BIOMETRICS 55, 1014–1021
December 1999

The Stratified Petersen Estimator with a Known Number of Unread Tags

Carl J. Schwarz,[1,*] Myra Andrews,[1] and Michael R. Link[2]

[1]Department of Statistics and Mathematics, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada
[2]LGL Limited, Sidney, British Columbia V8L 3Y8, Canada
*email: cschwarz@cs.sfu.ca

SUMMARY. The Petersen estimator estimator of abundance can be biased when the assumption of homogeneous capture probability or homogeneous recapture probability is violated. Often this heterogeneity is related to the time or place of capture or recapture, and if these can be stratified, the stratified Petersen estimator reduces the bias caused by this heterogeneity. In some experiments, not all the recovered tagged animals can be examined, and only a subsample has its stratum of release and recovery determined. We develop methods for this modified experiment and apply them to estimate the number of salmon returning to spawn in a river in British Columbia, Canada.

KEY WORDS: Capture–recapture; Mark–recapture; Petersen estimator; Salmon escapement; Salmon spawning.

1. Introduction

The Petersen estimator of abundance can be severely biased if the assumption of equal capture or of equal recapture probabilities among all animals is violated. Often heterogeneity in capture or recapture probabilities is related to the time when or the place where the animals are captured or recaptured. In these cases, the stratified Petersen method (Darroch, 1961; Seber, 1982; Plante, 1990; Banneheka, Routledge, and Schwarz, 1997; Plante, Rivest, and Tremblay, 1998; Schwarz and Taylor, 1998) can be used.

In this method, in each of s strata, a known number of animals (n_i^c) are captured, tagged with individually identifiable tags, and released. Recaptures take place in each of t strata. Tagged animals have their tags read, and the tags identify the stratum of release and of recapture so that the number (m_{ij}) of individuals marked in stratum i and recovered in stratum j may be determined. A count of untagged animals recovered in each of the recovery strata (u_j) is also made.

Normally, all tagged animals that are recaptured have their tags read. However, in some cases, the number of tagged animals recovered is too large or it is logistically impossible to read all of the tags. For example, our paper was motivated by a study where salmon are tagged as they return to spawn and recoveries are made by counting the number of tagged and untagged fish as they pass through a fish-way at a dam upstream. The tags are readily visible so that it is easy to count the number of tagged fish passing but it is difficult to extract tagged fish using a dip net to read the tags, so only a fraction of the tags can be read to determine the tag number and hence when the fish were initially tagged. In general, a count of the total number of tagged animals in the recovery samples is known, but only a random subsample of the recovered tagged animals have their tags read, leaving z_j animals whose stratum of release is unknown.

In this paper, we extend the Darroch–Plante model for the stratified Petersen to account for a known number of tags recovered but not read. We apply this model to estimate the number of salmon returning to spawn (the salmon escapement) in the Nass River, British Columbia, Canada.

2. Notation

We consider the case of s release strata and t recovery strata. A dot in a subscript implies summation over that subscript, e.g., $m_{i\bullet} = \Sigma_{j=1}^{t} m_{ij}$.

2.1 *Parameters*

- N total number of animals in the population over all release strata.
- ψ_i proportion of population in release stratum i ($i = 1, \ldots, s$). $\Sigma_{i=1}^{s} \psi_i = 1$.
- N_i number of animals in release stratum i ($i = 1, \ldots, s$). $N_i = N\psi_i = (\gamma_i + \Sigma_{j=1}^{t} \mu_{ij}\lambda_j)(1 + \beta_i)$. $\Sigma_{i=1}^{s} N_i = N$. Note that the sets $\{N, \{\psi_i\}\}$ and $\{N_i\}$ are two different but equivalent parameterizations for the number of animals in the release strata.
- c_i, r_j probability of initial capture in release stratum i ($i = 1, \ldots, s$) or capture in recovery stratum j ($j = 1, \ldots, t$).
- β_i odds of not capturing an animal in release stratum i ($i = 1, \ldots, s$). $\beta_i = (1 - c_i)/c_i$.
- θ_{ij} probability that an animal released in stratum i moves to recovery stratum j ($i = 1, \ldots, s$; $j = 1, \ldots, t$). We allow $\Sigma_{j=1}^{t} \theta_{ij} \leq 1$ to account for movement to areas

not sampled or for mortality between release and recovery.

μ_{ij} expected number of animals that are released in stratum i, move to recovery stratum j, and are recaptured ($i = 1,\ldots,s$; $j = 1,\ldots,t$). $\mu_{ij} = N\psi_i c_i \theta_{ij} r_j$.

λ_j probability that a tagged animal recovered in stratum j will have its tag read ($j = 1,\ldots,t$).

γ_i expected number of animals released in stratum i that do not have their tags read ($i = 1,\ldots,s$). $\gamma_i = N\psi_i c_i - \Sigma_{j=1}^{t} \mu_{ij}\lambda_j$.

2.2 *Statistics*

n_i^c number of animals captured, tagged, and released in stratum i ($i = 1,\ldots,s$).

m_{ij} number of animals tagged and released in stratum i that are recovered in stratum j and have their tags read ($i = 1,\ldots,s$; $j = 1,\ldots,t$).

u_j number of animals recovered in stratum j without tags ($j = 1,\ldots,t$).

z_j number of tagged animals recovered in stratum j whose tags are not read ($j = 1,\ldots,t$).

3. Model Development and Fitting

The statistics for this experiment can be arranged into a rectangular array as shown in Table 1. The key difference between this experiment and the usual stratified Petersen experiment is that the stratum of release for the $\{z_j\}$ tagged animals recovered is not known.

We make the same assumptions as outlined in Darroch (1961), Seber (1982), Plante (1990), Banneheka et al. (1997), Schwarz and Taylor (1998), and Plante et al. (1998). In addition, we assume that the animals selected to have their tags read are a random sample from all tagged animals recovered. The expected values of the statistics are shown in Table 2.

Plante (1990) and Plante et al. (1998) showed that, in the ordinary stratified Petersen (where all tags are read), not all parameters are identifiable. We can rewrite the expected values shown in Table 2 in terms of new identifiable parameters as shown in Table 3. (Note that our γ_i is defined slightly differently than in these previous works.) In the case where $s \leq t$, the population sizes in the release strata are identifiable even if the population is not closed, i.e., some animals may leave the population (e.g., die between release and recovery or emigrate to recovery strata not sampled). This corresponds to many real situations as exemplified by Schwarz and Taylor (1998). (In the case $s \geq t$, only the population size in the recovery strata are estimable.)

We chose to parameterize in terms of the s parameters $\{\gamma_i\}$ rather than the set $\{N, \{\psi_i\}_{i=1,\ldots,s}\}$ (recall that $\Sigma_{i=1}^{s} \psi_i = 1$) because, as will be shown later, the estimating equations are simpler. There is a one-to-one transformation between these two parameter sets.

Up to this point, the development has paralleled that in Plante (1990) and Plante et al. (1998). The distribution of the statistics $\{m_{ij}\}$, $\{n_i^c - m_{i\bullet}\}$, and $\{u_j\}$ can be modelled either as multinomial counts conditional on the $n_\bullet^c$ using the expectations as shown in Table 3 (as was done by Plante [1990] and Plante et al. [1998]) or as independent Poisson counts—both lead to the same maximum likelihood estimates (MLEs).

However, at this point, difficulties arise in developing a likelihood function. The distribution of $\{z_j\}$ conditional on the set $\{n_i^c - m_{i\bullet}\}$ is a convolution of s multinomial distributions with complex cell probabilities and involves a high-dimensional summation that is not amenable to inclusion in a likelihood, i.e, this portion of the likelihood is

$$
\begin{aligned}
&L(\{z_j\} \mid \{n_i^c - m_{i\bullet}\}, \{\lambda_j\}, \{\mu_{ij}\}, \{\gamma_i\}) \\
&\quad = \sum_{z_{11}^*} \cdots \sum_{z_{st}^*} \prod_{i=1}^{s} \binom{n_i^c - m_{i\bullet}}{z_{i1}^*, \ldots, z_{it}^*, n_i^c - m_{i\bullet} - z_{i\bullet}^*} \\
&\qquad \times \left[\prod_{j=1}^{t} \left(\frac{\mu_{ij}(1-\lambda_j)}{\gamma_i} \right)^{z_{ij}^*} \right] \\
&\qquad \times \left(\frac{\gamma_i - \sum_{j=1}^{t} \mu_{ij}\left(1-\lambda_j\right)}{\gamma_i} \right)^{n_i^c - m_{i\bullet} - z_{i\bullet}^*}
\end{aligned}
$$

subject to $z_{\bullet j}^* = z_j$ and $z_{i\bullet}^* \leq n_i^c - m_{i\bullet}$.

Table 1
Observed statistics

Release stratum	Number released	Recovery stratum 1	2	...	t	Tags not read
1	n_1^c	m_{11}	m_{12}	...	m_{1t}	$n_1^c - m_{1\bullet}$
2	n_2^c	m_{21}	m_{22}	...	m_{2t}	$n_2^c - m_{2\bullet}$
...	...	...	...	...	...	...
s	n_s^c	m_{s1}	m_{s2}	...	m_{st}	$n_s^c - m_{s\bullet}$
Number recovered without tags		u_1	u_2	...	u_t	
Number tags recovered but not read		z_1	z_2	...	z_t	

Note: Number of tags not recovered = $n_\bullet^c - m_{\bullet\bullet} - z_\bullet$.

Table 2
Expected values of statistics shown in Table 1

Release stratum	Number released	Recovery stratum 1	...	Recovery stratum t	Tags not read
1	$N\psi_1 c_1$	$N\psi_1 c_1 \theta_{11} r_1 \lambda_1$	...	$N\psi_1 c_1 \theta_{1t} r_t \lambda_t$	$N\psi_1 c_1 - \sum_{j=1}^{t} N\psi_1 c_1 \theta_{1j} r_j \lambda_j$
2	$N\psi_2 c_2$	$N\psi_2 c_2 \theta_{21} r_1 \lambda_1$	...	$N\psi_2 c_2 \theta_{2t} r_t \lambda_t$	$N\psi_2 c_2 - \sum_{j=1}^{t} N\psi_2 c_2 \theta_{2j} r_j \lambda_j$
...	...	...	...	...	...
s	$N\psi_s c_s$	$N\psi_s c_s \theta_{s1} r_1 \lambda_1$	...	$N\psi_s c_s \theta_{st} r_t \lambda_t$	$N\psi_s c_s - \sum_{j=1}^{t} N\psi_s c_s \theta_{sj} r_j \lambda_j$
Number recovered without tags		$\sum_{i=1}^{s} N\psi_i (1-c_i)\theta_{i1} r_1$	...	$\sum_{i=1}^{s} N\psi_i (1-c_i)\theta_{it} r_t$	
Number tags recovered but not read		$\sum_{i=1}^{s} N\psi_i c_i \theta_{i1} r_1 (1-\lambda_1)$	...	$\sum_{i=1}^{s} N\psi_i c_i \theta_{it} r_t (1-\lambda_t)$	

Note: Expected number of tags not recovered $= \sum_{i=1}^{s} N\psi_i c_i \left[1 - \sum_{j=1}^{t} \theta_{ij} r_j\right]$.

For these reasons, we use a generalized estimating equation (GEE) approach (Liang and Zeger, 1986). GEEs have been used in capture–recapture contexts (e.g., Becker, 1984; Yip, 1991) through estimating functions based on martingales describing the capture and release process. In this context, we will derive estimating equations based on the moments of the observed statistics.

Using the GEE approach, the estimates are derived as the solutions to $\mathbf{D}^{\mathbf{T}}\mathbf{V}^{-1}(\mathbf{Y} - \boldsymbol{\pi}) = \mathbf{0}$, where $\mathbf{Y}$ is a vector of observed statistics, $\boldsymbol{\pi}$ is a vector of the expected values of $\mathbf{Y}$, $\mathbf{V}$ is a working covariance matrix of the observed statistics, and $\mathbf{D}$ is a matrix of partial derivatives $\partial\boldsymbol{\pi}/\partial\boldsymbol{\omega}'$, where $\boldsymbol{\omega}$ is the parameter set. The estimates are consistent even if $\mathbf{V}$ is not the true covariance matrix of the observed statistics and will be efficient if $\mathbf{V}$ is close to the true covariance of $\mathbf{Y}$.

In this case, the $(st + s + 2t + 1)$ components of $\mathbf{Y}$ are st components referring to $\{m_{ij}\}$, s components referring to $\{n_i^c - m_{i\bullet}\}$, t components referring to $\{u_j\}$, t components

Table 3
Expected value of observed statistics in terms of the identifiable parameters

Release stratum	Number released	Recovery stratum 1	...	Recovery stratum t	Tags not read
1	$\gamma_1 + \sum_{j=1}^{t} \mu_{1j}\lambda_j$	$\mu_{11}\lambda_1$	...	$\mu_{1t}\lambda_t$	γ_1
2	$\gamma_2 + \sum_{j=1}^{t} \mu_{2j}\lambda_j$	$\mu_{21}\lambda_1$	...	$\mu_{2t}\lambda_t$	γ_2
...	...	...	...	...	...
s	$\gamma_s + \sum_{j=1}^{t} \mu_{sj}\lambda_j$	$\mu_{s1}\lambda_1$	...	$\mu_{st}\lambda_t$	γ_s
Number recovered without tags		$\sum_{i=1}^{s} \beta_i \mu_{i1}$	...	$\sum_{i=1}^{s} \beta_i \mu_{it}$	
Number tags recovered but not read		$(1-\lambda_1)\mu_{\bullet 1}$	...	$(1-\lambda_t)\mu_{\bullet t}$	

Note: Expected number of tags not recovered $= \gamma_\bullet - \sum_{j=1}^{t} (1-\lambda_j)\mu_{\bullet j}$.

referring to $\{z_j\}$, and one component referring to $n_\bullet^c - m_{\bullet\bullet} - z_\bullet$. The components of $\boldsymbol{\pi}$ refer to the expectations of these statistics as shown in Table 3. The components of $\boldsymbol{\omega}$ refer to the $st + 2s + t$ parameters, $\{\mu_{ij}\}, \{\gamma_i\}, \{\beta_i\}$, and $\{\lambda_j\}$. The working covariance matrix is a diagonal matrix with diagonal elements corresponding to the expected values. This makes the assumption that the observed statistics are approximately Poisson distributed. This formulation leads to estimates that minimize the usual Pearson χ^2 measure of goodness-of-fit, i.e., minimum χ^2 estimates. Consequently, we expect our estimators to be nearly as efficient as the true MLEs.

This gives rise to the following estimating equations:

$$0 = \left(\frac{m_{ij}}{\mu_{ij}\lambda_j} - 1\right)\lambda_j + \left(\frac{u_j}{\sum_{a=1}^{s}\beta_a\mu_{aj}} - 1\right)\beta_i + \left(\frac{z_j}{(1-\lambda_j)\,\mu_{\bullet j}} - 1\right)(1-\lambda_j) - \left(\frac{n_\bullet^c - m_{\bullet\bullet} - z_\bullet}{\gamma_\bullet - \sum_{a=1}^{s}\sum_{b=1}^{t}(1-\lambda_b)\,\mu_{ab}} - 1\right)(1-\lambda_j)$$

for $\{\mu_{ij}\}$, $i = 1, \ldots, s$, $j = 1, \ldots, t$;

$$0 = \left(\frac{n_i^c - m_{i\bullet}}{\gamma_i} - 1\right) + \left(\frac{n_\bullet^c - m_{\bullet\bullet} - z_\bullet}{\gamma_\bullet - \sum_{a=1}^{s}\sum_{b=1}^{t}(1-\lambda_b)\,\mu_{ab}} - 1\right)$$

for $\{\gamma_i\}$, $i = 1, \ldots, s$;

$$0 = \sum_{j=1}^{t}\left[\left(\frac{u_j}{\sum_{a=1}^{s}\beta_a\mu_{aj}} - 1\right)\mu_{ij}\right]$$

for $\{\beta_i\}$, $i = 1, \ldots, s$;

$$0 = \sum_{i=1}^{s}\left[\left(\frac{m_{ij}}{\mu_{ij}\lambda_j} - 1\right)\mu_{ij}\right] - \left(\frac{z_j}{(1-\lambda_j)\,\mu_{\bullet j}} - 1\right)\mu_{\bullet j} + \left(\frac{n_\bullet^c - m_{\bullet\bullet} - z_\bullet}{\gamma_\bullet - \sum_{a=1}^{s}\sum_{b=1}^{t}(1-\lambda_b)\,\mu_{ab}} - 1\right)\mu_{\bullet j}$$

for $\{\lambda_j\}$, $j = 1, \ldots, t$.

This system of equations does not have an analytical solution and must be solved numerically. However, two approximate solutions are $\hat{\gamma}_i \approx n_i^c - m_{i\bullet}$ and $\hat{\lambda}_j \approx m_{\bullet j}/\,(m_{\bullet j} + z_{\bullet j})$. The former is analogous to the results in Plante et al. (1998), and the latter is intuitively appealing because it represents the ratio of the number of tags read to the total number of tags recovered in each recovery stratum.

Under this approximate solution, the remaining equations reduce to

$$0 = \frac{m_{ij}}{\mu_{ij}} + \frac{u_j\beta_i}{\sum_{a=1}^{s}\beta_a\mu_{aj}} - \beta_i + \frac{z_j}{\mu_{\bullet j}} - 1$$

$$0 = \sum_{j=1}^{t}\left[\left(\frac{u_j}{\sum_{a=1}^{s}\beta_a\mu_{aj}} - 1\right)\mu_{ij}\right],$$

which again are analogous to those in Plante (1990), with her m_{ij} replaced by $m_{ij} + z_j(\mu_{ij}/\mu_{\bullet j})$, representing an estimate of the number of tags released in stratum i and recovered in stratum j had all tags been read. The last equation has the same form as Plante (1990). As outlined in Schwarz and Taylor (1998), the $\{\beta_i\}$ are found to minimize the discrepancy between the u_j row and the linear combination of the rows of the μ_{ij} matrix. Initial values for the β_i can be found using a least-squares approach (Banneheka et al., 1997) after adjusting the m_{ij} to account for unread tags.

The variances of the estimated parameters are found as

$$V(\hat{\boldsymbol{\omega}}) = \left[\mathbf{D}^{\mathbf{T}}\mathbf{V}^{-1}\mathbf{D}\right]^{-1}\left[\mathbf{D}^{\mathbf{T}}\mathbf{V}^{-1}V(\mathbf{Y})\mathbf{V}^{-1}\mathbf{D}\right] \times \left[\mathbf{D}^{\mathbf{T}}\mathbf{V}^{-1}\mathbf{D}\right]^{-1},$$

where $V(\mathbf{Y})$ is the true variance–covariance matrix of $\mathbf{Y}$. The $V(\mathbf{Y})$ can be readily derived despite the difficulties in writing a likelihood for the z_j. The variance of the estimates can be estimated by replacing the parameters by their estimates in the above equation.

Estimates of the initial stratum sizes are then found as $\hat{N}_i = (\hat{\gamma}_i + \Sigma_{j=1}^{t}\hat{\mu}_{ij}\hat{\lambda}_j)(1 + \hat{\beta}_i)$. The estimated total population size is found by summing the individual stratum estimates as $\hat{N} = \Sigma_{i=1}^{s}(\hat{\gamma}_i + \Sigma_{j=1}^{t}\hat{\mu}_{ij}\hat{\lambda}_j)(1 + \hat{\beta}_i)$. The estimated variance of these derived parameters is found using the delta method and the estimated variances of the original parameters.

Note that $\hat{\gamma}_i \approx n_i^c - m_{i\bullet}$, $\widehat{\mu_{ij}\lambda_j} \approx m_{ij}$, and $1 + \hat{\beta}_i = 1/\hat{c}_i$, so that $\hat{N} \approx \Sigma_{i=1}^{s} n_i^c/\hat{c}_i$, a Horvitz–Thompson-type estimator. Huggins (1989) and Yip, Huggins, and Lin (1996) examined the performance of similar types of estimators for closed populations and found that these performed poorly when sample sizes were small and the capture probabilities poorly estimated. This is also expected to be true for our estimators, as confirmed by our simulation study.

A goodness-of-fit statistic can be found using a Pearson-type statistic as

$$\mathrm{X}^2 = \sum_{i=1}^{s}\sum_{j=1}^{t}\frac{\left(m_{ij} - \hat{\mu}_{ij}\hat{\lambda}_j\right)^2}{\hat{\mu}_{ij}\hat{\lambda}_j} + \sum_{i=1}^{s}\frac{(n_i^c - m_{i\bullet} - \hat{\gamma}_i)^2}{\hat{\gamma}_i} + \sum_{j=1}^{t}\frac{\left(u_j - \sum_{a=1}^{s}\hat{\beta}_a\hat{\mu}_{aj}\right)^2}{\sum_{a=1}^{s}\hat{\beta}_a\hat{\mu}_{aj}} + \sum_{j=1}^{t}\frac{\left(z_j - \hat{\mu}_{\bullet j}(1-\hat{\lambda}_j)\right)^2}{\hat{\mu}_{\bullet j}(1-\hat{\lambda}_j)}$$

$$+\frac{\left(n_{\bullet}^{c}-m_{\bullet\bullet}-z_{\bullet}-\left(\hat{\gamma}_{\bullet}-\sum_{a=1}^{s}\sum_{b=1}^{t}\hat{\mu}_{ab}(1-\hat{\lambda}_{b})\right)\right)^{2}}{\hat{\gamma}_{\bullet}-\sum_{a=1}^{s}\sum_{b=1}^{t}\hat{\mu}_{ab}(1-\hat{\lambda}_{b})}.$$

This will have an approximate χ^2_{t-s} distribution. Not all of the cells are independent of each other because there is some double counting of animals between the z_j and the $n_i - m_{i\bullet}$. However, the z_j will be conditionally independent of the other cells and so, in large samples, the approximation should be valid.

Models where constraints are imposed on the parameters (e.g., equal tag reading rate in all recovery strata) can be fit by using variants of the chain rule.

Because there is no formal likelihood, there is no simple procedure for model selection and testing except, perhaps, comparing the change in the goodness-of-fit statistic. Williams (1970) outlined a bootstrap-type procedure for this that could be used to discriminate between models. Wald-type tests could also be constructed using the estimates and the estimated variance–covariance matrix.

As noted by Schwarz and Taylor (1998), extensive pooling of the rows or column counts may be required, and there is no objective method of assessing which poolings are optimal. Schwarz and Taylor (1998) also noted that the estimates of β_i are very unstable if the expected counts are small, leading to unstable estimates for N. They suggest that the expected values of m_{ij} should not be less than 5–10.

Specialized software to fit the above model using the estimating equations has been written in S-Plus and is available from the first author. This software also will fit simpler models where the tag-reading rate is assumed to be equal for all the recovery strata, where the tag application rate is equal for all release strata, or where the parameters can be modelled using covariates. In the middle model above, the estimate of the total population size is algebraically equal to the simple Petersen estimator, which is known to be consistent when all the initial capture probabilities are equal.

In the absence of specialized software for this experiment, approximate solutions can be found using software for the stratified Petersen with all tags read (e.g., SPAS from Arnason et al., 1996) by initially distributing the unread tags into their respective rows in the same proportion as the m_{ij} to their column sums. Then SPAS can be used to obtain estimates of the μ_{ij}, which can then be used to revise estimates to reapportion the z_j. This should be repeated until convergence. Of course, the final variance estimates reported by SPAS will be too small.

4. Example

The Nass River is located in northern British Columbia, Canada, and supports several stocks of sockeye salmon. From mid-June to early September 1995, daily samples of up to 400 returning sockeye salmon were captured with two fish wheels, tagged with individually numbered tags, and released back to the river (Link and Gurak, 1997). The fish migrate up river to spawn at several sites. The fish arrive at the entrance to one spawning area, Meziadin Lake, approximately 3 weeks later. Here fish travel through a fishway and counting chute, where they can be observed, counted, and removed, using a dip-net, to read their tags. Approximately 1500–10,000 fish pass through the fishway each day. Because the tags are quite visible, it is relatively easy to count the number of tagged fish, but it is more difficult to recapture the fish, and not all tagged fish can be captured.

Here, the population is not closed, as fish can be removed by a fishery after passing the fish wheels to spawn in many other sites not sampled. However, we can still obtain estimates of the number of fish passing the fish wheels. Because the m_{ij} matrix based on the daily counts is quite sparse, releases and recoveries were pooled into 8 release and 10 recovery strata as shown in Table 4.

The estimates of the parameters from the full model with separate initial capture probabilities for each release stratum and separate tag reading rates for each recovery stratum are shown in Tables 5 and 6. Although the goodness-of-fit statistic shows some lack of fit ($\chi^2 = 16.4$ with 2 d.f.), the majority of

Table 4

Summary statistics for sockeye salmon returning to spawn in the Nass River, British Columbia, Canada

Release stratum	Number released	Recovery stratum 1	2	3	4	5	6	7	8	9	10	Tags not read
1	1070	118	68	8	7	1	0	0	0	0	0	868
2	1919	32	245	88	45	14	3	2	2	0	0	1488
3	2487	0	114	251	216	89	11	3	4	1	1	1797
4	1103	0	0	6	93	221	65	19	9	1	0	689
5	763	0	0	0	0	52	98	47	20	4	0	542
6	638	0	0	0	0	0	1	30	119	56	8	424
7	628	0	0	0	0	0	0	0	61	171	18	378
8	209	0	0	0	0	0	0	0	0	8	38	163
Number recovered without tags		13,047	54,291	33,389	11,740	13,411	10,037	8448	16,156	21,178	5772	
Number tags recovered but not read		198	914	877	85	87	105	13	59	108	32	

Note: Number of tags not recovered = 3,871.

Table 5
Estimates of parameters from fitting the full model

Release stratum	$\hat{N}_i$	Estimates of $\hat{\mu}_{ij}$ 1	2	3	4	5	6	7	8	9	10	$\hat{\gamma}_i$	$\hat{\beta}_i$
1	36,883.4	273.6	213.8	27.5	8.9	1.2	0.0	0.0	0.0	0.0	0.0	867.4	33.5
2	102,511.6	74.1	772.0	297.7	58.9	17.1	4.4	2.9	2.3	0.0	0.0	1487.0	52.4
3	48,792.9	0.0	357.8	875.7	271.2	109.3	17.3	3.6	5.0	1.4	1.5	1795.8	18.6
4	31,941.5	0.0	0.0	20.8	118.1	271.0	100.2	23.8	11.1	1.5	0.0	688.5	27.9
5	35,316.7	0.0	0.0	0.0	0.0	63.6	145.6	64.4	23.6	5.9	0.0	541.6	45.2
6	56,086.6	0.0	0.0	0.0	0.0	0.0	1.4	53.3	127.6	86.1	11.8	423.7	86.7
7	31,760.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	71.2	254.5	26.7	377.7	49.6
8	12,811.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	12.0	56.3	162.9	60.3
$\hat{\lambda}_j$		0.431	0.318	0.287	0.809	0.812	0.629	0.886	0.785	0.690	0.670		

$\hat{N} = 356{,}104.1$

the lack-of-fit comes from two cells, the deviations between the observed and expected counts are usually below 4%, and large samples sizes make even minor deviations detectable. The estimates of the λ_j indicate that a simpler model would not be tenable (the goodness-of-fit statistic increases to 1008.2 with 11 d.f.). The estimates of the β_i indicate that tagging rates varied from about 1/90 per day to 1/30 per day, and again a simpler model is not tenable (the goodness-of-fit statistic increases to 433.6 with 9 d.f.).

The estimate of the overall run is 356,104 fish with an estimated standard error (SE) of 4749 fish. Note that the estimated SE of the overall run is comparable to the SE of the estimated individual strata sizes—this is caused by the fact that the latter are very highly correlated.

The pooled Petersen (i.e., ignoring stratification) is approximately 3.7% smaller. The apparent bias is statistically significant but not of biological significance. The negative bias is not unexpected. As noted by Seber (1982, Section 3.2.2, p. 86) and Schwarz and Taylor (1998), a positive correlation between the capture and recapture probabilities among fish leads to a negative bias in the Petersen estimate. Ironically, the capture probability of the fish wheel-caught fish is often the highest when large numbers of fish are migrating due to optimal water conditions and what might be density-dependent catchability may be positively correlated with abundance. Similarly, at the fishway, more effort is expended when a large number of fish are expected to arrive, and the capture probability is also larger.

Two simulation studies were performed to (1) assess the performance of the proposed estimator in the current example and (2) to assess the performance of the estimators when recapture rates are much smaller and data is sparser.

One hundred simulated datasets from a population based on the estimates in Tables 5 were generated, and estimates and estimated standard errors were obtained for each data set. All point estimates appear to unbiased, but there may be a slight underestimate of the actual standard error for the strata sizes and the overall population size. The good performance

Table 6
Estimated SE of parameter estimates from fitting the full model

Release stratum	SE($\hat{N}_i$)	SE($\hat{\mu}_{ij}$) 1	2	3	4	5	6	7	8	9	10	SE($\hat{\gamma}_i$)	SE($\hat{\beta}_i$)
1	3095.2	18.7	24.5	9.7	3.3	1.2	0.0	0.0	0.0	0.0	0.0	29.4	3.1
2	5676.8	12.3	38.3	27.0	7.2	4.5	2.6	1.8	1.7	0.0	0.0	38.5	3.2
3	4735.8	0.0	30.3	38.6	16.3	11.3	5.2	2.0	2.5	1.4	1.5	42.2	1.9
4	3346.3	0.0	0.0	8.5	11.0	16.6	11.7	5.1	3.7	1.5	0.0	26.2	3.1
5	3919.9	0.0	0.0	0.0	0.0	8.5	13.2	7.8	5.4	2.9	0.0	23.3	5.4
6	5535.8	0.0	0.0	0.0	0.0	0.0	1.5	5.9	10.4	10.7	4.1	20.6	9.3
7	3694.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	9.1	16.9	6.0	19.4	6.2
8	2188.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.1	8.2	12.8	11.3
SE($\hat{\lambda}_j$)		0.027	0.013	0.013	0.018	0.018	0.030	0.026	0.027	0.024	0.048		

SE($\hat{N}$) = 4,749.4

of the estimators is not unexpected given the relatively large number of tagged fish recovered.

A second set of simulated datasets with approximately 1/10 of the recovery effort observed in the example was generated using a similar procedure. The point estimators remained unbiased, but as expected, the variability of the estimators increased by about a factor of $(10)^{1/2}$. The estimated standard errors are much more variable, and there appears to be significant bias in the estimated standard errors for the stratum and overall population sizes. In sparse data, many of the m_{ij} are zero, and Plante et al. (1998) note that the corresponding $\hat{\mu}_{ij} = 0$, and these sampling zeroes have the same effect as structural zeroes in the variance estimation, i.e., the observed zeroes are treated as known parameters and so the estimated variances are biased downward.

5. Discussion

In this experiment, the use of estimating equations provides a simple method to obtain estimates compared to using a formal likelihood, which is not tractable. As noted in the paper, our estimates are a form of minimum χ^2 estimators and, asymptotically, will be close to fully efficient. The disadvantage of this approach is that model selection (through likelihood ratio tests and Akaike information criterion) is not readily done, although some modification using changes in the goodness-of-fit statistic as a surrogate for changes in the likelihood could be possible. Our approach is also related to that of quasi-likelihood (Wedderburn, 1974).

The estimates could also have been obtained using the EM algorithm, where the missing data refers to the allocation of tags counted but not read to the cells in the m_{ij} array. The evaluation and maximization of the complete-data likelihood then is exactly the same as that of Plante et al. (1998). However, it is still difficult to evaluate the expectation of the complete-data likelihood over the unobservable components subject to the complex observed constraints. However, an approximate solution can be obtained as outlined at the end of Section 3.1.

It is interesting to estimate the effect on the estimates of not reading all the tags. If we replace the observed m_{ij} by $m_{ij}^* = m_{ij} + z_j(\hat{\mu}_{ij}/\hat{\mu}_{\bullet j})$ to simulate what would happen if all tags were read and then analyze this simulated data using SPAS, the final estimates of the population size are unchanged, but the SE of the overall run size decreases by approximately 8%. It is not likely cost effective to increase the sampling effort to get this increased precision in this experiment. Using our software, simulated data could be analyzed to help select the optimal level of sampling at both the capture and recapture stages.

The ability to partially count tagged recoveries has two other immediate applications. First, it is quite common in fisheries to tag fish by injecting them with a small wire, about 2 mm in size, coded with information about the time and location of release and to simultaneously batch mark them with a fin clip. When fish are recovered, it is relatively easy to identify the marked fish from the fin clip but more tedious and time consuming to extract and read the coded-wire tag. For example, recoveries are often obtained from recreational fishers who deposit the head into specially marked barrels at fish cleaning stations, and the heads are dissected at the laboratory. Now only a portion of the fish need to be examined further, but the total count of the recoveries can be incorporated.

Second, field conditions may make it difficult to reliably identify tagged fish. Using methods similar to those outlined in Rajwani and Schwarz (1997), a second survey could be conducted to estimate the number of tags overlooked, i.e., an estimate of the z_j could be obtained. These could be used to improve the estimates based on the initial, faulty, survey, but of course, the variance of the estimates needs to be adjusted in a fashion similar to that in Rajwani and Schwarz (1997).

Last, in this particular experiment, the pooled Petersen estimator had a statistically significant but biologically meaningless negative bias. However, it is not difficult to construct examples where the bias can be as great as 40%. As well, the stratified Petersen provides estimates of the capture efficiency of the fish wheels that are being used to evaluate the performance of an in-season mark–recapture method (Link and Gurak, 1997). Similarly, Schwarz and Taylor (1998) compared estimates from a stratified mark–recapture experiment to those obtained by a hydroacoustic method to ascertain why the estimates of the total run varied by a factor of two between the two methods.

Acknowledgements

This work was supported by a Natural Science and Engineering Research Council of Canada (NSERC) research grant to CJS. The sockeye salmon data were provided by the Nisga'a Tribal Council and LGL Limited.

Résumé

L'estimateur d'effectif de Petersen peut être biaisé quand l'hypothèse d'homogénéité des probabilités de capture ou de recapture est violée. Souvent, cette hétérogénéité est liée à la date ou au lieu de capture ou de recapture; si ceux-ci peuvent être stratifiés, l'estimateur de Petersen stratifié réduit le biais induit par cette hétérogénéité. Dans certaines expériences, on ne peut examiner tous les animaux marqués qui sont retrouvés, et on ne dispose donc de la strate de lâcher et de retour que pour un sous-échantillon. Nous développons des méthodes pour une telle modification et les appliquons à l'estimation du nombre de saumons revenant frayer dans un cours d'eau de Colombie britannique, au Canada.

References

Arnason, A. N., Kirby, C. W., Schwarz, C. J., and Irvine, J. R. (1996). Computer analysis of marking data from stratified populations for estimation of salmonid escapements and the size of other populations. *Canadian Technical Report of Fisheries and Aquatic Sciences* **2106**.

Banneheka, S. G., Routledge, R. D., and Schwarz, C. J. (1997). Stratified two-sample tag–recovery census of closed populations. *Biometrics* **53,** 1212–1224.

Becker, N. G. (1984). Estimating population size from capture–recapture experiments in continuous time. *Australian Journal of Statistics* **26,** 1–7.

Darroch, J. N. (1961). The two-sample capture–recapture census when tagging and sampling are stratified. *Biometrika* **48,** 241–260.

Huggins, R. M. (1989). On the statistical analysis of capture experiments. *Biometrika* **76,** 133–140.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73,** 13–22.

Link, M. R. and Gurak, A. C. (1997). The 1995 Fishwheel Project on the Nass River, B.C. *Canadian Manuscript Report of Fisheries and Aquatic Sciences* **2422,** xi + 99 p.

Plante, N. (1990). Estimation de la taille d'une population animale à l'aide d'un modèle de capture–recapture avec stratification. M.Sc. thesis, Université Laval, Laval, Canada.

Plante, N., Rivest, L.-P., and Tremblay, G. (1998). Stratified capture–recapture estimation of the size of a closed population. *Biometrics* **54,** 47–60.

Rajwani, K. and Schwarz, C. J. (1997). Adjusting for missing tags in salmon escapement surveys. *Canadian Journal of Fisheries and Aquatic Sciences* **54,** 800–808.

Schwarz, C. J. and Taylor, C. G. (1998). The use of the stratified Petersen estimator in fisheries management: Estimating the number of pink salmon (*Oncorhynchus gorbuscha*) spawners in the Fraser River. *Canadian Journal of Fisheries and Aquatic Sciences* **55,** 281–297.

Seber, G. A. F. (1982). *The Estimation of Animal Abundance and Related Parameters,* 2nd edition. London: Griffen.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* **61,** 439–437.

Williams, D. A. (1970). Discussion on "A method of discriminating between models." *Journal of the Royal Statistical Society, Series B* **32,** 350.

Yip, P. (1991). A martingale estimating equation for a capture–recapture experiment in discrete time. *Biometrics* **47,** 1081–1088.

Yip, P. S., Huggins, R. M., and Lin, D. Y. (1996). Inference for capture–recapture experiments in continuous time with variable capture rates. *Biometrika* **83,** 477–483.

Received June 1998. *Revised February* 1999.
Accepted March 1999.