An adaptive algorithm for estimating inclusion probabilities and performing Horvitz-Thompson criterion in complex designs

Lorenzo Fattorini

Dipartimento di Metodi Quantitativi, Università di Siena,

P.za S. Francesco 8, 53100 Siena, Italy

Abstract

An algorithm for estimating inclusion probabilities and applying the Horvitz-Thompson criterion is considered for complex sampling designs, when the computation of the actual inclusion probabilities is prohibitive. The inclusion probabilities are estimated by means of independent replications of the sampling scheme. In turn, the number of replications is determined on the basis of the stability of the resulting estimates as well as on the basis of their precision, checked by means of the Bennet inequality. The number of replications is adaptively increased until a suitable level of precision is reached in a sustainable computational time. Details on the FORTRAN routines adopted for implementing the algorithm are given. The procedure is checked using three artificial examples.

Keywords: Complex sampling schemes; Horvitz-Thomson estimation; Replications; Empirical inclusion probabilities; Bennet inequality.

1. Introduction.

The Horvitz-Thompson criterion provides a very general tool for estimating population totals. For any population and any design the Horvitz-Thompson estimator is unbiased with variance depending on first- and second-order inclusion probabilities. Moreover, for any design ensuring the positivity of second-order inclusion probabilities, the sampling variance is unbiasedly estimated by means of the Horvitz-Thompson variance estimator. In the case of fixed-size designs, the Sen-Yates-Grundy statistic may be alternatively used (Hedayat and Shina 1991, pp. 51-2). When some second-order inclusion probabilities are zero or if their computation is prohibitive, the estimator referred to as the Hansen-Hurvitz variance estimator is most widely adopted. Wolter (1985, pp. 43-5) has proven that the Hansen-Hurvitz estimator is conservative for most fixed-size designs.

In the presence of complex sampling schemes, when the computation of the first-order inclusion probabilities is prohibitive, the Horvitz-Thompson estimator cannot be used. For example, closed expressions for the first-order inclusion probabilities are generally lacking in two-phase sampling schemes, where, as pointed out by Sarndal et al. (1992, p.346), the actual values of the inclusion probabilities of any order arise as a result of the expectations (with respect to all the possible first-phase samples) of the second-phase inclusion probabilities conditioned to the selected first-phase sample. In these cases, the Horvitz-Thompson estimator cannot always be used and the so-called double expansion estimator (or π^*-estimator in the parlance of Sarndal et al. 1992) is alternatively adopted. Closed expressions for the first-order inclusion probabilities are also lacking in most schemes in which units are sequentially selected without replacement (usually referred to as successive sampling or drawn-by-drawn sampling).

In these cases, the practical computation of the inclusion probabilities involves enumerating all the possible samples and all the orderings in which units enter the samples. Obviously, the procedure becomes prohibitive as soon as population and sample sizes become moderately large. On the other hand, successive sampling schemes are of great practical validity when researchers do not have a precise idea of the number of units which they are able to sample in the time available for the survey. Indeed, in these cases, units may be suitably drawn sequentially until the end of the survey period. Moreover, successive sampling schemes may be adopted in spatial sampling in order to avoid the selection of contiguous units.

In the presence of with-frame populations and if the sampling scheme does not depend on unknown characteristics of the population, Horvitz-Thompson estimation may be performed even in the presence of very complex sampling schemes. Indeed, in these cases, the sampling scheme adopted in the field can be independently replicated enough times and the first- and second-order inclusion probabilities can be artificially estimated on the basis of the proportion of times in which the units or the pairs of units enter the selected samples. Thus, as suggested by Fattorini (2006), the resulting *empirical inclusion probabilities* may be adopted in the Horvitz-Thompson statistic as well as in the variance estimators instead of their true counterparts. Fattorini (2006) derives the theoretical properties of the corresponding estimators, which will be referred to as *empirical estimators*, showing that, as the number of replicated samples increases, the empirical estimators are asymptotically equivalent to those adopting true inclusion probabilities.

Now, once a sample has been selected in the field, the very practical problem to be faced in implementing the procedure, is to determine a suitable number of replications

in such a way that the resulting empirical estimate is sufficiently near the (unknown) estimates obtained with that sample by using true inclusion probabilities. In the present paper, a computational algorithm is proposed in which, for any replicable sampling scheme, the number of replications is heuristically determined on the basis of the stability of the empirical estimates as well as on their precision, checked by means of the Bennet (1962) inequality. The number of replications is adaptively increased until a suitable level of precision is reached in a sustainable computational time.

Section 2 contains some notations and preliminary results, while the proposed algorithm is described in section 3. The algorithm is subsequently checked in section 4 with a small artificial population adopted by Skalsky (1994) to illustrate a sequential areal sampling suggested for estimating the size of wildlife animal populations, in section 5, where a two-phase version of Pareto order sampling (Rosén 1997a,b, 2000) is applied to the population of Swedish municipalities (Sarndal et al 1992, Appendix B) for estimating the total revenues from the 1985 municipal taxation and in section 6, where the algorithm is applied to an artificial population of 400 spatial units, in which units are selected with probabilities proportional to their sizes but avoiding the selection of contiguous units.

2. Preliminaries and notations.

Consider a with-frame population of N units and denote by $y_1, y_2, \ldots, y_N$ the values of a nonnegative survey variable associated with each unit. Denote the population total by T and consider a without-replacement sampling scheme inducing first-order inclusion probabilities $\pi_j > 0$ $(j = 1,2,\ldots,N)$ as well as second-order inclusion probabilities

$\pi_{jh}\ (h > j = 1,2,\ldots,N)$. With the units referred to by their indexes, the selected sample S may be viewed as a sub-set of size *n* from the basic set $\{1,2,\ldots,N\}$.

When the computation of the first-order inclusion probabilities is prohibitive, the Horvitz-Thompson estimator

$$\hat{T} = \sum_{j \in \mathsf{S}} \frac{y_j}{\pi_j} \tag{1}$$

cannot be used. However, if the sampling scheme does not depend on unknown characteristics of the population, further M samples can be independently selected from the population frame by repeating the same rules as those used to select S. Subsequently, the survey variable is recorded in the field only for the units in S while the M samples $\mathsf{S}_1, \mathsf{S}_2, \ldots, \mathsf{S}_M$ are used to estimate the inclusion probabilities. An invariably positive estimator of π_j is given by

$$p_j = \frac{X_j + 1}{M + 1},\ j = 1,2,\ldots,N$$

where X_j is the number of times unit j enters the M samples. As $M \to \infty$, p_j constitutes a consistent estimator of π_j. Thus, Fattorini (2006) suggests the use of the empirical Horvitz-Thompson estimator

$$\hat{T}_M = \sum_{j \in \mathsf{S}} \frac{y_j}{p_j} \tag{2}$$

as a very natural modification of (1). The author shows that $\hat{T}_M$ converges almost surely to $\hat{T}$ as M increases. Moreover

$$\left|E(\hat{T}_M) - T\right| \leq \frac{T}{(M+2)\pi_0}$$

and

$$\left|MSE(\hat{T}_M) - \text{var}(\hat{T})\right| \leq \frac{9E(\hat{T}^2)}{(M+2)\pi_0} \tag{3}$$

where *MSE* denotes mean squared error and $\pi_0 = \min_{1 \leq j \leq N} \pi_j$.

As for the estimation of the variance of $\hat{T}_M$, if $\pi_{jh} > 0$ for each $h > j = 1,2,\ldots,N$, an invariably positive estimator of π_{jh} is given by

$$p_{jh} = \frac{X_{jh} + 1}{M + 1}, \quad h > j = 1,2,\ldots,N$$

where X_{jh} denotes the number of times units *j* and *h* enter the *M* samples jointly. In this case the empirical Horvitz-Thompson variance estimator

$$v_M^2 = \sum_{j \in S} y_j^2 \left(\frac{1}{p_j^2} - \frac{1}{p_j} \right) + 2 \sum_{j \in S} \sum_{h > j} y_j y_h \left(\frac{1}{p_j p_h} - \frac{1}{p_{jh}} \right)$$

constitutes an asymptotically unbiased estimator of $\text{var}(\hat{T}_M)$, in the sense that

$$\left|E(v_M^2) - \text{var}(\hat{T}_M)\right| \leq \frac{T^2}{(M+2)\pi_{00}}$$

where $\pi_{00} = \min_{h>j} \pi_{jh}$ (see Fattorini, 2006). Alternatively, if $\pi_{jh} = 0$ for some $h > j = 1,2,\ldots,N$, the empirical Hansen-Hurvitz variance estimator

$$w_M^2 = \frac{1}{n(n-1)} \sum_{j \in S} (t_j - \hat{T}_M)^2$$

where $t_j = ny_j / p_j$, tends to be an asymptotically conservative estimator of $\text{var}(\hat{T}_M)$, in the sense that

$$\left|E(w_M^2) - \text{var}(\hat{T}_M).\right| \leq \frac{9T^2}{(M+2)\pi_0} + \Delta$$

(see Fattorini 2006), where, as argued by Wolter (1985, pp.43-5), the quantity

$$\Delta = T^2 - 2\frac{n}{n-1} \sum_{j=1}^{N} \sum_{h>j} y_j y_h \frac{\pi_{jh}}{\pi_j \pi_h}$$

is positive for most fixed-size designs.

Expression (3) is very general and shows that the computational effort needed for bounding the efficiency loss of $\hat{T}_M$ with respect to $\hat{T}$ increases with the efficiency of $\hat{T}$ as well as in the presence of small inclusion probabilities. Thus, when very precautionary values of the coefficient of variation of $\hat{T}$ and π_0 are supposed (for example of order 10^{-2} and 10^{-4} respectively), small losses may be ensured only at the prohibitive cost of some billions of replications. Moreover, even greater effort may be needed for bounding the bias of v_M^2 when some second-order inclusion probabilities are suspected to be very small (for example of order 10^{-7} or 10^{-8}). On the other hand, some simulations demonstrate that, in practical situations, some millions of replications usually suffice for obtaining $\hat{T}_M$ comparable with $\hat{T}$ (Fattorini 2006). Accordingly, once a sample S has been selected, the problem reduces to determining a suitable value for M bounding to a desired level $\alpha > 0$ the probability that, conditional on S, the relative difference between $\hat{T}_M$ and $\hat{T}$ is greater than a desired precision $\varepsilon > 0$.

3. A Bennet inequality-based algorithm.

On the basis of the results reported in the Appendix and which are mainly derived from the Bennet (1982) inequality, an adaptive algorithm, which I will refer to as the BENNET algorithm is constructed. This algorithm is very general and necessitates only the following initial inputs: *i*) the labels identifying the sampled units $j \in \mathsf{S}$ together with the corresponding values of the survey variable y_j; *ii*) an option for choosing between the two variance estimators v_M^2 and w_M^2; *iii*) an option for choosing between the two precision criteria $P(\varepsilon, M, \mathsf{S})$ and $\Pi(\varepsilon, M, \mathsf{S})$ defined in the Appendix. Moreover, since the BENNET algorithm is constructed to work with any replicable

sampling scheme, it also needs an algorithm supplied by the user with the aim of replicating the sampling scheme originally adopted to select S. Obviously the supplied algorithm must be inclusive of a pseudo-random number generator and all the parameters needed to replicate the sample selection (e.g. the initial seeds for starting the random number generator or the values of the size variable when sampling with probability proportional to size is adopted).

In order to describe the algorithm, suppose at first that the empirical Horvitz-Thompson variance estimator v_M^2 is adopted to estimate the sampling variance of $\hat{T}_M$ and the inequality of type (A1) is adopted to check the precision. Define a *step* as a sequence of L consecutive samples independently selected from the population by means of the scheme which originally generated S, in such a way that, at the end of step i, $\hat{T}_i$ and v_i^2 constitute the couple of empirical estimates performed on the basis of the overall $M = i \times L$ selected samples. Thus, quoting from the Appendix, denote by $\hat{M}_i(\varepsilon)$ the value of M_0 and by $\hat{P}_i(\varepsilon)$ the value of the right-hand side of (A1), both evaluated (for a given ε) at the end of step i by means of the empirical inclusion probabilities estimated via the $L \times i$ replications. Moreover, define $\delta > 0$ as the maximum acceptable value for the relative differences between the empirical estimates calculated in different steps and define $k(i)$ as the number of consecutive steps previous to step i, in which the relative differences turn out to be less than δ, *i.e*

$$k(i) = \max\left\{ j : \bigcap_{l=i-j}^{i} \left(\frac{\left|\hat{T}_l - \hat{T}_{i-j}\right|}{\hat{T}_{i-j}} < \delta, \frac{\left|v_l^2 - v_{i-j}^2\right|}{v_{i-j}^2} < \delta \right) \right\}$$

Finally, define K as the minimum acceptable value for $k(i)$.

In accordance with these definitions, the algorithm proceeds as follows. At the beginning of stage 1, two values ε_1 and α_1 are established for ε and α, respectively. Then, starting with step $i=1$, the algorithm determines the number of replications which ensures stable empirical estimates for both $\hat{T}_M$ and v_M^2 as well as the required level of accuracy for $\hat{T}_M$. Accordingly, stage 1 is concluded at step i_1, where

$$i_1 = \min\left\{ i : k(i) \geq K, L \times i \geq \hat{M}_i, \hat{P}_i(\varepsilon_1, \mathsf{S}) \leq \alpha_1 \right\}$$

If stage 1 ends in a sustainable computational time, better accuracy levels may be attempted by decreasing ε_1 and/or α_1 to ε_2 and/or α_2 and then performing stage 2. Stage 2 starts at step i_1+1, ends at step i_2 and is carried out by performing the same operations in stage 1, with ε_2 and α_2 instead of ε_1 and α_1. The procedure is definitively stopped at a stage such that the level of precision, compared with the computational effort , is judged suitable.

It is worth noting that the procedure remains conceptually the same when v_M^2 is replaced by the empirical Hansen-Hurvitz variance estimator w_M^2 or when the inequality of type (A1) is replaced by (A6), in such a way that $\hat{\Pi}_i(\varepsilon)$ is computed instead of $\hat{P}_i(\varepsilon)$. However, it should be noticed that the use of w_M^2, involving only the empirical first-order inclusion probabilities, may provide some savings in the computational time needed to achieve stable estimates. Moreover, for small ε values, it is at once apparent from relations (A2) and (A7) in the Appendix that, while working

with inequality (A1) gives rise to a number of replications $M = O(\varepsilon^{-2})$, the more severe inequality (A6) necessitates a greater computational effort $M > O(\varepsilon^{-2})$. More precisely, the computational effort needed in this case is shown to be $M = O(\varepsilon^{-2} \ln \varepsilon^{-2})$.

The FORTRAN implementation of the BENNET algorithm is performed by means of the subroutine BENNET.FOR. The BENNET subroutine makes use of the subroutine SAMPLE.FOR (or other names), furnished by the user for replicating the sampling scheme. The SAMPLE subroutine is called by means of the instruction

```
call sample(nbig,nsmall,ind)
```

where `nbig` and `nsmall` denote population and sample sizes while `ind` is the vector containing the labels of the sampled units. Any other argument which may be necessary to the SAMPLE subroutine must be passed through the COMMON instruction. The BENNET subroutine also makes use of the subroutine HTM.FOR to compute the empirical Horvitz-Thompson estimates, as well as of the subroutines BENN1.FOR and BENN2.FOR to work with inequalities (A1) and (A6), respectively. All these subroutines are available by request from the author.

Finally, as to the use of the BENNET subroutine in the forthcoming applications, the function RANDOM.FOR by Whichmann and Hill (1982) was adopted to generate pseudo-random numbers. Three integer seeds were necessary before the first entry of the function, which were passed by means of the COMMON instruction. In all the applications, the initial values of the seeds were invariably set at 13, 17 and 23. The computations were performed by a Pentium IV processor (2.40 Ghz).

4. A simple artificial example.

Quoting from Skalski (1994), consider a study area of size $A = 8$, partitioned into $N = 4$ spatial units of size $A_1 = A_2 = 1$, $A_3 = 2$ and $A_4 = 4$. The values of the survey variable within the units are $y_1 = 4$, $y_2 = 6$, $y_3 = 10$ and $y_4 = 44$, with a resulting population total $T = 64$. Moreover, suppose that a sample of size $n = 2$ is selected from the population by means of a draw-by-draw scheme in which, at the first drawing, the selection is performed with probabilities A_j / A $(j = 1,2,3,4)$, while in the second drawing, the selection is performed with probabilities $A_j /(A - A_{j_1})$ $(j \neq j_1)$, where j_1 denotes the label of the unit selected in the first drawing. Skalski (1994) devotes this example to biologists to illustrate the use of the Horvitz-Thompson estimator in schemes in which, at each drawing, the spatial units are selected with probabilities proportional to their sizes. In general, given a population of *N* plots with sizes $A_1, \ldots A_N$ the probability of selecting plot *j* at first drawing is

$$\tau_1(j) = \frac{A_j}{A} \quad , \quad j = 1, \ldots, N$$

while the probability of selecting plot *j* at drawing *i*, stated that the plots $j_1, \ldots, j_{i-1}$ have been selected at the first $i - 1$ drawing turns out to be

$$\tau_i(j / j_1, \ldots, j_{i-1}) = \frac{A_j}{A - (A_{j_1} + \ldots + A_{j_{i-1}})} \quad , \quad j \neq j_1 \ldots \neq j_{i-1}, \; i = 2, \ldots, n$$

In accordance with Hedayat and Sinha (1991, Ch.V) these schemes will be referred to as the PPS scheme, to distinguish them from the more familiar ΠPS schemes in which the

first-order inclusion probabilities, rather than the selection probabilities, are proportional to a size variable.

Owing to the very small values of *N* and *n*, the example is very unrealistic but very simple, and allows for the exact computations of the first- and second-order inclusion probabilities, which in PPS schemes quickly become prohibitive as soon as *N* and *n* become moderately large, since, as pointed out in the Introduction, it should involve enumerating all the possible samples and all the orderings in which units enter the samples. After few computations, the first-order inclusion probabilities turn out to be $\pi_1 = \pi_2 = 0.30952$, $\pi_3 = 0.57143$ and $\pi_4 = 0.80952$, while those of second order are $\pi_{12} = 0.03571$, $\pi_{13} = \pi_{23} = 0.07738$, $\pi_{14} = \pi_{24} = 0.19643$ and $\pi_{34} = 0.41667$. Accordingly, the Horvitz-Thompson estimator of the population total has a standard deviation $sd(\hat{T}) = 15.04767$ and an efficiency with respect to simple random sampling without replacement of 6.27.

Owing to its simplicity, the example was adopted here as a first check of the BENNET algorithm. Suppose that the sample $\mathsf{S} = \{1,4\}$ was selected from the population. If the first- and second-order inclusion probabilities were used, the Horvitz-Thompson estimate would turn out to be $\hat{T} = 67.27643$ while the Horvitz-Thompson estimate of the standard deviation would be 17.05574. On the other hand, if the inclusion probabilities were unknown (as actually happens as soon as *n* and *N* become moderately large), empirical estimation would be necessary. To this purpose, the BENNET subroutine was used with $L = 1{,}000$, $\delta = 0.001$, $K = 100$ and initial values of ε and α both equal to 0.1. In turn, the BENNET subroutine makes use of the subroutine PPS.FOR, which was supplied in order to effectively perform PPS sampling. The PPS subroutine is available by request from the author.

At first, the inequality of type (A1) was adopted to check the accuracy of the resulting estimates. Stage 1 ended after 2.24 seconds of computational time, at step $i_1 = 1{,}339$. Owing to the quickness of computations, greater precision was sought by using ε_2 and α_2 both equal to 0.01. With these accuracy levels, stage 2 ended at step $i_2 = 1{,}482$ after 0.23 seconds. Once again, greater precision was aspired to by setting $\alpha_3 = 0.001$. Then, stage 3 ended at step $i_3 = 2{,}175$ in 1.16 seconds. Finally, ε_4 was decreased to 0.001, and after 5 minutes and 23.35 seconds of computations, the procedure was definitively ended at step $i_4 = 196{,}863$, where for $k(i_4) = 114{,}367$ consecutive steps, the relative differences between the empirical estimates turned out to be less than $\delta = 0.001$. The final empirical Horvitz-Thompson estimate was 67.2779, while the empirical Horvitz-Thompson estimate of the standard deviation was 17.06143. The basic steps of the estimation are reported in Table 1.

The same search strategy was repeated by using the more severe inequality of type (A6). Obviously, in this case much more computational effort was necessary to satisfy the same accuracy levels. In this case, the procedure was stopped at step $638{,}989$, in a total computational time of 17 minutes and 41.89 seconds, while the final empirical Horvitz-Thompson estimate was 67.2767 and the empirical Horvitz-Thompson estimate of standard deviation was17.0564. The basic steps of the estimation are reported in Table 2.

Table 1. Empirical estimation performed on the artificial population quoted from Skalski (1994), with $L = 1{,}000$, $\delta = 0.001$ and *K*=100. Precision was checked by means of inequality of type (A1). The empirical Horvitz-Thompson variance estimator was adopted.

Stage	steps	ε	α	$k(i)$	$\hat{T}_i$	v_i
1	1,339	0.1	0.1	100	67.2473	16.9669
2	1,482	0.01	0.01	100	67.2512	16.9810
3	2,175	0.01	0.001	100	67.2426	17.0253
4	196,863	0.001	0.001	114,367	67.2779	17.0614
-	∞	-	-	-	67.2764	17.0557

Table 2. Empirical estimation performed on the artificial population quoted from Skalski (1994), with $L = 1{,}000$, $\delta = 0.001$ and *K*=100. Precision was checked by means of inequality of type (A6). The empirical Horvitz-Thompson variance estimator was adopted.

Stage	steps	ε	α	$k(i)$	$\hat{T}_i$	v_i
1	1,339	0.1	0.1	100	67.2473	16.9669
2	4,691	0.01	0.01	585	67.2642	17.0585
3	5,299	0.01	0.001	1,193	67.2684	17.0572
4	638,989	0.001	0.001	5,56493	67.2767	17.0564
-	∞	-	-	-	67.2764	17.0557

5. An application of two-phase Pareto sampling

Order sampling constitutes an effective and straightforward class of sampling schemes proposed by Rosen (1997a) for obtaining fixed-size designs with first-order inclusion probabilities $\pi_1,\ldots,\pi_N$ approximately equal to a set of target probabilities $\lambda_1,\ldots,\lambda_N$. Quoting from Rosen (1997b, 2000), given a distribution *F* with support $(0,a)$ (where *a* may equal ∞), the order sampling is performed by generating *N iid* random numbers on $(0,1)$, say $u_1,\ldots,u_N$, computing the *ranking values* $r_1,\ldots,r_N$, where $r_j = F^{-1}(u_j)/F^{-1}(\lambda_j)$, and then selecting the units corresponding to the *n* smallest ranking values. Order sampling has been theoretically investigated by Rosen (1997a) who shows that under very general conditions the π_j s converge to the corresponding values of the λ_j s as the sample size increases. Moreover, among the possible choices of *F*, Pareto sampling (which arises when *F* is the standard Pareto distribution on R^+ with a unit shape parameter) has been proved to be optimal (see Rosen 1997b). Under Pareto sampling the ranking variables are obviously given by

$$r_j = \frac{u_j}{\lambda_j}\frac{1-\lambda_j}{1-u_j} \quad , \; j = 1,\ldots,N \, ,$$

In the presence of populations partitioned into a large number, say *M*, of small clusters of neighbouring units (in the spatial sense or in a covariate space), a two-phase Pareto sampling scheme may be suitable for obtaining a nearly-ΠPS design in which samples also tend to be representative of the cluster structure. More precisely, in the first phase, a sample of *M* units, say S_0, is obtained by selecting one unit per cluster with a

probability proportional to the size variable. Then, in the second phase, a final sample of $n < M$ units is selected by means of Pareto sampling with target probabilities $\lambda_j(\mathsf{S}_0) = n(a_j/T_0)$ for any $j \in \mathsf{S}_0$, where a_j represents the size of unit j and T_0 denotes the total of sizes for the units of S_0. Moreover, r_j is set to be 0 whenever $\lambda_j(\mathsf{S}_0) \geq 1$ in such a way that in this case unit *j* is selected with a probability of one.

Even if Aires (2004) introduces FORTRAN algorithms to calculate first- and second-order inclusion probabilities for the Pareto scheme, these procedures are of no utility in a two-phase framework. Indeed, the inclusion probabilities arising from these routines should be computed for all the possible first-phase samples in order to obtain the actual inclusion probabilities of the two-phase scheme. Thus, the use of empirical inclusion probabilities seems to be the only way to perform Horvitz-Thompson estimation.

Now, consider the well-known MU284 population consisting of the 284 Swedish municipalities labelled from 1 to 284 and partitioned into $M = 50$ small clusters of neighbouring units by Sarndal et al. (1992, Appendix B,C). Suppose the revenue (in millions of kronor) from the 1985 municipal taxation is the survey variable, from which $T = 69{,}578$, and let the inhabitants in the same year be the size variable. Moreover, suppose that a first-phase sample of 50 units is formed by selecting one municipality per cluster with probability proportional to its inhabitants and a final sample S of size $n = 20$ is selected out of these 50 units by means of Pareto sampling. Finally suppose S consists of the following municipality labels:

5,8,16,23,29,41,47,80,114,117,123,137,152,160,198,199,206,211,240,244,

Then, in order to perform empirical estimation, the BENNET subroutine was used with $L = 1{,}000$, $\delta = 0.001$, $K = 100$ with initial values of ε and α both equal to 0.1. The

FORTRAN implementation of the two-phase Pareto sampling was performed by means of the subroutine PARETO2.FOR which is available by request from the author.

Accuracy was checked by inequality of type (A1), while since the design does not ensure invariably positive second-order inclusion probabilities (which obviously vanished for the units of the same cluster), the conservative variance estimator w_M^2 was adopted.

The computation was stopped at stage 4, with $\varepsilon_4 = 0.01$ and $\alpha_4 = 0.001$, after $M_3 = 71{,}766{,}000$ independent selections and a total computational time of 153 minutes and 54.25 seconds. The final empirical Horvitz-Thompson estimate of total revenues was 71,312.83 millions of kronor with an empirical Hansen-Hurvitz estimate of standard deviation of 11,385.76 and a coefficient of variation of 0.16. The basic steps of the estimation are reported in Table 3.

Table 3. Empirical estimation of total revenue (millions of kronor) from the 1985 taxation in Swedish municipalities, with $L = 1{,}000$, $\delta = 0.001$ and *K*=100. Precision was checked by means of inequality of type (A1). The empirical Hansen-Hurvitz variance estimator was adopted.

Stage	Steps	ε	α	$k(i)$	$\hat{T}_i$	w_i
1	2,646	0.1	0.1	100	71,313.65	11,346.96
2	2,648	0.05	0.01	102	71,310.01	11,344.57
3	3,117	0.05	0.001	100	71,284,96	11,329.10
4	71,766	0.01	0.001	522	71,312.83	11,385.76

6. An application to spatial sampling.

When population units are plots partitioning a study region, simple random sampling may prove to be an inefficient design. Indeed, as Cochran (1977) points out, adjacent units are often more alike than units that are far apart, thus giving a poor contribution to the sample information. Moreover, in other situations, an opposite trend may be present in which adjacent units greatly differ from each other. These problems are well recognized in a model-based setting as positive or negative spatial autocorrelation problems.

In order to handle the presence of spatial autocorrelation in a design-based approach, Wywial (1996) proposes a design in which the probability of selecting a sample may increase or decrease linearly with the number of contiguous units in the sample. Obviously, when the probabilities decrease (increase), the design tends to avoid (prefer) the selection of contiguous units. More recently, for handling the presence of a positive spatial autocorrelation, Bryant et al. (2002) prove the existence of a design with equal first-order inclusion probabilities, in which the second-order inclusion probabilities vanish for contiguous units and are otherwise constant. The design constitutes a two-dimensional generalization of the balanced sampling plan excluding contiguous units, previously proposed by Hedayat et al. (1988) for populations of units arranged in a one-dimensional array. Unfortunately, both the proposals give no insight about the unit drawing change mechanism corresponding to these designs. Thus the design implementation becomes prohibitive even for moderate values of N and n, since it involves listing all the possible samples and their corresponding probabilities and then selecting one of these samples.

An alternative procedure which seems to be more attractive from a practical point of view is proposed by Fattorini and Ridolfi (1997). The authors suggest modifying the simple random sampling without replacement in such a way that, at each drawing, the probabilities of selecting those units that are adjacent to the previously selected ones are reduced or increased according to a prefixed factor $\beta \geq 0$. Thus, at the first drawing, each unit has the same probability $\tau_1(j) = 1/N$ $(j = 1,\ldots,N)$ of being selected. Subsequently, conditional on the first $i-1$ selected units $j_1,\ldots,j_{i-1}$, the remaining $N-i+1$ units are selected at drawing i $(i = 2,\ldots,n)$ with probability

$$\tau_i(j \mid j_1,\ldots,j_{i-1}) = \begin{cases} \dfrac{\beta}{N-i+1+(\beta-1)N(j_1,\ldots,j_{i-1})} & \text{if} \quad j \in C(j_1,\ldots,j_{i-1}) \\ \dfrac{1}{N-i+1+(\beta-1)N(j_1,\ldots,j_{i-1})} & \text{otherwise} \end{cases} \tag{4}$$

where $C(j_1,\ldots,j_{i-1})$ denotes the set of units which are contiguous to at least one of the $i-1$ units $j_1,\ldots,j_{i-1}$ and $N(j_1,\ldots,j_{i-1})$ denotes the cardinality of the set. It is at once apparent from (4) that the design exists whenever $\beta > 0$ and it reduces to simple random sampling when $\beta = 1$. Moreover, when $\beta > 0$ each sample of size *n* has a positive probability of being selected, in such a way that $\pi_{jh} > 0$ for all $h > j = 1,\ldots,N$. When $\beta = 0$, the second-order inclusion probabilities vanish for contiguous units and the design exists providing that $N-i+1 > N(j_1,\ldots,j_{i-1})$ for each $i = 2,\ldots,N$, a condition ensuring that, at each drawing, the remaining units are more numerous than those that are contiguous to at least one of the previously selected units. Finally, as $\beta \to \infty$, (4) reduces to

$$\tau_i(j \mid j_1,\ldots,j_{i-1}) = \begin{cases} \dfrac{1}{N(j_1,\ldots,j_{i-1})} & \text{if} \quad j \in C(j_1,\ldots,j_{i-1}) \\ 0 & \text{otherwise} \end{cases}$$

in such a way that some second-order inclusion probabilities may vanish for very distant units.

Obviously, when units are of equal size, the variance of the population values simply measures the spatial variation of the interest variable across the study area. On the other hand, when the units are of different sizes $A_1,\ldots,A_N$, the population variance is inflated by the variation of plot size, thus rendering inadvisable the use of simple random sampling. In these cases, if the interest variable is likely to increase with size, a suitable solution may be the use of a ΠPS design in which the first-order inclusion probabilities are proportional to the unit sizes. Obviously, this may be straightforwardly performed by adopting one of the 43 sampling schemes listed in Brewer and Hanif (1983). However, in many situations, the presence of unequal-sized units may concur with the presence of a spatial autocorrelation. Thus, in order to take into account both these problems, Barabesi et al. (1997) suggest modifying the sampling scheme proposed by Skalski (1994) in which, at each drawing the probability of selecting a unit is proportional to its size (see Section 4). Once again the modification is performed by reducing or increasing the selection probabilities of those units that are adjacent to the previously-selected ones according to a factor $\beta \geq 0$. Thus, at the first drawing, $\tau_1(j) = A_j / A$ $(j = 1,\ldots,N)$. Subsequently, conditional on the first $i-1$ selected units $j_1,\ldots,j_{i-1}$, the remaining $N-i+1$ units are selected with probability

$$\tau_i(j \mid j_1,\ldots,j_{i-1}) = \begin{cases} \dfrac{\beta A_j}{A - A_{j_1} - \ldots - A_{j_{i-1}} + (\beta-1)H(j_1,\ldots,j_{i-1})} & \text{if} \quad j \in C(j_1,\ldots,j_{i-1}) \\ \dfrac{A_j}{A - A_{j_1} - \ldots - A_{j_{i-1}} + (\beta-1)H(j_1,\ldots,j_{i-1})} & \text{otherwise} \end{cases} \tag{5}$$

where $H(j_1,\ldots,j_{i-1})$ now denotes the total size of the units which are contiguous to at least one of the $i-1$ units $j_1,\ldots,j_{i-1}$. Also in this case the design exists for any $\beta > 0$ and $\pi_{jh} > 0$ for all $h > j = 1,\ldots,N$, while for $\beta = 0$, the second-order inclusion probabilities vanish for contiguous units and the design exists providing that $N - i + 1 > N(j_1,\ldots,j_{i-1})$ for each $i = 2,\ldots,N$. Moreover, as $\beta \to \infty$, (5) reduces to

$$\tau_i(j \mid j_1,\ldots,j_{i-1}) = \begin{cases} \dfrac{1}{H(j_1,\ldots,j_{i-1})} & \text{if} \quad j \in C(j_1,\ldots,j_{i-1}) \\ 0 & \text{otherwise} \end{cases}$$

so that some second-order inclusion probabilities may vanish for very distant units. Fattorini and Ridolfi (1997) give some guidelines for the β choice, showing that, in presence of a marked positive or negative autocorrelation, the efficient choices are $\beta = 0$ or $\beta = \infty$, respectively.

Both these sampling schemes are easy to implement in practice. However the Horvitz-Thompson estimator is inapplicable even for moderate values of *N* and *n*, since, once again, the computation of first- and second-order inclusion probabilities should involve enumerating all the possible samples and all the orderings in which the units enter the sample. Accordingly, the inclusion probabilities may be estimated by means of an appropriate number of replications.

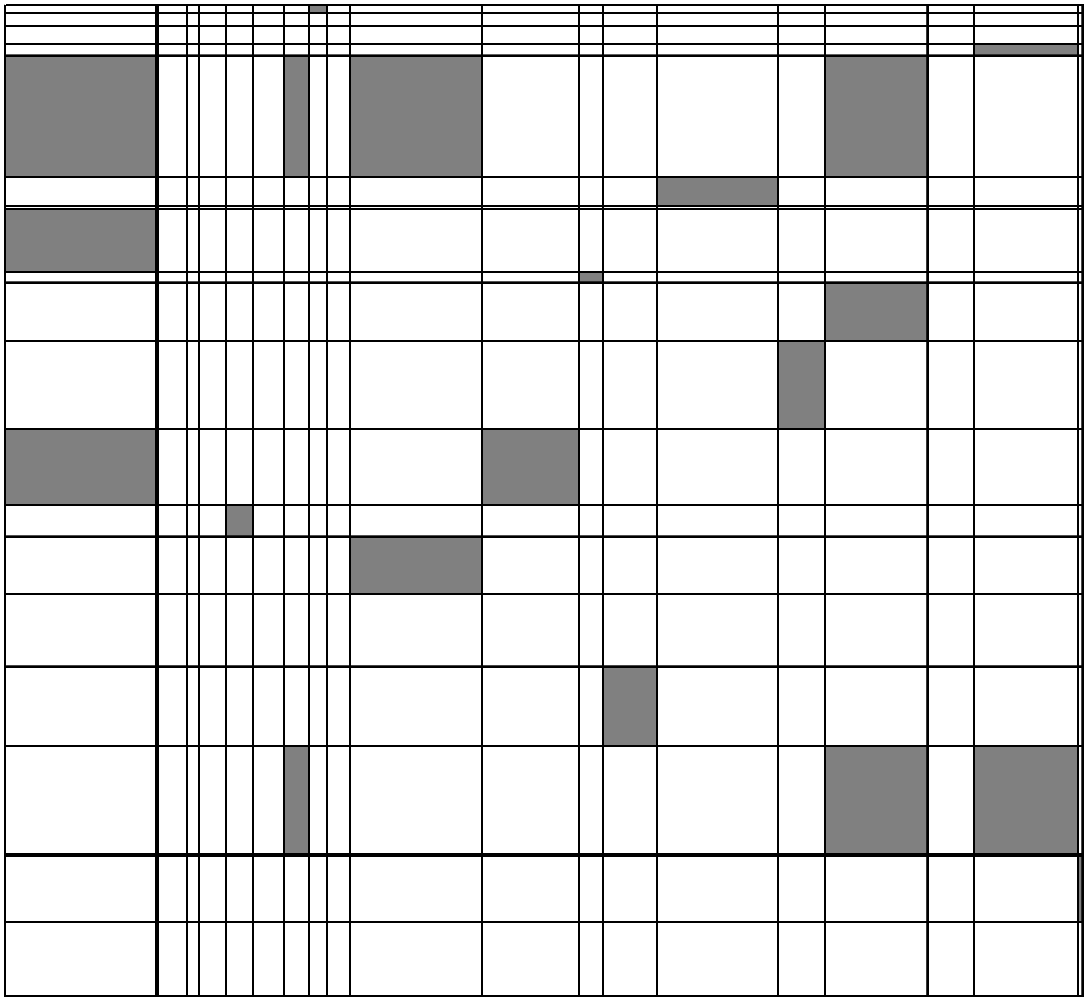
In order to check the validity of this strategy, a simulation study was performed by Fattorini (2006) on two artificial population of $N = 400$ spatial units. Population 1 was made up of 400 quadrats of equal size corresponding to the square $(0,1) \times (0,1)$, while Population 2 was made up of 400 rectangles of different sizes, generated within the square $(0,1) \times (0,1)$ by means of 19 horizontal and vertical random lines (see Figure 1). A survey variable was spread over the square with intensity $4uv$ $(0 < u, v, < 1)$, in such a way that its cumulative value within each unit was given by integrating the intensity over the unit. In both the populations $T = 1$. The variance in Population 1 was 0.000005 while the variance in Population 2 increased 4.2 times to 0.000021 and the correlation coefficient between the size and the survey variable in Population 2 was 0.633129. Finally, the Moran spatial autocorrelation index in Population 1 was 0.9113, which reduced to 0.285409 in Population 2, since the actual level of autocorrelation was masked by the variability in unit sizes. From both the populations, 1000 samples of size $n = 20$ and $n = 40$ were selected by means of the Fattorini and Ridolfi (1997) and Barabesi et al. (1997) sampling schemes. In order to handle the positive spatial autocorrelation, the schemes were implemented with $\beta = 0$ and $\beta = 0.01$. Moreover, in order to achieve a sustainable simulation time, the inclusion probabilities of each sample were estimated by means of just $M = 1000$ replications. The simulation results showed that, even with an inadequate value of M, negligible levels of bias are always incurred by $\hat{T}_M$ along with remarkable gains in efficiency relative to simple random sampling without replacement.

Now, suppose that a sample S of size $n = 20$ consisting of the following spatial units:

5,8,16,23,29,41,47,80,114,117,123,137,152,160,198,199,206,211,240,244,

has been selected from Population 2 by means of the Barabesi et al. (1997) scheme with $\beta = 0$, i.e with selection probabilities proportional to the sizes of the units but avoiding the selection of those units contiguous to the units previously selected. As highlighted in Figure 1, in which the selected units are grey, the scheme provides an even distribution of these units throughout the study area.

Figure 1. Graphical representation of Population 2 by Fattorini (2006). The selected units are shown in grey.



Then, in order to perform empirical estimation, the BENNET subroutine was used with $L = 1{,}000$, $\delta = 0.001$, $K = 100$ with initial values of ε and α both equal to 0.1. The FORTRAN implementation of the Barabesi et al (1997) scheme was performed by means of the subroutine BFRSAMP.FOR which is available by request from the author. Accuracy was checked by inequality of type (A1), while since the second-order inclusion probabilities obviously vanished for contiguous units, the conservative variance estimator w_M^2 was adopted.

The computation was stopped at stage 4, with $\varepsilon_4 = 0.05$ and $\alpha_4 = 0.001$, after $M_4 = 6{,}818{,}000$ independent selections and a total computational time of 21 hours, 17 minutes and 32.23 seconds. The final empirical Horvitz-Thompson estimate of the total was 1.1785 with an empirical Hansen-Hurvitz estimate of standard deviation of 0.2194 and an estimated coefficient of variation of 0.19. The basic steps of the estimation are reported in Table 4.

Table 4. Empirical estimation of total for the Population 2 by Fattorini (2006), with $\mathbf{L = 1{,}000}$, $\boldsymbol{\delta = 0.001}$ and *K*=100. Precision was checked by means of inequality of type (A1). The empirical Hansen-Hurvitz variance estimator was adopted.

Stage	Steps	ε	α	$k(i)$	$\hat{T}_i$	w_i
1	1,393	0.1	0.1	100	1.1783	0.2191
2	3,585	0.05	0.05	162	1.1794	0.2194
3	4,989	0.05	0.01	100	1.1786	0.2193
4	6,818	0.05	0.001	1428	1.1785	0.2194

Acknowledgements

The author is grateful to Professor Luca Pratelli for his helpful suggestions in the theoretical aspects of the work.

Appendix. Two results from the Bennett inequality

Given a selected sample S and an arbitrary $\varepsilon > 0$, denote by $\pi_0(\mathsf{S}) = \min_{j \in \mathsf{S}} \pi_j$, $M_0 = (2 + 2\varepsilon)/\{\varepsilon \pi_0(\mathsf{S})\}$ and

$$c(\varepsilon) = \left(\frac{\varepsilon}{2 + 2\varepsilon} + 1 \right) \ln\left(\frac{\varepsilon}{2 + 2\varepsilon} + 1 \right) - \frac{\varepsilon}{2 + 2\varepsilon}$$

The following results hold

Proposition 1

For any $M > M_0$

$$P(\varepsilon, M, \mathsf{S}) = \Pr\left\{ \frac{\left| \hat{T}_M - \hat{T} \right|}{\hat{T}} > \varepsilon \right\} \le 2 \sum_{j \in \mathsf{S}} e^{-c(\varepsilon) M \pi_j} \tag{A1}$$

while, for an adequately small ε, a target probability level $P(\varepsilon, M, \mathsf{S}) = \alpha$ is ensured for any

$$\frac{16}{3\pi_0(\mathsf{S})} \ln\left(\frac{2}{\alpha} \right) \varepsilon^{-2} \le M \le \frac{16}{\pi_0(\mathsf{S})} \ln\left(\frac{2n}{\alpha} \right) \varepsilon^{-2} \tag{A2}$$

Proof

From (1) and (2), $P(\varepsilon, M, \mathsf{S})$ may be rewritten as

$$P(\varepsilon, M, \mathsf{S}) = \Pr\left\{\left|\sum_{j\in\mathsf{S}} y_j \left(\frac{1}{p_j} - \frac{1}{\pi_j}\right)\right| > \varepsilon \sum_{j\in\mathsf{S}} \frac{y_j}{\pi_j}\right\}$$

in such a way that

$$P(\varepsilon, M, \mathsf{S}) \le \Pr\left\{\sum_{j\in\mathsf{S}} y_j \frac{|p_j - \pi_j|}{p_j \pi_j} > \varepsilon \sum_{j\in\mathsf{S}} \frac{y_j}{\pi_j}\right\} \le \sum_{j\in\mathsf{S}} \Pr\left\{|p_j - \pi_j| > \varepsilon\, p_j\right\} =$$

$$= \sum_{j\in\mathsf{S}} \Pr\left\{|p_j - \pi_j| - \varepsilon(p_j - \pi_j) > \varepsilon\pi_j\right\} \le \sum_{j\in\mathsf{S}} \Pr\left\{(1+\varepsilon)|p_j - \pi_j| > \varepsilon\pi_j\right\} =$$

$$= \sum_{j\in\mathsf{S}} \Pr\left\{\left|\frac{M\pi_j - X_j}{M+1} - \frac{1-\pi_J}{M+1}\right| > \frac{\varepsilon\pi_j}{1+\varepsilon}\right\} \le \sum_{j\in\mathsf{S}} \Pr\left\{\left|\pi_j - \frac{X_j}{M}\right| + \frac{1}{M} > \frac{\varepsilon\pi_j}{1+\varepsilon}\right\} \qquad \text{(A3)}$$

Accordingly, if $M > M_0$, in such a way that $(\varepsilon\pi_j)/(2+2\varepsilon) > 1/M$ for each $j \in \mathsf{S}$, then from (A3) and from the well-known Bennett inequality (see Bennett 1962), it can be concluded that

$$\sum_{j\in\mathsf{S}} \Pr\left\{\left|\pi_j - \frac{X_j}{M}\right| > \frac{\varepsilon\pi_j}{2+2\varepsilon}\right\} \le 2\sum_{j\in\mathsf{S}} e^{-c(\varepsilon)M\pi_j}$$

As to the second part of the proposition, it is worth noting that

$$2e^{-c(\varepsilon)M\pi_0(\mathsf{S})} \leq 2\sum_{j\in\mathsf{S}} e^{-c(\varepsilon)M\pi_j} \leq 2ne^{-c(\varepsilon)M\pi_0(\mathsf{S})} \tag{A4}$$

where

$$\lim_{\varepsilon\to 0}\frac{c(\varepsilon)}{\varepsilon^2/8}=1$$

in such a way that, for an adequately small ε,

$$\frac{\varepsilon^2}{16}\leq c(\varepsilon)\leq\frac{3\varepsilon^2}{16}$$

Thus, from (A4) it also follows that

$$2e^{-\frac{3\varepsilon^2}{16}M\pi_0(\mathsf{S})} \leq 2\sum_{j\in\mathsf{S}} e^{-c(\varepsilon)M\pi_j} \leq 2ne^{-\frac{\varepsilon^2}{16}M\pi_0(\mathsf{S})} \tag{A5}$$

Accordingly, by equating the central term to a probability level α and solving for *M*, inequality (A5) gives rise to (A2)

Proposition 2

For any $M > M_0$

$$\Pi(\varepsilon, M, \mathsf{S}) = \Pr\left\{\bigcup_{H=0}^{\infty}\left(\frac{\left|\hat{T}_{M+H} - \hat{T}\right|}{\hat{T}} > \varepsilon\right)\right\} \le 2\sum_{j\in\mathsf{S}}\frac{e^{-c(\varepsilon)M\pi_j}}{1-e^{-c(\varepsilon)\pi_j}} \tag{A6}$$

while, for an adequately small ε, a target probability level $\Pi(\varepsilon, M, \mathsf{S}) = \alpha$ is ensured for any

$$\frac{16}{3\pi_0(\mathsf{S})}\ln(\varepsilon^{-2})\varepsilon^{-2} \le M \le \frac{16}{\pi_0(\mathsf{S})}\left\{\ln\left(\frac{32n}{\alpha\pi_0(\mathsf{S})}\right)+1\right\}\ln(\varepsilon^{-2})\varepsilon^{-2} \tag{A7}$$

Proof

From result (A.1) it immediately follows that

$$\Pr\left\{\bigcup_{H=0}^{\infty}\left(\frac{\left|\hat{T}_{M+H} - \hat{T}\right|}{\hat{T}} > \varepsilon\right)\right\} \le \sum_{H=0}^{\infty}\Pr\left\{\frac{\left|\hat{T}_{M+H} - \hat{T}\right|}{\hat{T}} > \varepsilon\right\} \le$$

$$\le 2\sum_{H=0}^{\infty}\sum_{j\in\mathsf{S}} e^{-c(\varepsilon)(M+H)\pi_j} = 2\sum_{j\in\mathsf{S}}\frac{e^{-c(\varepsilon)M\pi_j}}{1-e^{-c(\varepsilon)\pi_j}}$$

As to the second part of the proposition, it is worth noting that

$$2\frac{e^{-c(\varepsilon)M\pi_0(\mathsf{S})}}{1-e^{-c(\varepsilon)\pi_0(\mathsf{S})}} \le 2\sum_{j\in\mathsf{S}}\frac{e^{-c(\varepsilon)M\pi_j}}{1-e^{-c(\varepsilon)\pi_j}} \le 2n\frac{e^{-c(\varepsilon)M\pi_0(\mathsf{S})}}{1-e^{-c(\varepsilon)\pi_0(\mathsf{S})}} \tag{A8}$$

where

$$\lim_{\varepsilon \to 0} \frac{1 - e^{-c(\varepsilon)\pi_0(\mathsf{S})}}{(\varepsilon^2/8)\pi_0(\mathsf{S})} = 1$$

in such a way that, for an adequately small ε,

$$\frac{\varepsilon^2}{16}\pi_0(\mathsf{S}) \leq 1 - e^{-c(\varepsilon)\pi_0(\mathsf{S})} \leq \frac{3\varepsilon^2}{16}\pi_0(\mathsf{S})$$

Thus, from (A8) it also follows that

$$32\frac{e^{-\frac{3\varepsilon^2}{16}M\pi_0(\mathsf{S})}}{3\varepsilon^2\pi_0(\mathsf{S})} \leq 2\sum_{j\in\mathsf{S}}\frac{e^{-c(\varepsilon)M\pi_j}}{1-e^{-c(\varepsilon)\pi_j}} \leq 32n\frac{e^{-\frac{\varepsilon^2}{16}M\pi_0(\mathsf{S})}}{\varepsilon^2\pi_0(\mathsf{S})} \qquad \text{(A9)}$$

Accordingly, by equating the central term to a probability level α and solving for *M*, inequality (A9) gives rise to

$$\frac{16}{3\pi_0(\mathsf{S})}\ln\left(\frac{32}{3\alpha\pi_0(\mathsf{S})}\varepsilon^{-2}\right)\varepsilon^{-2} \leq M \leq \frac{16}{\pi_0(\mathsf{S})}\ln\left(\frac{32n}{\alpha\pi_0(\mathsf{S})}\varepsilon^{-2}\right)\varepsilon^{-2}$$

which, in turn, gives rise to (A7) if $\varepsilon \leq 1/\sqrt{e}$.

References

Aires N (2006) A guide to Fortran programs to calculate inclusion probabilities for conditional Poisson sampling and Pareto πps sampling designs. Comput Statist 19:337-345.

Barabesi L, Fattorini L, Ridolfi G (1997) Two-phase surveys of elusive populations. In: Proceedings of the Statistics Canada Symposium 97: New Direction in Surveys and Censuses. Statistics Canada, Ottawa, pp. 285-287.

Bennet G (1962) Probability inequalities for the sum of independent random variables. J Amer Statist Ass 57:33-45

Brewer KRW, Hanif M (1983) Sampling with Unequal Probabilities. Springer-Verlag , New York.

Bryant D, Chang Y, Rodger CA, Wei R (2002) Two-dimensional balanced sampling plans excluding contiguous units. Commun Statist-Theory Meth 31:1441-1455.

Chen XH, Dempster AP, Liu JS (1994) Weighted finite population sampling to maximize entropy. Biometrika 81:457-69.

Cochran WG (1977) Sampling Techniques. Wiley, New York.

Deville JC, Tillé Y (1998) Unequal probability sampling without replacement through a splitting method. Biometrika 85:89-101.

Fattorini L (2006) Applying the Horvitz-Thompson criterion in complex designs: a computer-intensive perspective for estimating inclusion probabilities. Biometrika 93:269-278.

Fattorini L, Ridolfi G (1997) A sampling design for areal units based on spatial variability. Metron 55: 59-72.

Hedayat AS, Sinha BK (1991) Design and Inference in Finite Population Sampling. Wiley, New York.

Hedayat AS, Rao CR, Stufken J (1988) Sampling plans excluding contiguous units. J Statist Plann Infer 19:159-170.

Rosén B (1997a) Asymptotic theory for order sampling. J Statist Plann Infer 62:135-158.

Rosén B (1997b) On sampling with probability proportional to size. J Statist Plann Infer. 62:159-191.

Rosén B (2000) On inclusion probabilities for order πps sampling, J Statist Plann Infer 90:117-143.

Sarndal CE, Swensson B, Wretman J (1992) Model Assisted Survey Sampling. Springer-Verlag, New York.

Skalski JR (1994) Estimating wildlife populations based on incomplete area surveys. Wildl Soc Bull 22:192-203.

Tillé Y (1996) An elimination procedure for unequal probability sampling without replacement. Biometrika 83:238-41.

Wichmann BA, Hill ID (1982) Algorithm AS 183. An efficient and portable pseudo-random number generator. Appl Statist 31:188-90.

Wolter KM (1985) Introduction to Variance Estimation. Springer-Verlag, New York.

Wywial J (1996) On space sampling. Statistics in Transition 2:1185-1191.