Fast Procedures For Selecting Unequal Probability Samples From a Stream

Yves Tillé
University of Neuchatel

2018
SMURF

Table of Contents

1. Introduction
2. Computation of the inclusion probabilities
3. Why is the problem complicated?
4. Method with fixed sampling rate
5. Balanced sampling
6. Methods with fixed sample size
   - Generalization of Chao's method
7. Some applications
   - Block reservoir method
   - Applications 4: Balanced Block Reservoir method
   - Two-pass method
8. Conclusions

Introduction

History

- First method with unequal probabilities without replacement: systematic sampling Madow (1949).
- Books Brewer and Hanif (1983), Tillé (2006).
- New applications in computer sciences (Duffield, 2004).
- Principle of elimination: Chao (1982) and Tillé (1996).
- Interaction with the pivotal method (Deville and Tillé, 1998; Grafström et al., 2012; Chauvet, 2012) or the Fuller method (Fuller, 1970).

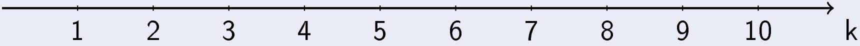
Notation

Notation

- Population U_N of size N.
- Sample s is a subset of U_N.
- Sampling design $p(.)$ is a probability distribution on all the samples such that
$$p(s) \geq 0 \text{ and } \sum_{s \subset U_N} p(s) = 1.$$
- S denotes a random sample $\Pr(S = s) = p(s)$
- Inclusion probability $\pi_k = \Pr(k \in S)$.
- Joint inclusion probability $\pi_{k\ell} = \Pr(\{k, \ell\} \subset S)$.

Notation

Stream



- The stream is so large that it is not possible to store the data.
- The nonselected units are deleted.

Stream end estimation

Stream end estimation

- Stream $U_N, U_{N+1}, U_{N+2}, \dots$
- not possible or difficult to save all the observations.
- Variable of interest y and y_k the value taken by this variable on unit k.
- Total of the values $Y = \sum_{k \in U_N} y_k$.
- Expansion estimator $\widehat{Y} = \sum_{k \in S} \frac{y_k}{\pi_k}$ (Narain, 1951; Horvitz and Thompson, 1952).

Notations

- Sequence of populations $U_1, U_2, \dots, U_n, \dots, U_i, \dots, U_N$.
- Sequence of samples $S_1, S_2, \dots, S_n, \dots, S_i, \dots, S_N$.
- $S_i \subset U_i$.
- Auxiliary variables $x_k > 0, k \in U_N$.

Reservoir method

With equal inclusion probabilities are equal, the reservoir method has been described in Knuth (1981, p. 144), McLeod and Bellhouse (1983) and Vitter (1985).

```
Definition : $k$ integer; $u$ real;
The first $n$ units are selected
Repeat for $k = n + 1, \ldots, N$
    $u$ = uniform random number $[0, 1]$;
    If $u < \frac{n}{k}$
        select unit $k$;
        a unit is removed from the sample ;
        the select unit $k$ takes the place of the removed unit ;
    otherwise unit $k$ is not selected.
```

Bad solutions

- Poisson sampling. Generate independent uniforms u_k in [0,1]. Units such that $u_k < \pi_k$ are selected. Sample size is not fixed. Its distribution is Poisson-Binomial (Hodges Jr. and Le Cam, 1960; Stein, 1990; Chen and Liu, 1997).

Bad solutions 2: Weighted sampling (importance sampling)

- At step 1, a unit is selected from population U_N with unequal drawing probability proportional to $p_k = x_k / \sum_{k \in U_N} x_k$.
- At step 2, if unit j has been selected at step 1, a unit is selected with probability $p_k/(1-p_j)$ for all $k \in U_N \backslash \{j\}$.
- And so on.
- This method is false.
- Proof for $n = 2$

$$\begin{aligned}\pi_k &= P(k \text{ selected at the first step}) \\ &\quad + P(k \text{ selected at the second step}) \\ &= p_k + \sum_{j \in U_N \backslash \{k\}} p_j \frac{p_k}{(1-p_j)}\end{aligned}$$

- π_k is not proportional to x_k and p_k
- The R "sample" function is false.

Bad solutions

Efraimidis (2015)

"In the first case, the relative weight of each item determines the probability that the item is in the final sample. In the second, the weight of each item determines the probability that the item is selected in each of the explicit or implicit item selections of the sampling procedure."

In the area of sampling the word 'weight' should be avoided because it is not clear if it refers to the drawing or to the inclusion probabilities.

Systematic sampling

Systematic sampling

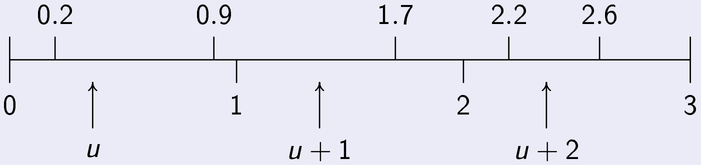
- Madow (1949) also called Dollar Unit Sampling (Leslie et al., 1979) or more generally Monetary Unit Sampling.
- The implementation. Compute the cumulated inclusion probabilities:

$$V_k = \sum_{l=1}^{k} \pi_k \text{ with } V_0 = 0.$$

- A uniform random variable u in $[0, 1]$.
- Next the units k such that the intervals $[V_{k-1} - u, V_k - u]$ contains an integer are selected.

Systematic sampling

$$\pi_1 = 0.2,\ \pi_2 = 0.7,\ \pi_3 = 0.8,\ \pi_4 = 0.5,\ \pi_5 = \pi_6 = 0.4,$$
$$V_0 = 0,\ V_1 = 0.2,\ V_2 = 0.9,\ V_3 = 1.7,\ V_4 = 2.2,\ V_5 = 2.6,\ V_6 = 3,$$
$$\text{and } u = 0.3658$$

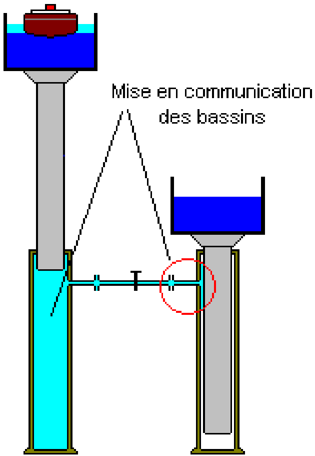


Pivotal method

Pivotal method

Pivotal method

from Michel Maigre©, web site of Région Wallone: Direction des voies hydrauliques, canal du centre.



Pivotal method

- Pivotal method Deville and Tillé (1998).
- Pick two units (denoted by i and j) in the population.

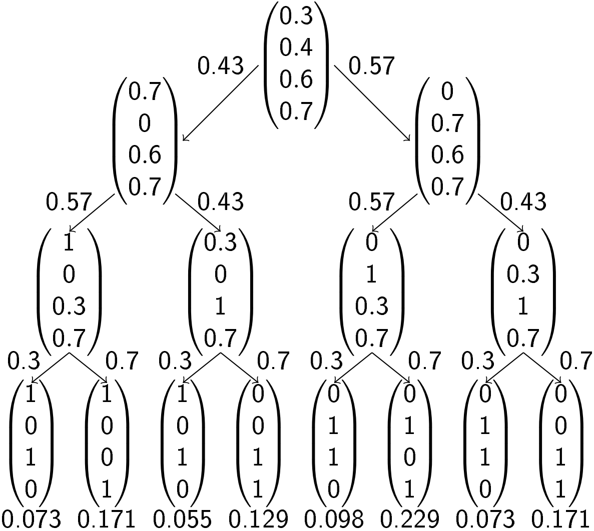
$$(\widetilde{\pi}_i, \widetilde{\pi}_j) = \begin{cases} (\min(1, \pi_i + \pi_j), \max(\pi_i + \pi_j - 1, 0)) & \text{with pr. } q \\ (\max(0, \pi_i + \pi_j - 1), \min(\pi_i + \pi_j, 1)) & \text{with pr. } 1 - q, \end{cases}$$

with

$$q = \frac{\min(1, \pi_i + \pi_j) - \pi_j}{2\min(1, \pi_i + \pi_j) - \pi_i - \pi_j}.$$

- One can check that $(\widetilde{\pi}_i, \widetilde{\pi}_j)$ contains at least one 0 or one 1, that $\mathrm{E}(\widetilde{\pi}_i, \widetilde{\pi}_j) = \mathrm{E}(\pi_i, \pi_j)$ and that $\widetilde{\pi}_i + \widetilde{\pi}_j = \pi_i + \pi_j$.
- The order pivotal method. Take the units in the order.

Pivotal method


$\begin{pmatrix}0.3\\0.4\\0.6\\0.7\end{pmatrix}$
0.43
0.57
$\begin{pmatrix}0.7\\0\\0.6\\0.7\end{pmatrix}$
$\begin{pmatrix}0\\0.7\\0.6\\0.7\end{pmatrix}$
0.57
0.43
0.57
0.43
$\begin{pmatrix}1\\0\\0.3\\0.7\end{pmatrix}$
$\begin{pmatrix}0.3\\0\\1\\0.7\end{pmatrix}$
$\begin{pmatrix}0\\1\\0.3\\0.7\end{pmatrix}$
$\begin{pmatrix}0\\0.3\\1\\0.7\end{pmatrix}$
0.3
0.7
0.3
0.7
0.3
0.7
0.3
0.7
$\begin{pmatrix}1\\0\\1\\0\end{pmatrix}$
$\begin{pmatrix}1\\0\\0\\1\end{pmatrix}$
$\begin{pmatrix}1\\0\\1\\0\end{pmatrix}$
$\begin{pmatrix}0\\0\\1\\1\end{pmatrix}$
$\begin{pmatrix}0\\1\\1\\0\end{pmatrix}$
$\begin{pmatrix}0\\1\\0\\1\end{pmatrix}$
$\begin{pmatrix}0\\1\\1\\0\end{pmatrix}$
$\begin{pmatrix}0\\0\\1\\1\end{pmatrix}$
0.073
0.171
0.055
0.129
0.098
0.229
0.073
0.171

Fuller's method

- Fuller (1971) method is the ordered pivotal method with a random start (see Tillé, 2017).
- Fast implementation of the Fuller method. A phantom unit '0' is added in the beginning with an inclusion probability $\pi_0 \sim U[0,1]$.

Balanced sampling

Bamboleo

Balanced sampling

- 'Rejective procedure': generate samples until obtaining a balanced sample.

  $$p\left(s \,\middle|\, \sum_{j=1}^{p}\left|\frac{\widehat{X}_j - X_j}{X_j}\right| < c\right),$$

  where

  $$X_j = \sum_{k\in U} x_{kj} \text{ and } \widehat{X}_j = \sum_{k\in U}\frac{x_{kj}}{\pi_k},$$

  and c is a positive small value. (Hájek, 1981; Legg and Yu, 2010).
- Problem: the extreme units have a small inclusion probability.

Cube method: Idea of the algorithm

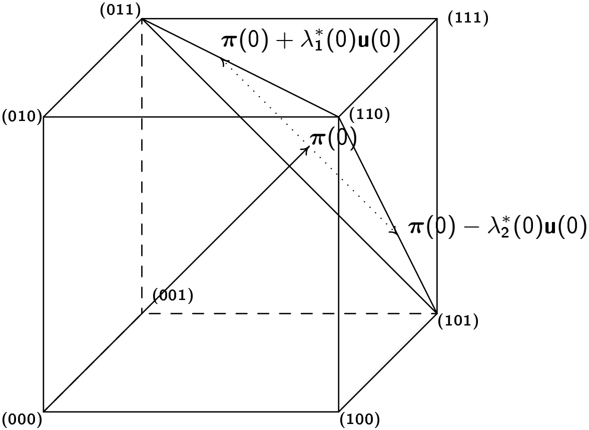

(011)
(111)
$\pi(0) + \lambda_1^*(0)\mathbf{u}(0)$
(010)
(110)
$\pi(0)$
$\pi(0) - \lambda_2^*(0)\mathbf{u}(0)$
(001)
(101)
(000)
(100)

Figure: Flight phase in a population of size $N = 3$ with a sample size constraint $n = 2$

Balanced sampling

- Cube method. Select a balanced sample, i.e.

$$\sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} \approx \sum_{k \in U} \mathbf{x}_k.$$

  (Deville and Tillé, 2004; Tillé and Favre, 2004; Deville and Tillé, 2005; Tillé and Favre, 2005; Chauvet and Tillé, 2006; Chauvet et al., 2011; Tillé, 2011; Breidt and Chauvet, 2011).
- The balancing constraints cannot be exactly satisfied (rounding problem).
- Two phases: Flight phase, landing phase.
- During the flight phase the balancing equation is exactly balanced.
- The landing phase is used to fix the rounding problem.
- Inplementaton in R language (Tillé and Matei, 2015; Grafström and Lisic, 2016) and in SAS (Rousseau and Tardieu, 2004; Chauvet and Tillé, 2005).

Fast Cube method

Fast Cube method

- Cube method select samples that are balanced

$$\sum_{k\in S}\frac{\mathbf{x}_k}{\pi_k}\approx\sum_{k\in U}\mathbf{x}_k.$$

- Fastcube: Chauvet and Tillé (2006) is a generalization of the ordered pivotal method. If consist of running the cube method on only $p+1$ units at each steps.
- The units that are not selected in the stream can be forgotten.
- Allows the selection of a balanced sample in a stream.

Inclusion probabilities

- Auxiliary variable x with x_k be the value taken by x on unit $k > 0$.
  - *Fixed sampling rate*. Coefficient of proportionality $\tau > 0$. Then

    $$\pi_k = \min(1, \tau x_k). \tag{1}$$

    Sample size not controlled $\mathrm{E}(n) = \sum_{k \in U_N} \pi_k$.
  - *Fixed sample size*. Inclusion probabilities updated for $U_n, U_{n+1}, \dots, U_i, \dots, U_N, \dots,$

    $$\pi_k(U_i, n) = \min(1, x_k \, \tau_i), \tag{2}$$

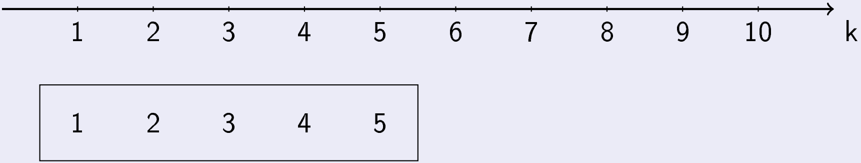
    where τ_i is obtained by solving

    $$\sum_{k \in U_i} \min(1, x_k \, \tau_i) = n.$$

Chao's method

Chao's method

- Chao's method is a resevoir method (Chao, 1982; Sugden et al., 1996).
- The first sample contains the first $n = 5$ units.



1 2 3 4 5 6 7 8 9 10 k

1 2 3 4 5

- Each time a new unit appears, it can enter in the reservoir and a unit of the reservoir is removed.

Chao Method

Chao Method

- Generalization of the Reservoir method to unequal inclusion probabilities.
- The only reservoir method that really satisfies given inclusion probabilities was proposed by Chao (1982) (see also Sugden et al., 1996).
- Sequence $U_n \subset U_{n+1} \subset \cdots \subset U_i \subset \cdots \subset U_N$.
- $\pi_k(U_i, n)$ denotes the inclusion probabilities computed on U_i for a sample of size n.
- Chao's method begins by selecting the first n population units U_n, which is called the reservoir.

Fixed size methods

Example of $\pi_k(U_i, n)$ $N = 12, n = 5$

x	0.70	2.6	0.48	0.40	0.21	0.73	0.15	0.43	0.53	0.05	0.34	3.70
U_5	1	1	1	1	1							
U_6	1	1	0.88	0.73	0.39	1						
U_7	1	1	0.77	0.64	0.34	1	0.24					
U_8	0.91	1	0.62	0.51	0.28	0.94	0.19	0.56				
U_9	0.77	1	0.53	0.44	0.24	0.80	0.16	0.48	0.58			
U_{10}	0.76	1	0.52	0.43	0.23	0.79	0.16	0.47	0.57	0.05		
U_{11}	0.70	1	0.48	0.40	0.21	0.72	0.15	0.43	0.52	0.05	0.34	
U_{12}	0.53	1	0.36	0.30	0.16	0.54	0.11	0.32	0.39	0.04	0.25	1

Chao Method

- At each step the reservoir is updated as follows.
  - At step $i-1$, the reservoir is denoted by S_{i-1}.
  - At step $i = n+1, \dots, N$, unit i is included with probability $\pi_i(U_i, n)$.
  - If unit i is selected, one of the unit of the reservoir is removed with probability

$$a_{ki} = \frac{1}{\pi_i(U_i, n)} \left[1 - \frac{\pi_k(U_i, n)}{\pi_k(U_{i-1}, n)}\right], k = 1, \dots, i-1.$$

  - It is indeed possible to proof that

$$\sum_{k \in S_{i-1}} a_{ki} = 1.$$

This is magic!

Chao Method

- Cohen et al. (2009) have shown that the unselected units can be forgotten.

Generalization of the Chao Method

Flight phase of the cube method

- The flight phase of the cube method is a procedure that provide a quasi-sample.
$$\boldsymbol{\psi} = (\psi_1, \dots \psi_N)^\top,$$
where
- $\displaystyle\sum_{k \in U} \frac{\mathbf{x}_k}{\pi_k} \psi_k = \sum_{k \in U} \frac{\mathbf{x}_k}{\pi_k} \pi_k = \sum_{k \in U} \mathbf{x}_k.$
- $\mathrm{E}(\boldsymbol{\psi}) = \mathrm{E}(\boldsymbol{\pi}).$
- $\#\{0 < \psi_k < 1\} \leq p$, where p is the dimension of $\mathbf{x}_k$.

Generalization of the Chao Method

Generalization of the Chao Method

- Possibility to generalize the Chao method with the notion of a quasi-sample.
- The reservoir is a quasi-sample.
- The number of non integer component of ψ is less that the dimension of $\mathbf{x}_k$.
- It is possible to update the quasi sample when a new units appear in the stream.
- Units that are not selected are forgotten.
- The final sample is balanced.

Applications 1: Preserving two variables for unequal probabilities

Preserving two variables

- Two variables $u_k > 0$ and $v_k > 0$
- Compute $x_k = \max(u_k, v_k)$.
- Quasi-sample ψ_k^1 can be selected with inclusion probabilities $\pi_k^1 = \min(1, \tau_1 x_k)$ balanced on $\mathbf{z}_k = (u_k, v_k, x_k)^\top$.
- Next subsample.
- For instance to be proportional to u_k

$$\pi_k^2 = \frac{\min(1, \tau_1 u_k)}{\pi_k^1}.$$

Applications 2: Block reservoir method

Block reservoir method

- Treatment of the reservoir method by block of H new units.
- The reservoir is updated is updated in function of the block.

Applications 3: Balanced Reservoir method

Balanced Reservoir method

- The Chao method can be generalized to balanced sampling.
- The reservoir is a quasi-samples.
- When a new unit is selected, update the quasi-sample.
- One can forget the non selected values.

Applications 4: Balanced Block Reservoir method

Balanced Reservoir method

- The Chao method can be generalized to balanced sampling.
- The reservoir is a quasi-samples.
- Select a block of new units, update the quasi-sample.
- One can forget the non selected values.
- All these methods can be applied on a stream.
- The information about the units that are not selected can be forgotten.

Two-pass method

Two-pass method

- First pass, the size is considerably reduces.
- Select a quasi-sample with inclusion probabilities $\pi_k^1 = \pi_k(U_k, n), k = n + 1, \dots, N.$
- First sample size $n + n \ln \frac{N}{n}$
- This quasi-sample ψ_k^1 is balanced on two auxiliary variables $\mathbf{z}_k = (\pi_k^1, x_k)^\top.$
- At the first pass the inclusion probabilities can be chosen freely.
- Second pass, fixed sample size or balanced. $\pi_k^2 = \pi_k(U_N, n), k = 1, \dots, N.$
- Drawing probabilities $\frac{\pi_k^2 \psi_k^1}{\pi_k^1}$

Conclusions

Conclusions

- Chao method can be generalized.
- Consistent with balanced sampling.
- Consistent with treatment by block.
- Consistent with several phase sampling.
- Very large set of variants of the original algorithm.

Bibliography I

Breidt, F. J. and Chauvet, G. (2011). Improved variance estimation for balanced samples drawn via the Cube method. *Journal of Statistical Planning and Inference*, 141:479–487.

Brewer, K. R. W. and Hanif, M. (1983). *Sampling with Unequal Probabilities*. Springer, New York.

Chao, M.-T. (1982). A general purpose unequal probability sampling plan. *Biometrika*, 69:653–656.

Chauvet, G. (2012). On a characterization of ordered pivotal sampling. *Bernoulli*, 18(4):1099–1471.

Chauvet, G., Bonnéry, D., and Deville, J.-C. (2011). Optimal inclusion probabilities for balanced sampling. *Journal of Statistical Planning and Inference*, 141(2):984–994.

Chauvet, G. and Tillé, Y. (2005). *Fast SAS macros for balancing samples: user's guide*. Software Manual, University of Neuchâtel.

Chauvet, G. and Tillé, Y. (2006). A fast algorithm of balanced sampling. *Journal of Computational Statistics*, 21:9–31.

Chen, X.-H. and Liu, J. S. (1997). Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Statistica Sinica*, 7:875–892.

Cohen, E., Duffield, N., Kaplan, H., Lund, C., and Thorup, M. (2009). Stream sampling for variance-optimal estimation of subset sums. In *Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1255–1264. Society for Industrial and Applied Mathematics.

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

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

Deville, J.-C. and Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128:569–591.

Duffield, N. (2004). Sampling for passive internet measurement: A review. *Statistical Science*, pages 472–498.

Efraimidis, P. S. (2015). Weighted random sampling over data streams. In Zaroliagis, C., Pantziou, G., and Kontogiannis, S., editors, *Algorithms, Probability, Networks, and Games: Scientific Papers and Essays Dedicated to Paul G. Spirakis on the Occasion of His 60th Birthday*, pages 183–195. Springer International Publishing, Cham.

Bibliography II

Fuller, W. A. (1970). Sampling with random stratum boundaries. *Journal of the Royal Statistical Society*, B32:209–226.

Fuller, W. A. (1971). A procedure for selecting nonreplacement unequal probabilities samples. unpublished manuscript seen by courtesy of the author.

Grafström, A. and Lisic, J. (2016). *BalancedSampling: Balanced and spatially balanced sampling*. R package version 1.5.2.

Grafström, A., Lundström, N. L. P., and Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics*, 68(2):514–520.

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

Hodges Jr., J. L. and Le Cam, L. (1960). The Poisson approximation to the Poisson binomial distribution. *Annals of Mathematical Statistics*, 31:737–740.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685.

Knuth, D. E. (1981). *The Art of Computer Programming (Volume II): Seminumerical Algorithms*. Addison-Wesley, Reading, MA.

Legg, J. C. and Yu, C. L. (2010). Comparison of sample set restriction procedures. *Survey Methodology*, 36:69–79.

Leslie, D. A., Teitlebaum, A. D., and Anderson, R. J. (1979). *Dollar-unit sampling: a practical guide for auditors*. Copp Clark Pitman.

Madow, W. G. (1949). On the theory of systematic sampling, II. *Annals of Mathematical Statistics*, 20:333–354.

McLeod, A. I. and Bellhouse, D. R. (1983). A convenient algorithm for drawing a simple random sampling. *Applied Statistics*, 32:182–184.

Narain, R. D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3:169–174.

Rousseau, S. and Tardieu, F. (2004). La macro SAS CUBE d'échantillonnage équilibré, Documentation de l'utilisateur. Technical report, Insee, Paris.

Bibliography III

Stein, C. (1990). Application of Newton's identities to a generalized birthday problem and to the Poisson-Binomial distribution. Technical Report TC 354, Department of Statistics, Stanford University.

Sugden, R. A., Smith, T. M. F., and Brown, R. P. (1996). Chao's list sequential scheme for unequal probability sampling. *Journal of Applied Statistics*, 23:413–421.

Tillé, Y. (1996). An elimination procedure of unequal probability sampling without replacement. *Biometrika*, 83:238–241.

Tillé, Y. (2006). *Sampling Algorithms*. Springer, New York.

Tillé, Y. (2011). Ten years of balanced sampling with the cube method: an appraisal. *Survey Methodology*, 37:215–226.

Tillé, Y. (2017). Fast implementation and generalization of Fuller's unequal probability sampling method. Institut de Statistique, Université de Neuchâtel.

Tillé, Y. and Favre, A.-C. (2004). Co-ordination, combination and extension of optimal balanced samples. *Biometrika*, 91:913–927.

Tillé, Y. and Favre, A.-C. (2005). Optimal allocation in balanced sampling. *Statistics and Probability Letters*, 74:31–37.

Tillé, Y. and Matei, A. (2015). *sampling: Survey Sampling*. R package version 2.7.

Vitter, J. S. (1985). Random sampling with a reservoir. *ACM Transactions on Mathematical Software*, 11(1):37–57.