Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Using balanced sampling in creel surveys

by Ibrahima Ousmane Ida, Louis-Paul Rivest and Gaétan Daigle

Release date: December 20, 2018

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at STATCAN.infostats-infostats.STATCAN@canada.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard table symbols

The following symbols are used in Statistics Canada publications:

. not available for any reference period
.. not available for a specific reference period
... not applicable
0 true zero or a value rounded to zero
0^s value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p} preliminary
^{r} revised
x suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E} use with caution
F too unreliable to be published
* significantly different from reference category ($p < 0.05$)

Published by authority of the Minister responsible for Statistics Canada

An HTML version **is also available.**

Cette publication est aussi disponible en français.

Survey Methodology, December 2018
Vol. 44, No. 2, pp. 239-252
Statistics Canada, Catalogue No. 12-001-X

Using balanced sampling in creel surveys

Ibrahima Ousmane Ida, Louis-Paul Rivest and Gaétan Daigle[1]

Abstract

These last years, balanced sampling techniques have experienced a recrudescence of interest. They constrain the Horvitz Thompson estimators of the totals of auxiliary variables to be equal, at least approximately, to the corresponding true totals, to avoid the occurrence of bad samples. Several procedures are available to carry out balanced sampling; there is the cube method, see Deville and Tillé (2004), and an alternative, the rejective algorithm introduced by Hájek (1964). After a brief review of these sampling methods, motivated by the planning of an angler survey, we investigate using Monte Carlo simulations, the survey designs produced by these two sampling algorithms.

Key Words: Balanced sampling; Creel surveys; Cube method; Multistage sampling; Rejective algorithm; Monte Carlo simulation.

1 Introduction

Creel surveys provide the foundation for estimating the impact of recreational fishing (Pollock, Jones and Brown, 1994). They are conducted to estimate total catch, fishing effort, and catch rate for various species at several locations (Hoenig, Jones, Pollock, Robson and Wade, 1997). As they focus on fish of interest to recreational anglers, they provide useful information for the management and economic contribution of sport fisheries (Minnesota Department of Natural Resources, 2011).

Two methods are used to contact anglers in creel surveys, either the site access or the roving method. In site access, an agent waits at a location that the anglers must go through when they leave the site and interviews them when they depart (Robson and Jones, 1989). With the roving method the agent moves through the survey area and contacts anglers while they are fishing (United States Environmental Protection Agency, 1998). As the agent cannot be on location for the whole survey, survey sampling is used to select the periods when he will be on site, interviewing fishermen.

In practice creel surveys can face several operational constraints especially when they involve many sites as an agent can only be at one site at a given time. Accommodating all these constraints can be a real challenge when planning a survey. This paper discusses balanced sampling in this context. By framing some operational constraints as balancing equations in a multi-stage sampling design, one should be able to ensure that the sample selected meets the necessary requirements.

Balanced sampling is reviewed in Tillé (2011). A popular method to select a balanced sample is the cube method of Deville and Tillé (2004). An alternative is to select repeatedly several unbalanced samples until, by chance, a sample that approximately meets the balancing equations is drawn. This is the rejective method introduced by Hájek (1964), see also Fuller (2009) and Legg and Yu (2010). In a creel survey, the number of balancing equations is typically large. The implementation of the cube method in this context is discussed

1. Ibrahima Ousmane Ida, Louis-Paul Rivest and Gaétan Daigle, Université Laval. E-mail: louis-paul.rivest@mat.ulaval.ca.

in Chauvet (2009) and Hasler and Tillé (2014). See Vallée, Ferland-Raymond, Rivest and Tillé (2015) for a recent application of these methods in the context of a forest inventory. A recent paper in this area by Chauvet, Haziza and Lesage (2015) investigates the properties of the balanced samples obtained using a rejective method.

The objectives of this paper are twofold. First, the operational constraints for a creel survey of striped bass (*Morone saxatilis*) carried out in the Gaspé Peninsula are presented. Then we will show how balanced sampling, implemented using the cube method, can be used to plan a survey fulfilling most of the constraints. The last section of the paper compares the rejective method to the cube method in the context of creel surveys.

In Section 2, balanced sampling is presented using either the cube method or rejective sampling. Section 3 introduces operational constraints for a creel survey and shows how they can be met using balanced sampling with the cube method. In Section 4, the cube method is compared with the rejective algorithm in the context of a resource inventory where the balancing equations only involve indicator variables. Discussions of the results are presented in the Section 5.

2 Balanced sampling

Suppose that U is a finite population of size N that is sampled with a design having selection probabilities given by $\{\pi_i\colon\ i = 1, \ldots, N\}$. If x is an auxiliary variable known for all population units, then the sample is balanced on x if the Horvitz-Thompson estimator for the total of x is equal to the known total of x. In other words, for any balanced sample s, the following equation has to be satisfied,

$$\sum_{i \in s} \frac{x_i}{\pi_i} = \sum_{i=1}^{N} x_i. \tag{2.1}$$

For the surveys considered here, we balance on indicator variables $I_i\,(\omega)$ equal to 1 if unit i is of type ω and 0 otherwise. If all the units i for which $I_i\,(\omega)$ is equal to 1 have the same selection probability π_ω, then equation (2.1) reduces to $\sum_{i \in s} I_i\,(\omega) \big/ \pi_\omega = \sum_{i=1}^{N} I_i\,(\omega)$. In this context the balancing equation simply requests that the number of sampled units of type ω, $n_\omega = \sum_{i \in s} I_i\,(\omega)$, is equal to its expectation,

$$n_\omega = \sum_{i=1}^{N} I_i\,(\omega)\,\pi_\omega. \tag{2.2}$$

To implement balanced sampling we use the cube method of Deville and Tillé (2004), and the extension of Hasler and Tillé (2014) to cope with highly stratified populations. In Section 4 this method is compared with the implementation of the rejection method proposed by Fuller (2009). In the context of this study, we are balancing on T types of units; we want the sampled numbers of units for the T types, $\tilde{n} = (n_1, \ldots, n_T)^\top$, to be equal to their expectations, $E\,(\tilde{n})$, under the sampling design. Under rejective sampling, the sample is said to be balanced if

$$Q_{T,n} = (\tilde{n} - E\,(\tilde{n}))^\top\ [\mathrm{Var}\,(\tilde{n})]^{-1}\ (\tilde{n} - E\,(\tilde{n})) < \gamma^2 \tag{2.3}$$

where $\text{Var}(\tilde{n})$ represents the design based covariance matrix of $\tilde{n}$ and γ^2 is a tolerance value that determines the balancing condition. Samples that do not meet the balancing equation $Q_{T,n} < \gamma^2$ are simply rejected.

3 A creel survey for striped bass in the Gaspé Peninsula

The Gaspé Peninsula is on the Canadian East Coast in the Province of Québec. In 2015 a creel survey for striped bass was conducted in this peninsula as recreational striped bass fishing had just been reintroduced after a long moratorium.

The study area, presented in Figure 3.1, is scattered over more than 250 kms, on the Gaspé Peninsula coast. The survey is carried out by a single wildlife agent; it is not possible for him to visit two distant sites on the same day. For that reason, neighboring sites are grouped into three sectors as shown in Figure 3.1. We consider the survey for the 33 holidays. The survey variable is the fishing effort, in number of hours of fishing. As some sites attract more fishermen than others, the number of visits to site l of sector i has to be proportional to its importance x_{il} as given in Table 3.1. In addition, for the purpose of the survey, a day is divided into three periods (AM, PM, EV), where EV stands for evening, and six subperiods (AM1, AM2, PM1, PM2, and EV1, EV2). For instance AM1 goes from 8:00 to 10:00 while AM2 is from 10:00 to 12:00. A working day contains two periods and four subperiods. For instance if the agent works AM and PM, then he has a free evening. Thus during a working day he is able to visit four sites, two per working period.

The survey population on a day consists of 54 quadruplets, $(\text{sector} \times \text{period} \times \text{subperiod} \times \text{site})$, 4 of which are sampled. To denote population units the following indices are useful:

i) $h = 1, \ldots, H = 33$ represents the days;

ii) $i = 1, 2, 3$ stands for the sectors in Figure 3.1;

iii) $j = 1, 2, 3$ denotes a period within a day;

iv) $k = 1, 2$ represents the subperiods within a period;

v) $l = 1, 2, 3$ represents the sites, see Figure 3.1, within a sector.

The goal is to estimate the fishing effort for combination of subperiod (6 levels) and site (9 levels). We want to plan a survey with a predetermined sample size for the 54 cells of the cross-classified table. The basic selection probabilities are

$$\pi_{hijkl} = \frac{2x_{il}}{3x_{\bullet\bullet}}, \tag{3.1}$$

where replacing i or l by $\bullet$ means that a summation is taken on the corresponding index. Observe that the sum of π_{hijkl} over the indices (i, j, k, l) is equal to 4, the number of units visited by the wildlife technician on a single day.

At a first glance, the sample could possibly be drawn in a single stage using selection probabilities (3.1) by balancing on the 54 site by subperiod indicator variables. This is not feasible because of operational

constraints. The first one is that on a single day the technician visits sites from the same sector to limit the traveling between sites. The second constraint is that on a working day the technician is off duty for the two subperiods of the same period. In order to meet these operational constraints we propose, in the next section, a design having three levels of sampling where sectors are selected at level 1, periods are selected at level 2 and sites are selected at level 3.

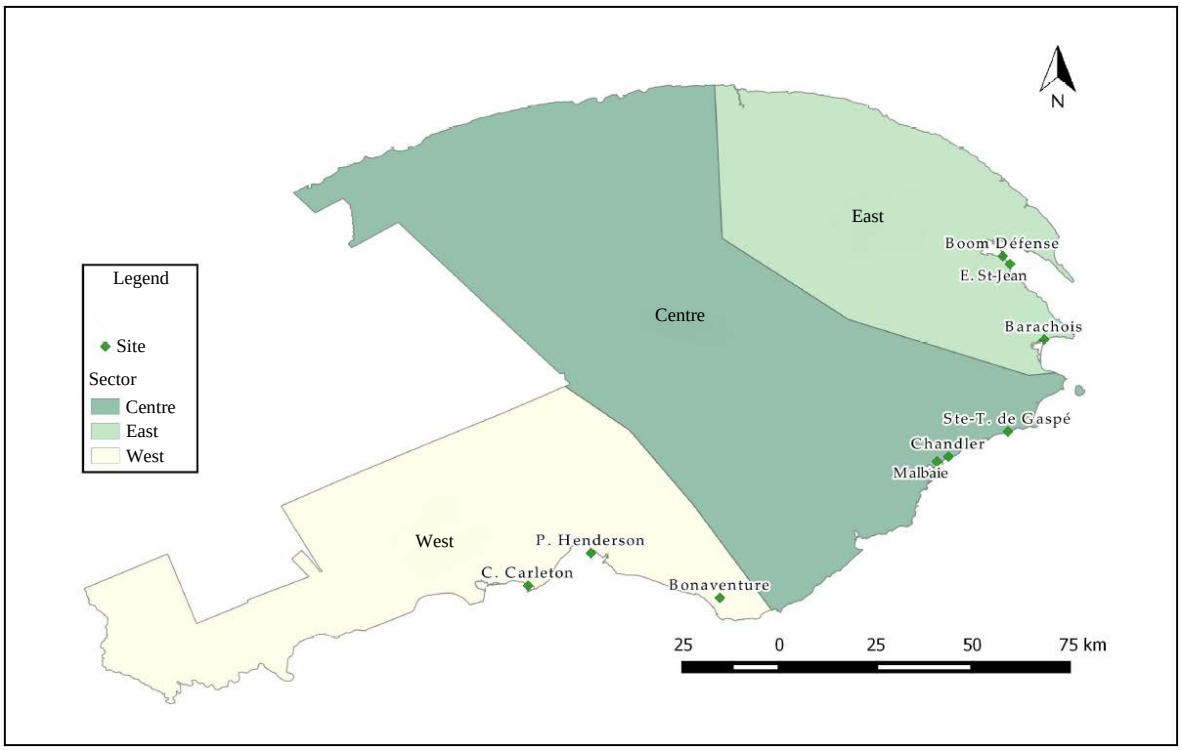

Figure 3.1 The 9 sites to be surveyed for striped bass.

Table 3.1
Average and expected number of visits to each site

Sector	Site	x_{il}	$E(n_{il})$	$\bar{n}_{il}$	$\mathrm{Sd}_{n_{il}}$
East $(i = 1)$	Boom Défense $(l = 1)$	2	20.308	20.286	0.850
	E. St-Jean $(l = 2)$	1	10.154	10.153	0.621
	Barachois $(l = 3)$	2	20.308	20.296	0.881
Centre $(i = 2)$	Ste-T. de Gaspé $(l = 4)$	1	10.154	10.176	0.865
	Malbaie $(l = 5)$	1	10.154	10.155	0.880
	Chandler $(l = 6)$	1	10.154	10.162	0.881
West $(i = 3)$	Bonaventure $(l = 7)$	2	20.308	20.311	1.004
	P. Henderson $(l = 8)$	1	10.154	10.153	0.681
	C. Carleton $(l = 9)$	2	20.308	20.309	1.016

3.1 A balanced multi-stage design for creel survey

This section describes the three stages of the survey that ensures that the operational constraints presented in the previous section are met. It also gives, for each stage, the balancing variables.

The first stage is stratified by day; for each day a single sector is drawn with selection probabilities $x_{i\bullet}/x_{\bullet\bullet}$. At level two, for each sector selected at level 1, two periods are selected out of 3 using simple random sampling (i.e., with selection probabilities 2/3). At level three, a sector*period is stratified by subperiod and one site is selected for each subperiod, the selection probabilities are $x_{il}/x_{i\bullet}$. In summary the selection probabilities at the three levels are

$$\pi_{hi}^{(1)} = \frac{x_{i\bullet}}{x_{\bullet\bullet}}, \qquad \pi_{j|i}^{(2)} = \frac{2}{3}, \qquad \pi_{l|ijk}^{(3)} = \frac{x_{il}}{x_{i\bullet}}.$$

As expected the product $\pi_{hi}^{(1)} \times \pi_{j|i}^{(2)} \times \pi_{l|ijk}^{(3)}$ is equal to (3.1), the target selection probability.

The goal is still to get a sample with predetermined sample sizes for the 54 site by subperiod combinations. Thus balanced sampling needs to be implemented at each stage. At level 1 we need to balance on the indicator variables for the three sectors while at level 2 balancing on the 9 indicator variables for the sector by period combinations is needed. Balancing at level 3 is slightly more complicated as it involves several strata.

At level 2, $33 \times 2 = 66$ sector*periods have been selected. Each one is stratified by subperiod so we are facing 132 strata at level 3 and one site is selected from each one. Balancing is needed with respect to the 54 site by subperiod indicator functions. This is a complex problem and the balancing constraints (2.3) involve the inverse of a large variance covariance matrix. Thus to implement a rejective algorithm in this context one would need an alternative to criterion (2.3) for accepting a sample. For now we discuss the implementation of balanced sampling for this design with the cube method. Comparisons between the cube method and rejective sampling in the context of a simplified creel survey are presented in Section 4.

Among the 132 third stage strata, the number of strata for one subperiod, say AM2, in sector i is an integer close to $22x_{i\bullet}/x_{\bullet\bullet}$ that depends on the stage 2 sample. This integer plays the role of $\sum_{i=1}^{N} I_i(\omega)$ in equation (2.2) for balancing the sites of sector i at stage 3 while, for the l^{th} site, the probability in (2.2) is $\pi_\omega = x_{il}/x_{i\bullet}$. The stage 3 calibration equations for the 54 site by subperiod indicator functions can be described in a similar way. Clearly, it is not possible to meet exactly the 54 balancing equations and the cube method will give a sample that is approximately balanced.

The approximation occurs at the landing phase of the algorithm where balancing constraints are dropped in order to complete the selection of the sample, as introduced in Deville and Tillé (2004). As the stage 3 sample is highly stratified, we use the implementation of the landing phase in the function `balancedstratification2` developed in Hasler and Tillé (2014), with a small correction that prevents it from stopping when the sample is already balanced at the start of the landing phase. In the matrix of balancing constraints, the site constraints were given more importance than those which make visits to

each site equally distributed among subperiods at level 3. They were the last ones to be dropped at the landing phase of the cube method.

To investigate how a failure to meet all balancing equations impacted the sample design, we generated $B = 10{,}000$ random replications of the balanced sample. The number of visits n_{il} to site (i, l) was noted. Table 3.1 compares the average $\bar{n}_{il}$ of n_{il} over the Monte Carlo replications,

$$\bar{n}_{il} = \frac{1}{B}\sum_{b=1}^{B} n_{il}^{(r)},$$

to its expectation, $E(n_{il})$. For all practical purposes, the two are equal and a failure to meet some balancing equations has no impact on the site selection probabilities. Table 3.1 also reports the standard deviations

$$\mathrm{Sd}_{n_{il}} = \left\{\frac{1}{B-1}\sum_{b=1}^{B}\left(n_{il}^{(b)} - \bar{n}_{il}\right)^2\right\}^{1/2}. \tag{3.2}$$

Most of the standard deviations are less than 1 in Table 3.1. Thus the absolute differences between target and realized sample sizes are less than or equal to 2 for most Monte Carlo samples.

Table 3.2 gives the expected number of visits in the 6 subperiods; they are all equal to 22, up to two decimal points, with standard deviations less than 0.2. Thus the period and subperiod constraints are met. Table 3.3 gives a realized sample for the first five days of the creel survey. It shows a harmonious permutation of sectors at level 1, periods at level 2, and sites at level 3 through the days because of the way in which the sample design was constructed. Given a balanced sample produced by the cube algorithm, an arbitrary permutation of the days gives an alternative balanced sample. Indeed the sampling design is invariant to a relabeling of the days. For instance, with the sample of Table 3.3 the technician has to travel from the western to the eastern sector between days 4 and 5. To avoid this long trip one could interchange days 1 and 5: the first two days would then be spent in the eastern sector and between days 4 and 5 the technician would travel from the western to the central sector. The alternative and the original samples have the same estimated totals for the calibration variables.

Table 3.2
Average and expected number of visits at each subperiod

Period	Subperiod	$E(n_{jk})$	$\bar{n}_{jk}$	$\mathrm{Sd}_{n_{jk}}$
Morning ($j = 1$)	8h00-10h00 ($k = 1$)	22	22.000	0.000
	10h00-12h00 ($k = 2$)	22	22.000	0.000
Afternoon ($j = 2$)	12h00-15h00 ($k = 3$)	22	21.999	0.184
	15h00-18h00 ($k = 4$)	22	21.999	0.184
Evening ($j = 3$)	18h00-20h30 ($k = 5$)	22	22.001	0.184
	20h30-23h00 ($k = 6$)	22	22.001	0.184

Table 3.3
Units selected in a balanced sample for the first five days

H	Sector	Period	Subperiod	Site
1	Centre ($i = 2$)	Afternoon ($j = 2$)	12h00-15h00 ($k = 3$)	Chandler ($l = 6$)
			15h00-18h00 ($k = 4$)	Malbaie ($l = 5$)
		Evening ($j = 3$)	18h00-20h30 ($k = 5$)	Chandler ($l = 6$)
			20h30-23h00 ($k = 6$)	Ste-T. de Gaspé ($l = 4$)
2	East ($i = 1$)	Morning ($j = 1$)	8h00-10h00 ($k = 1$)	E. St-Jean ($l = 2$)
			10h00-12h00 ($k = 2$)	Boom Défense ($l = 1$)
		Evening ($j = 3$)	18h00-20h30 ($k = 5$)	Barachois ($l = 3$)
			20h30-23h00 ($k = 6$)	E. St-Jean ($l = 2$)
3	Centre ($i = 2$)	Morning ($j = 1$)	8h00-10h00 ($k = 1$)	Malbaie ($l = 5$)
			10h00-12h00 ($k = 2$)	Ste-T. de Gaspé ($l = 4$)
		Afternoon ($j = 2$)	12h00-15h00 ($k = 3$)	Malbaie ($l = 5$)
			15h00-18h00 ($k = 4$)	Chandler ($l = 6$)
4	West ($i = 3$)	Morning ($j = 1$)	8h00-10h00 ($k = 1$)	P. Henderson ($l = 8$)
			10h00-12h00 ($k = 2$)	Bonaventure ($l = 7$)
		Afternoon ($j = 2$)	12h00-15h00 ($k = 3$)	C. Carleton ($l = 9$)
			15h00-18h00 ($k = 4$)	C. Carleton ($l = 9$)
5	East ($i = 1$)	Afternoon ($j = 2$)	12h00-15h00 ($k = 3$)	Boom Défense ($l = 1$)
			15h00-18h00 ($k = 4$)	Barachois ($l = 3$)
		Evening ($j = 3$)	18h00-20h30 ($k = 5$)	Boom Défense ($l = 1$)
			20h30-23h00 ($k = 6$)	Barachois ($l = 3$)

3.2 Estimation of the fishing effort and of its variance

Once the survey is completed, the sample is a set of site × subperiod $\{(h, i, j, k, l)\}$ with sampling weights equal to the inverse of the selection probabilities given in (3.1). As the balancing equations for the 54 cells of the site by subperiod cross-classified table are not met exactly, we propose, following Deville and Tillé (2004), calibrating the survey weights on the total, H, of the indicator variables for these 54 cells. All the sampled units in cell (i, j, k, l) have the same weight, namely $1/\pi_{ijkl}$ where $\pi_{ijkl} = \pi_{hijkl}$, defined in (3.1), does not depend on h. The calibrated weight for a sampled unit in cell (i, j, k, l) is

$$w_{ijkl}^{(c)} = \frac{1}{\pi_{ijkl}} \times \frac{H}{n_{ijkl}/\pi_{ijkl}} = \frac{H}{n_{ijkl}},$$

where n_{ijkl} is the sample size for cell (i, j, k, l); it is the number of days for which site l of sector i has been visited during subperiod k of period j. In general n_{ijkl} is a random variable. When the samples are perfectly balanced, (2.2) implies that $n_{ijkl} = H\pi_{ijkl}$; the calibrated and basic weights are then equal. Now if y_{hijkl} represents the fishing effort for population unit (h, i, j, k, l), the fishing effort in cell (i, j, k, l) is $Y_{Uijkl} = \sum_h y_{hijkl}$. Its calibrated estimator is $\hat{Y}_{ijkl} = H\bar{y}_{sijkl}$ where $\bar{y}_{sijkl}$ is the average fishing effort for the

n_{ijkl} units sampled for that cell of the cross classified table. An estimator for the total fishing effort is obtained by summing the cells' estimated totals.

The evaluation of a design based variance estimator for the calibrated estimator of the total fishing effort is complex. A simple variance estimator for the estimated total for a single cell of the cross-classified table is available. The sample of days selected for cell (i, j, k, l) is a Bernoulli sample with selection probabilities π_{ijkl}, neglecting the balancing constraints. Thus by conditioning on the sample size, n_{ijkl}, $\hat{Y}_{ijkl}$ is H times the sample mean of a simple random sample. It is a design-unbiased estimator whose variance can be estimated using the formula for the variance of an estimated total in a simple random sampling design. We claim that these results are still valid when the balancing constraints are taken into account since the balanced sample design is invariant to a relabelling of the days. The estimated fishing efforts for the 54 cells of the cross-classified table are however dependent and it seems difficult to come up with a conditionally unbiased design based variance estimator for their total. A model based estimator seems to be only approach available for this total.

For the survey actually conducted in 2015, the methods used to estimate fishing effort and total catch are among those proposed in Pollock et al. (1994). It was a roving survey and the fishing effort at a sampled site was calculated as the average number of anglers on the site during the subperiod times the length, in hours, of the subperiod. Fishing efforts were estimated using calibrated weights; additional results are available in (Daigle, Crépeau, Bujold and Legault, 2015).

4 Comparison of the cube method and the rejective algorithm

Chauvet et al. (2015) have studied the cube method and the rejective algorithm by examining different aspects of these balancing techniques. They balanced on continuous auxiliary variables and they documented how the balancing algorithm impacted the selection probabilities and the sampling properties of estimators of population totals. The goal of this section is to compare the two sampling algorithms in a resource inventory where the balancing equations only involve indicator variables. This comparison is carried out in the context of a simplified creel survey with a stratified two stage design. The days represent strata $h = 1, \ldots, H$, the sectors are defined as primary units $i = 1, 2, 3$ and sites, indexed by j, are the secondary units. This sampling plan is similar to the design exposed in Section 3.1 except that periods and subperiods do not enter in the sampling design.

On each day two out of 3 sectors are selected and within each one 2 sites are sampled; thus 4 units are selected each day. The site importance variable x_{ij} determines the inclusion probabilities $\pi_{hij} = (2x_{i\bullet}/x_{\bullet\bullet}) \times (2x_{ij}/x_{i\bullet}) = \pi_{hi} \times \pi_{hj|i}$ for the two stages. As two out of three units are selected at each level, the joint selection probabilities are completely determined by $\{(\pi_{hi}, \pi_{hj|i}) : i, j = 1, 2, 3\}$ for the two stages; see the Appendix. If Z_{hij} stands for the indicator variables taking the value 1 if site (i, j) is sampled on day h and 0 otherwise then the entries of 9×9 variance covariance matrix for $\{Z_{hij} : i, j = 1, 2, 3\}$ are given by

$$\mathrm{Cov}\left(Z_{hij}, Z_{hi'j'}\right) = \begin{cases} \pi_{hij} - \pi_{hij}^2 & \text{if } i = i' \text{ and } j = j' \\ \pi_{hi}\ \pi_{hjj'|i} - \pi_{hij}\ \pi_{hij'} & \text{if } i = i' \text{ and } j \neq j' \\ \pi_{hii'}\ \pi_{hj|i}\ \pi_{hj'|i'} - \pi_{hij}\ \pi_{hi'j'} & \text{if } i \neq i' \end{cases} \quad (4.1)$$

where $\pi_{hii'}$ represents the joint selection probability of sectors i and i' on a single day, $\pi_{hj|i}$ is the probability for selecting site j, in sector i, at stage 2 and $\pi_{hjj'|i}$ is the joint selection probability of sites j and j' in sector i. All these probabilities are evaluated using the size measure x. Details are available in the appendix, see also Ousmane Ida (2016). The corresponding matrix $\mathrm{Var}(\tilde{n})$ in (2.3) is singular as one of the 9 constraints is redundant; thus in (2.3) a generalized inverse of the covariance matrix was used and γ^2, in (2.3), was set equal to 2.73 and 7.34, the 5^{th} and the 50^{th} percentiles of the χ_8^2 distribution.

4.1 Simulations on the comparison of the cube method and of the rejective algorithm

To investigate the impact of the algorithm on the sampling properties of survey estimators we simulated, for each unit, a fishing effort for site (i, j) on day h, y_{hij}, using independent Poisson random variables with mean $15 \times x_{ij}$. The total fishing effort for site (i, j) is then

$$Y_{Uij} = \sum_{h=1}^{H} y_{hij}.$$

A calibrated estimator, as defined in Section 3.2, for the fishing effort in site (i, j) is $\hat{Y}_{ij} = H\,\overline{y}_{sij}$, the average fishing effort for the n_{ij} units sampled at site (i, j) times H.

To compare the balancing algorithms, we used designs with $H = 12$ strata and two importance variables x, one with a small variation between site and one with a medium variation. Under each scenario we generated $B = 100{,}000$ random replications of a balanced sample by using the cube methods on one hand, and two rejective algorithms on the other. The inclusion probabilities for site (i, j) was estimated by

$$\hat{\pi}_{ij} = \frac{1}{B \times H} \sum_{b=1}^{B} n_{ij}^{(b)}.$$

This estimator assumes that the inclusion probabilities π_{hij} are constant in h. This holds true because the sample design is invariant to a relabelling of the days, see Section 3.1.

As argued in Section 3.2, the calibrated estimator $\hat{Y}_{ij}$ is design unbiased under the two selection algorithms. We compare their standard deviations,

$$\mathrm{Sd}_{\hat{Y}_{ij}} = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{Y}_{ij}^{(b)} - \overline{\hat{Y}}_{ij} \right)^2 \right\}^{1/2},$$

where $\overline{\hat{Y}}_{ij}$ is the average of the B simulated values. The sample size standard deviations were also calculated using (3.2). Observe that $\hat{\pi}_{ij} = \overline{n}_{ij} / H$. The simulation results are presented in Tables 4.1, 4.2 and 4.3.

Table 4.1
Comparison of the cube method (CM) and of two rejective algorithms (*R* 5% and *R* 50%) when *x* has a low variation

				CM		*R* 5%		*R* 50%	
Sector	**Site**	x_{ij}	π_{ij}	$\overline{\hat{\pi}}_{ij}$	$\mathrm{Sd}_{\hat{Y}_{ij}}$	$\overline{\hat{\pi}}_{ij}$	$\mathrm{Sd}_{\hat{Y}_{ij}}$	$\overline{\hat{\pi}}_{ij}$	$\mathrm{Sd}_{\hat{Y}_{ij}}$
$i = 1$	$j = 1$	3	0.500	0.500	16.56	0.503	16.86	0.505	17.40
	$j = 2$	2	0.333	0.333	22.20	0.329	23.35	0.328	25.07
	$j = 3$	3	0.500	0.500	23.99	0.503	24.47	0.505	25.15
$i = 2$	$j = 4$	2	0.333	0.333	25.80	0.329	26.93	0.326	29.11
	$j = 5$	2	0.333	0.333	33.97	0.329	35.54	0.326	38.28
	$j = 6$	2	0.333	0.333	27.65	0.329	28.87	0.326	31.10
$i = 3$	$j = 7$	3	0.500	0.500	22.50	0.502	22.88	0.502	23.66
	$j = 8$	3	0.500	0.500	20.02	0.502	20.20	0.502	20.94
	$j = 9$	4	0.667	0.667	22.01	0.674	21.98	0.679	22.25

Table 4.2
Comparison of the cube method (CM) and of two rejective algorithms (*R* 5% and *R* 50%) when *x* has a medium variation

				CM		*R* 5%		*R* 50%	
Sector	**Site**	x_{ij}	π_{ij}	$\overline{\hat{\pi}}_{ij}$	$\mathrm{Sd}_{\hat{Y}_{ij}}$	$\overline{\hat{\pi}}_{ij}$	$\mathrm{Sd}_{\hat{Y}_{ij}}$	$\overline{\hat{\pi}}_{ij}$	$\mathrm{Sd}_{\hat{Y}_{ij}}$
$i = 1$	$j = 1$	3	0.500	0.500	25.52	0.505	25.78	0.507	26.60
	$j = 2$	2	0.333	0.333	25.25	0.330	26.26	0.329	28.16
	$j = 3$	3	0.500	0.500	21.12	0.505	21.36	0.507	22.03
$i = 2$	$j = 4$	1	0.167	0.167	29.17	0.158	32.45	0.149	31.19
	$j = 5$	2	0.333	0.333	13.73	0.329	14.38	0.326	15.49
	$j = 6$	2	0.333	0.333	32.82	0.329	34.22	0.326	36.91
$i = 3$	$j = 7$	2	0.333	0.333	16.84	0.329	17.52	0.325	18.85
	$j = 8$	4	0.667	0.667	18.68	0.672	18.70	0.678	18.89
	$j = 9$	5	0.833	0.833	8.06	0.844	7.81	0.854	7.67

Table 4.3
Standard deviations of the sample sizes obtained with the cube method (CM) and with two rejective algorithms (*R* 5%, *R* 50%)

		x has a low variation				*x* has a medium variation			
Sector	**Site**	***x***	**CM**	***R* 5%**	***R* 50%**	***x***	**CM**	***R* 5%**	***R* 50%**
$i = 1$	$j = 1$	3	0.000	0.894	1.371	3	0.000	0.891	1.371
	$j = 2$	2	0.000	0.854	1.295	2	0.000	0.831	1.294
	$j = 3$	3	0.000	0.896	1.377	3	0.000	0.891	1.374
$i = 2$	$j = 4$	2	0.130	0.828	1.293	1	0.144	0.654	1.013
	$j = 5$	2	0.195	0.832	1.298	2	0.170	0.831	1.290
	$j = 6$	2	0.179	0.826	1.296	2	0.141	0.830	1.297
$i = 3$	$j = 7$	3	0.339	0.859	1.366	2	0.342	0.835	1.294
	$j = 8$	3	0.381	0.859	1.367	4	0.350	0.807	1.294
	$j = 9$	4	0.319	0.822	1.288	5	0.248	0.655	1.010

In Tables 4.1 and 4.2, the cube method maintains the selection probabilities and yields a total estimator with the smallest standard deviations. Taking γ^2 equal to the 50^{th} percentile of the χ^2_8 distribution for the rejective algorithm yields the poorer results, both in terms of selection probabilities and of the standard deviations of $\overline{y}_{sij}$. The largest biases for the selection probabilities occur at the extreme x values in Table 4.2. The selection probability for site $j = 4$ is underestimated by 11% with the rejective method based on the 50^{th} percentile and by 5% with the 5^{th} percentile. The probability is over estimated in the sites with the large values for x.

In Tables 4.1 and 4.2, the standard deviation for $\hat{Y}_{ij}$ is, in most cases, smallest for the cube method and largest for the rejection algorithm based on the 50^{th} percentile. The standard deviations for the rejective algorithm are up to 10% larger than the ones for the cube method. In Table 4.2, the largest gain in efficiency of the cube method with respect to the R 5% rejective algorithm (equal to the ratio of standard deviations squared) is 23%; it occurs when $j = 4$ and $x = 1$. These standard deviations are driven by the variability in sample sizes n_{ij}. Table 4.3 gives the sample sizes' standard deviations. Since the expected number of visits to sector 1 and to sites 1, 2, and 3 are integers, the cube method is able to get sample sizes equal to their expectations for this sector and the sample sizes standard deviations are 0. This is not possible in sectors 2 and 3 as the expected sample sizes for these sectors are not integer valued. In general, the rejective algorithms give sample sizes whose standard deviations are much more variable than those for the cube method. This makes the rejective algorithm total estimators more variable than those obtained with the cube method.

The conditional variance estimator for fishing effort $\hat{Y}_{ij}$ in site (i, j) proposed in Section 3.2 is

$$v\left(\hat{Y}_{ij}\right) = \frac{H^2\left(1 - n_{ij}/H\right)}{n_{ij}} \sum_{h \in s_{ij}} \frac{\left(y_{hij} - \overline{y}_{sij}\right)^2}{n_{ij} - 1}.$$

The conditional sampling properties, given n_{ij}, of this variance estimator were investigated in the Monte Carlo study with $B = 10{,}000$ balanced samples for the three sample designs. For each site and for each sample size n_{ij} the conditional variance $\text{Var}\left(\hat{Y}_{ij} \mid n_{ij}\right)$ and the conditional expectation of the variance estimator $\text{E}\left\{v\left(\hat{Y}_{ij}\right)\right\}$ were evaluated using the Monte Carlo samples for which the sample size for site (i, j) was n_{ij}. The conditional relative bias of the variance estimator, $\text{E}\left\{v\left(\hat{Y}_{ij}\right)\right\} \Big/ \text{Var}\left(\hat{Y}_{ij} \mid n_{ij}\right) - 1$, was then calculated. The conditional relative biases were then aggregated by weighting each sample size n_{ij} using its frequency in the 10,000 Monte Carlo samples; the results are in Table 5.1.

In Table 5.1, the aggregated relative biases are less than 3% in absolute value for the three selection algorithms. This validates the conditional variance estimator proposed in Section 3.2 for a single cell of the cross-classified table. The conditional variances of sums such as $\hat{Y}_{ij} + \hat{Y}_{ij'}$ is more complicated as it involves joint selection probabilities; the estimation of these variances is not considered here. See Breidt and Chauvet (2011) for a discussion of variance estimation with the cube method.

Table 5.1
Aggregated conditional bias, in percentage, of the conditional variance estimator $v\left(\hat{Y}_{ij}\right)$ obtained with the cube method and two rejective algorithms (R 5%, R 50%)

		x has a low variation				x has a medium variation			
Sector	**Site**	x	**CM**	R 5%	R 50%	x	**CM**	R 5%	R 50%
$i = 1$	$j = 1$	3	1	-3	3	3	-1	1	1
	$j = 2$	2	2	-1	-2	2	3	1	-2
	$j = 3$	3	-1	0	1	3	0	-1	0
$i = 2$	$j = 4$	2	-2	2	0	1	1	-1	-2
	$j = 5$	2	1	-1	-1	2	2	2	3
	$j = 6$	2	0	3	-2	2	0	0	-3
$i = 3$	$j = 7$	3	1	-3	2	2	0	-3	-1
	$j = 8$	3	2	1	1	4	0	0	0
	$j = 9$	4	-1	1	-2	5	-2	-1	1

The conclusion of this Monte Carlo investigation is that the rejective algorithm changes the selection probabilities: sites with small importance are under represented in the rejective samples while the cube method is very good at preserving the selection probabilities. Under both algorithms the calibrated estimator for the total of y in a domain is unbiased. Smaller variances are however obtained with the cube algorithm as it gives domain sample sizes that are less variable than the rejective algorithm.

5 Discussion

In the context of creel surveys, balanced sampling techniques such as the cube method or the rejective algorithm are used to ensure a predetermined sample size in small domains of the survey population. The cube method is very effective at doing so especially in complex survey designs with several stages of sampling. It does not change the selection probabilities and it yields domain sample sizes that are very close their target values. The rejective method, on the other hand, changes the selection probabilities slightly and produce domain sample sizes that are more variable. With a large number of constraints, Fuller's rejective sampling scheme is not really applicable as it requires the evaluation and the inversion of a large covariance matrix in (2.3); alternative acceptation criteria for a sample need to be investigated.

Acknowledgements

We thank the Associate Editor and the referees for their constructive comments on the first version of this manuscript. The assistance and the suggestions of Valérie Bujold, Michel Legault and of Hélène Crépeau who participated to the initial phase of this project is gratefully acknowledged. This project benefitted from the financial assistance of the Canada Research Chair in Statistical Sampling and Data Analysis and form a discovery grant (5244/2012) from the Natural Sciences and Engineering Research Council of Canada.

Appendix

Calculation of the joint selection probabilities when $N = 3$

Consider a population of size 3 and let π_1, π_2, and π_3 be the marginal selection probabilities when drawing a sample of size $n = 2$. The joint selection probabilities π_{ij}, $i \neq j = 1, 2, 3$ satisfy

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} \pi_{12} \\ \pi_{13} \\ \pi_{23} \end{pmatrix} = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \end{pmatrix}.$$

Thus

$$\begin{pmatrix} \pi_{12} \\ \pi_{13} \\ \pi_{23} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & 1 & -1 \\ 1 & -1 & 1 \\ -1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \end{pmatrix}.$$

Using these equations, the entries of the covariance matrix (4.1) can be evaluated using the stage 1 and the stage 2 selection probabilities.

References

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.

Chauvet, G. (2009). Stratified balanced sampling. *Survey Methodology*, 35, 1, 115-119. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2009001/article/10888-eng.pdf.

Chauvet, G., Haziza, D. and Lesage, É. (2015). Examining some aspects of balanced sampling in surveys. *Statistica Sinica*.

Daigle, G., Crépeau, H., Bujold, V. and Legault, M. (2015). Enquête de la pêche sportive au bar rayé en Gaspésie en 2015. Technical report.

Deville, J.-C., and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91(4), 893-912.

Fuller, W.A. (2009). Some design properties of a rejective sampling procedure. *Biometrika*, 96(4), 933-944.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, 35, 1491-1523.

Hasler, C., and Tillé, Y. (2014). Fast balanced sampling for highly stratified population. *Computational Statistics & Data Analysis*, 74, 81-94.

Hoenig, J.M., Jones, C.M., Pollock, K.H., Robson, D.S. and Wade, D.L. (1997). Calculation of catch rate and total catch in roving surveys of anglers. *Biometrics*, 306-317.

Legg, J.C., and Yu, C.L. (2010). A comparison of sample set restriction procedures. *Survey Methodology*, 36, 1, 69-79. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2010001/article/11249-eng.pdf.

Minnesota Department of Natural Resources (2011). Creel surveys.

Ousmane Ida, I. (2016). L'échantillonnage équilibré par la méthode du cube et la méthode réjective. Master's thesis, Université Laval.

Pollock, K., Jones, C. and Brown, T. (1994). Angler survey methods and their applications in fisheries management. *American Fisheries Society special publication (USA)*.

Robson, D., and Jones, C.M. (1989). The theoretical basis of an access site angler survey design. *Biometrics*, 83-98.

Tillé, Y. (2011). *Sampling Algorithms*. New York: Springer.

United States Environmental Protection Agency (1998). *Guidance for Conducting Fish and Wildlife Consumption Surveys*. EPA, Washington, DC.

Vallée, A.-A., Ferland-Raymond, B., Rivest, L.-P. and Tillé, Y. (2015). Incorporating spatial and operational constraints in the sampling designs for forest inventories. *Environmetrics*, 26(8), 557-570.