilde{m}}^{1,2},$$

where $m = 10^5$ is the number of runs, $c_{\tilde{m}}^{1,2} = |s_{1\tilde{m}} \cap s_{2\tilde{m}}|$, and $s_{1\tilde{m}}$, $s_{2\tilde{m}}$, are the samples drawn in the $\tilde{m}^{th}$ simulation run. The Monte Carlo variance of the overlap between samples was also reported in the tables below

$$V_{sim}(c) = \frac{1}{m-1}\sum_{\tilde{m}=1}^{m}\left(c_{\tilde{m}}^{1,2} - E_{sim}(c)\right)^2.$$

1. First study – MU284 population

The first study is based on the MU284 population used in Example 2. As before, the inclusion probabilities $\pi_{1k}$ are computed using the variable P75 (population in 1975 in thousands), and $\pi_{2k}$ using the variable P85 (population in 1985 in thousands). In each simulation, samples of expected sizes 10 and 6, respectively, are drawn as described in the methods a)-e).

**Case A)**

We use the 'MU284 static population' described in Example 2 ($N = 48$). Table 8 shows the expected overlap $E_{sim}(c)$ and variance $V_{sim}(c)$ for each of the five methods. Like Poisson sampling with PRN, the mixed method shows a very good expected overlap, which equals the theoretical AUB. The sequential method also provides a very good expected overlap, but smaller than the AUB and the Pareto sampling with PRN performance. For this example, the mixed method performs very well, providing in each run a realized overlap equal to the AUB. As expected, Poisson sampling with PRN shows the larger overlap variance, while PAR, SEQ and MIX methods perform equally, giving similar values for the $V_{sim}(c)$.

**Case B)**

We use the 'MU284 dynamic population' described in Example 3. Samples of expected sizes $n_1 = 10, n_2 = 6$ respectively are drawn from this population of size $N = 72$. Table 9 shows the expected overlap and variance for each method. As in Case A, the mixed method shows an expected overlap equal to the theoretical AUB, and provides better performance than Pareto sampling with PRN. The sequential method performs worse than Pareto sampling concerning the expected overlap. The discrepancies for $V_{sim}(c)$ in Table 9 (except IND and POI) are too small to show a real difference between the methods; they may be due to the process of random simulation.

*Table 8. Expected overlap and variance in the first Monte Carlo study based on $10^5$ simulation runs - MU284 static population*

| Method | $E_{sim}(c)$ | $V_{sim}(c)$ |
|---|---|---|
| IND | 3.05 | 1.51 |
| POI | 6.00 | 4.13 |
| PAR | 5.98 | 0.01 |
| SEQ | 5.83 | 0.15 |
| MIX | 6.00 | 0.00 |
| AUB | 6.00 | |

*Table 9. Expected overlap and variance in the first Monte Carlo study based on $10^5$ simulation runs - MU284 dynamic population*

| Method | $E_{sim}(c)$ | $V_{sim}(c)$ |
|---|---|---|
| IND | 1.55 | 0.78 |
| POI | 2.79 | 1.94 |
| PAR | 2.76 | 1.04 |
| SEQ | 2.55 | 1.00 |
| MIX | 2.79 | 0.99 |
| AUB | 2.79 | |

2. Second study – artificial population I

The second study is based on an artificial population of size $N = 1,000$. No births and no deaths are assumed. In each simulation, samples of expected sizes 100 and 250, respectively, are drawn using the same methods as before (a-e). The inclusion probabilities $\pi_{1k}$ and $\pi_{2k}$ are randomly generated using the $U(0,1)$ distribution and are normalized to sum to 100 and 250, respectively. Table 10 shows the expected overlap and the variance for each method.

The results given in Table 10 show very good performances of the proposed methods. The first proposed method shows a good coordination level of samples, but performs worse than the mixed method. The mixed method yields results approximately equal to the AUB. The difference between 87.01 ($E_{sim}(c)$ for MIX) and 86.99 (the AUB) in Table 10 is due to the difference between $\pi_{2k}$ and $\tilde{\pi}_{2k}$. Indeed, $E_{sim}(c)$ for MIX is an unbiased estimator of $\sum_{k\in U} \min(\pi_{1k}, \tilde{\pi}_{2k})$ and not of the $\sum_{k\in U} \min(\pi_{1k}, \pi_{2k})$, the AUB.

3. Third study – artificial population II

Finally, we consider an extreme situation for the mixed method (the same example is given in Section 5.1). It is a case where it is not always possible to directly draw a sample $s_2$ using the parameters $p_{2k|s_1}$. Instead, a sample $s_2$ is selected using the parameters $p^*_{2k|s_1}$. We have $N = 6, n_1 = n_2 = 3$, $\boldsymbol{\pi}_1 = (0.3, 0.3, 0.3, 0.7, 0.7, 0.7)'$, $\boldsymbol{\pi}_2 = (0.4, 0.4, 0.4, 0.4, 0.4, 1)'$. Table 11 gives the expected overlap and variance for each method. The mixed method shows an expected overlap larger than the AUB because the first-order inclusion probabilities are not respected exactly for the second design. The estimated first-order

*Table 10. Expected overlap and variance in the second Monte Carlo study based on $10^5$ simulation runs - artificial population I*

| Method | $E_{sim}(c)$ | $V_{sim}(c)$ |
|---|---|---|
| IND | 24.75 | 16.83 |
| POI | 86.99 | 76.15 |
| PAR | 86.90 | 10.26 |
| SEQ | 86.44 | 10.51 |
| MIX | 87.01 | 9.74 |
| AUB | 86.99 | |

*Table 11. Expected overlap and variance in the third Monte Carlo study based on $10^5$ simulation runs - artificial population II*

| Method | $E_{sim}(c)$ | $V_{sim}(c)$ |
|---|---|---|
| IND | 1.62 | 0.44 |
| POI | 2.40 | 1.33 |
| PAR | 2.33 | 0.32 |
| SEQ | 2.32 | 0.35 |
| MIX | 2.48 | 0.30 |
| AUB | 2.40 | |

inclusion probabilities $\left(\hat{\tilde{\pi}}_{2k}\right)$ for the second design are the following: 0.3722, 0.3729, 0.3727, 0.4417, 0.4405, 1.0000. By computing

$$\sum_{k \in U} \min\left(\pi_{1k}, \hat{\tilde{\pi}}_{2k}\right)$$

we obtain the expected overlap of the mixed method of 2.4822, matching the value of $E_{sim}(c)$ for the mixed method given in Table 11.

## 7. Discussion and Conclusions

Two methods of coordinating CP samples or approximate CP samples have been proposed. The two samples can be drawn sequentially or in different time occasions. We have focused on positive coordination, but the negative case can be also implemented using the two proposed methods.

The first method is based on the list-sequential implementation of CP sampling. It is a PRN method and has the advantage of preserving the second sampling design exactly (both samples are CP). It provides a good level of expected overlap as shown in Section 6, but smaller than the AUB. This is mainly due to the differences between the selection and the inclusion probabilities.

The second method is an approximate one, because the second sampling design is not exactly respected. For small populations and samples, there are differences between the inclusion probabilities provided by the proposed sampling design on the second occasion and those of the corresponding CP sampling. In our examples, these differences seem to vanish as the population and sample size increase.

The parameters $p_{2k|s_1}$ cannot be used as inclusion probabilities for a fixed size design because they do not in general sum up to the desired sample size. By instead using them as parameters, they can be rescaled to sum up to $n_2$. As for Conditional Poisson sampling, the inclusion probabilities will differ from the parameters, but the difference vanishes as the population and sample sizes increase (for CP sampling see Aires 2000a). For very small examples, such as Example 1 in Subsection 5.2, there can be larger differences. This is why this procedure only gives an approximate CP sample. On average, the rescaling cancels out, so we achieve approximately correct inclusion probabilities. However, the mixed method should not be used on overly small populations/samples. For small populations, we can use the exact list-sequential method shown in Section 4.

Since the true inclusion probabilities for the second design cannot be directly computed, but in our examples they are close to $\pi_{2k}$ and $\pi_{2k\ell}$, the latter can be used in estimations. As our simulation studies in Subsection 5.3 show, there are small differences in terms of ARB and RMSE in estimations especially for large populations/samples between the proposed design when $\pi_{2k}$ and $\pi_{2k\ell}$ are used and the CP sampling.

The mixed method provides the optimal degree of coordination since by construction it tries to achieve the AUB. The expected overlap provided by this method can be larger or smaller than the AUB. In order to achieve the AUB, we need to have $\pi_k^{1,2} = \min(\pi_{1k}, \pi_{2k})$ for all $k$. This means, for example, that all units $k \in s_1$ for which $\pi_{1k} \leq \pi_{2k}$ must be included in $s_2$ with conditional probability 1. As the third study in Section 6 shows, this may lead to more than $n_2$ units with conditional inclusion probabilities equal to 1. Such probabilities can never be respected by a design of fixed-size $n_2$. Thus, in general the AUB cannot be achieved by any fixed-size design.

The mixed method shows a high performance comparable to Poisson sampling with PRN. It has the advantage of allowing fixed sample sizes compared to Poisson sampling with PRN. Due to this fact, the mixed method provides a smaller overlap variance than Poisson sampling with PRN, as also shown in our simulations. Compared to Pareto sampling with PRN, the mixed method performs better in simulations, but it has the disadvantage of providing only an approximate CP sample in the second selection. On the other hand, Pareto sampling does not possess the maximum-entropy property for given first-order inclusion probabilities.

Based on the criterion to achieve the AUB, the second sampling in the mixed method is an optimal sampling design for the first one. In our paper, the first sample is a CP sample. It is possible to apply the mixed method for any type of fixed-size sampling design used in the first selection. Hence, the method allows to use of, for example, a balanced sample in the first selection. The second sampling is always optimal for the first one if the conditional inclusion probabilities given in the first part of Remark 4 are respected.

## 8. References

Aires, N. 2000a. "Comparisons Between Conditional Poisson Sampling and Pareto πps sampling designs." *Journal of Statistical Planning and Inference* 82: 115.

Aires, N. 2000b. *Techniques to Calculate Exact Inclusion Probabilities for Conditional Poisson Sampling and Pareto πps Sampling Designs*. Doctoral thesis. Chalmers University of Technology and Göteborg University, Sweden. Available at: http://www.math.chalmers.se/Stat/Research/Preprints/Doctoral/2000/1.pdf (accessed October 2015).

Arthnari, T. and Y. Dodge. 1981. *Mathematical Programming in Statistics*. New York: Wiley.

Berger, Y. 1998. "Rate of Convergence for Asymptotic Variance for the Horvitz-Thompson Estimator." *Journal of Statistical Planning and Inference* 74: 149–168. Doi: http://dx.doi.org/10.1016/S0378-3758(98)00107-4.

Bondesson, L., I. Traat, and A. Lundqvist. 2006. "Pareto Sampling Versus Sampford and Conditional Poisson Sampling." *Scandinavian Journal of Statistics* 33: 699–720.

Brewer, K., L. Early, and S. Joyce. 1972. "Selecting Several Samples From a Single Population." *Australian Journal of Statistics* 3: 231–239. Doi: http://dx.doi.org/10.1111/j.1467-842X.1972.tb00899.x.

Broström, G. and L. Nilsson. 2000. "Acceptance-Rejection Sampling From the Conditional Distribution of Independent Discrete Random Variables, Given Their Sum." *Statistics* 34: 247–257. Doi: http://dx.doi.org/10.1080/02331880008802716.

Causey, B.D., L.H. Cox, and L.R. Ernst. 1985. "Application of Transportation Theory to Statistical Problems." *Journal of the American Statistical Association* 80: 903–909. Doi: http://dx.doi.org/10.1080/01621459.1985.10478201.

Chen, S., A. Dempster, and J. Liu. 1994. "Weighted Finite Population Sampling to Maximize Entropy." *Biometrika* 81: 457–469. Doi: http://dx.doi.org/10.1093/biomet/81.3.457.

Chen, S. and J. Liu. 1997. "Statistical Applications of the Poisson-Binomial and Conditional Bernoulli Distributions." *Statistica Sinica* 7: 875–892. Available at: http://www3.stat.sinica.edu.tw/statistica/oldpdf/A7n44.pdf (accessed October 2015).

Cotton, F. and C. Hesse. 1992. *Tirages Coordonnés d'Échantillons*. Technical Report E9206, Direction des Statistiques Économiques, INSEE, Paris, France.

Deville, J.-C. 2000. *Note Sur l'Algorithme de Chen, Dempster et Liu*. Technical Report, CREST-ENSAI, Rennes, France.

Deville, J.-C. and Y. Tillé. 2004. "Efficient Balanced Sampling: The Cube Method." *Biometrika* 91: 893–912. Doi: http://dx.doi.org/10.1093/biomet/91.4.893.

Dupacová, J. 1979. "A Note on Rejective Sampling." In *Contributions to Statistics: Jaroslav Hájek Memorial Volume*, edited by J. Jurecková, 71–78. Prague: Academia.

Ernst, L.R. 1996. "Maximizing the Overlap of Sample Units For Two Designs With Simultaneous Selection." *Journal of Official Statistics* 12: 33–45.

Ernst, L.R. 1998. "Maximizing and Minimizing Overlap When Selecting a Large Number of Units per Stratum Simultaneously For Two Designs." *Journal of Official Statistics* 14: 297–314.

Ernst, L.R. 1999. "The Maximization and Minimization of Sample Overlap Problems: a Half Century of Results." Bulletin of the International Statistical Institute, Proceedings, Tome LVIII, Book 2, 293–296.

Ernst, L.R. and M.M. Ikeda. 1995. "A Reduced-size Transportation Algorithm For Maximizing the Overlap Between Surveys." *Survey Methodology* 21: 147–157.

Ernst, L.R. and S.P. Paben. 2002. "Maximizing and Minimizing Overlap When Selecting Any Number of Units Per Stratum Simultaneously For Two Designs With Different Stratications." *Journal of Official Statistics* 18: 185–202.

Fattorini, L. 2006. "Applying the Horvitz-Thompson Criterion In Complex Designs: a Computer-intensive Perspective for Estimating Inclusion Probabilities." *Biometrika* 93: 269–278. Doi: http://dx.doi.org/10.1093/biomet/93.2.269.

Fattorini, L. 2009. "An Adaptive Algorithm For Estimating Inclusion Probabilities and Performing the Horvitz-Thompson Criterion In Complex Designs." *Computational Statistics* 24: 623–639. Doi: http://dx.doi.org/10.1007/s00180-009-0149-9.

Grafström, A. 2010. "Entropy of Unequal Probability Sampling Designs." *Statistical Methodology* 7: 84–97. Doi: http://dx.doi.org/10.1016/j.stamet.2009.10.005.

Hájek, J. 1964. "Asymptotic Theory of Rejective Sampling With Varying Probabilities From a Finite Population." *Annals of Mathematical Statistics* 35: 1491–1523.

Hájek, J. 1981. *Sampling from a Finite Population*. New York: Marcel Dekker.

Haziza, D., F. Mecatti, and J.N.K. Rao. 2008. "Evaluation of Some Approximate Variance Estimators Under the Rao-Sampford Unequal Probability Sampling Design." *METRON - International Journal of Statistics* LXVI: 91–108. Available at: http://www.researchgate.net/profile/Fulvia_Mecatti/publication/227458252_Evaluation_of_some_approximate_variance_estimators_under_the_Rao-Sampford_unequal_probability_sampling_design/links/00b7d5268caea36e5d000000.pdf (accessed October 2015).

Keytz, N. 1951. "Sampling With Probabilities Proportional to Size: Adjustment For Changes In the Probabilities." *Journal of American Statistics Association* 46: 105–109. Doi: http://dx.doi.org/10.1080/01621459.1951.10500773.

Kish, L., and A. Scott. 1971. "Retaining Units After Changing Strata and Probabilites." *Journal of the American Statistical Association* 66: 461–470. Doi: http://dx.doi.org/10.1080/01621459.1971.10482286.

Mach, L., P.T. Reiss, and I. Schiopu-Kratina. 2006. "Optimizing the Expected Overlap of Survey Samples Via the Northwest Corner Rule." *Journal of the American Statistical Association* 101: 1671–1679. Doi: http://dx.doi.org/10.1198/016214506000000320.

Matei, A. and C. Skinner. 2009. "Optimal Sample Coordination Using Controlled Selection." *Journal of Statistical Planning and Inference* 139: 3112–3121. Doi: http://dx.doi.org/10.1016/j.jspi.2009.02.012.

Matei, A. and Y. Tillé. 2005a. "Evaluation of Variance Approximations and Estimators In Maximum Entropy Sampling With Unequal Probability and Mixed Sample Size." *Journal of Official Statistics* 21: 543–570.

Matei, A. and Y. Tillé. 2005b. "Maximal and Minimal Sample Co-ordination." *Sankhyā* 67: 590–612.

Nedyalkova, D., J. Pea, and Y. Tillé. 2008. "Sampling Procedures For Coordinating Stratified Samples: Methods Based On Microstrata." *International Statistical Review* 76: 368–386. Doi: http://dx.doi.org/10.1111/j.1751-5823.2008.00057.x.

Nedyalkova, D., L. Qualité, and Y. Tillé. 2009. "General Framework For the Rotation of Units In Repeated Survey Sampling." *Statistica Neerlandica* 63: 269–293. Doi: http://dx.doi.org/10.1111/j.1467-9574.2009.00423.x.

Ohlsson, E. 1995. "Coordination of Samples Using Permanent Random Numbers." In *Business Survey Methods*, edited by B.G. Cox, D.A. Binder, B.N. Chinnapa, A. Christianson, M.J. Colledge, and P.S. Kott. 153–169. New York: Wiley.

Ohlsson, E. 1998. "Sequential Poisson Sampling." *Journal of Official Statistics* 14: 149–162.

Ohlsson, E. 2000. Coordination of PPS Samples Over Time. In Proceedings of the Second International Conference on Establishment Surveys, June 17–21, 2000 Buffalo, New York. 255–264. Alexandria, VA: American Statistical Association.

Raj, D. 1968. *Sampling Theory*. New York: McGraw-Hill.

Rosén, B. 1991. *Variance estimation for systematic pps-sampling*. Technical Report 1991:15, Statistics Sweden.

Rosén, B. 1997a. "Asymptotic Theory For Order Sampling." *Journal of Statistical Planning and Inference* 62: 135–158. Doi: http://dx.doi.org/10.1016/S0378-3758(96)00185-1.

Rosén, B. 1997b. "On Sampling With Probability Proportional To Size." *Journal of Statistical Planning and Inference* 62: 159–191. Doi: http://dx.doi.org/10.1016/S0378-3758(96)00186-3.

Rosén, B. 2000. "On Inclusion Probabilties For Order πps Sampling." *Journal of Statistical Planning and Inference* 90: 117–143. Doi: http://dx.doi.org/10.1016/S0378-3758(00)00104-X.

Saavedra, P. 1995. "Fixed Sample Size PPS Approximations With a Permanent Random Number." In Proceedings of the Section on Survey Research Methods, August 13–17, 1995, Orlando, Florida. 697–700. Alexandria, VA: American Statistical Association.

Schiopu-Kratina, I., J.M. Fillion, L. Mach, and P.T. Reiss. 2014. "Maximizing the Conditional Overlap in Business Surveys." *Journal of Statistical Planning and Inference* 149: 98–115. Doi: http://dx.doi.org/10.1016/j.jspi.2014.02.002.

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

Thompson, M.E. and C. Wu. 2008. "Simulation-based Randomized Systematic PPS Sampling Under Substitution of Units." *Survey Methodology* 34: 3–10. Available at: http://www.researchgate.net/publication/228887421_Simulation-based_randomized_systematic_PPS_sampling_under_substitution_of_units (assessed October 2015).

Tillé, Y. 2006. *Sampling Algorithms*. New York: Springer.

Tillé, Y. and D. Haziza. 2010. "An Interesting Property of the Entropy of Some Sampling Designs." *Survey Methodology* 36: 229–231. Available at: http://www.researchgate.net/publication/228454231_An_Interesting_Property_of_the_Entropy_of_Some_Sampling_Designs (accessed October 2015).

Traat, I., L. Bondesson, and K. Meister. 2004. "Sampling Design and Sample Selection Through Distribution Theory." *Journal of Statistical Planning and Inference* 123: 395–413. Doi: http://dx.doi.org/10.1016/S0378-3758(03)00150-2.

Received June 2013
Revised January 2015
Accepted January 2015