2007

SALMONNb: a program for computing cohort-specific effective population sizes (*Nb*) in Pacific salmon and other semelparous species using the temporal method

Robin Waples
NOAA, robin.waples@noaa.gov

Michele Masuda
Alaska Fisheries Science Center, michele.masuda@noaa.gov

Jerome Pella
Alaska Fisheries Science Center

Molecular Ecology Notes (2007) 7, 21–24 doi: 10.1111/j.1471-8286.2006.01567.x

PROGRAM NOTE

SALMONNb: a program for computing cohort-specific effective population sizes (N_b) in Pacific salmon and other semelparous species using the temporal method

ROBIN S. WAPLES,* MICHELE MASUDA† and JEROME PELLA†
*Northwest Fisheries Science Center, 2725 Montlake Blvd. East, Seattle, WA 98112, USA, †Auke Bay Laboratory, Alaska Fisheries Science Center, 11305 Glacier Highway, Juneau, AK 99801, USA

Abstract

We describe a new method and a computer program SALMONNb to calculate the effective number of breeders (N_b) per year in semelparous species with variable age at maturity. The existing temporal method for the 'salmon' life history produces an estimate of the harmonic mean N_b in the two sampled years. SALMONNb reads genotypic data in standard formats and computes yearly N_b values by combining information from pairwise comparisons of samples taken in 3 or more years. Simulations show that the new method produces unbiased estimates of yearly N_b, and precision is inversely proportional to true effective size.

Keywords: age-structured, genetics, least-squares, N_e/N ratio, semelparous

Received 24 May 2006; revision received 20 July 2006; accepted 21 August 2006

Spurred by growing interest among evolutionary biologists and conservation biologists in studying natural populations that are (or might be) quite small, genetic methods for estimating effective population size (N_e) have been energetically applied in recent years (reviewed by Wang 2005 and Leberg 2005). The most widely used genetic approach for estimating contemporary N_e is known as the temporal method (Krimbas & Tsakas 1971; Nei & Tajima 1981) because it involves comparisons of population allele frequencies at two or more points in time. Temporal estimates of N_e can be tricky to interpret, especially if one wants to identify the specific time period(s) to which the estimates apply (Waples 2005). This would be important, for example, if one wanted to estimate the ratio of effective size to census size (N_e/N), or if one wanted to pinpoint the time periods during which particularly rapid genetic changes would be expected to occur.

In the standard (discrete generation) temporal method, $\hat{N}_e$ represents the harmonic mean N_e across the period of sampling. In many species, however, the quantity that can be estimated directly is the effective number of breeders per year (N_b) rather than the effective size per generation (N_e). Recently, Waples (2005) showed that when the temporal method is applied to species with a life history like Pacific salmon, and each sample includes individuals from only a single cohort, the result [$\hat{N}_{b(i,j)}$] can be interpreted as an estimate of the harmonic mean N_b in the parental years (years i and j) for the two sampled cohorts. In addition to Pacific salmon, this result could apply to other semelparous species with variable age at maturity, such as many monocarpic plants and crustaceans with diapausing eggs (Waples 2006).

Because a single temporal estimate in the salmon model integrates information from N_b in two different years, there is a resulting ambiguity in relating the estimate to effective size in any given year. Suppose, however, that three or more temporal samples are available from the same population. In this case, more than one pairwise estimate will yield information about effective size in any given year, and this provides a basis for estimating N_b in individual brood years from the combined sampling information. Here, we describe an algorithm to estimate N_b in specific years for species with salmon life history, given genetic data from three or more points in time. We also describe a computer program, SALMONNb, which implements this algorithm.

Correspondence: Robin Waples, Fax: (206) 860–3335;
E-mail: robin.waples@noaa.gov

Consider a series of genetic samples taken in K different years. Samples from years i and j produce an estimate $\hat{N}_{b(i,j)}$, which relates to the harmonic mean effective size in years i [$N_{b(i)}$] and j [$N_{b(j)}$]:

$$\hat{N}_{b(i,j)} \approx \frac{2}{1/N_{b(i)} + 1/N_{b(j)}} \Rightarrow \frac{2}{\hat{N}_{b(i,j)}} \approx \frac{1}{N_{b(i)}} + \frac{1}{N_{b(j)}}. \quad (1)$$

Our objective here is to decompose the pairwise estimates $\hat{N}_{b(i,j)}$ into separate estimates of $N_{b(i)}$ and $N_{b(j)}$. The K samples allow a total of $K(K-1)/2$ pairwise comparisons, and for any given year $K-1$ of these pairwise estimates will be informative about effective size. For sample $i = 1$ we have

$$\hat{N}_{b(1,2)} + \hat{N}_{b(1,3)} + \cdots + \hat{N}_{b(1,K)} \approx \frac{2}{1/N_{b(1)} + 1/N_{b(2)}} + \frac{2}{1/N_{b(1)} + 1/N_{b(3)}} + \cdots + \frac{2}{1/N_{b(1)} + 1/N_{b(K)}},$$

with comparable equations for every other year.

If we replace the $N_{b(i)}$ with the estimates $\hat{N}_{b(i)}$, Equation 1 can be rewritten as $\Delta_{i,j} = 2/\hat{N}_{b(i,j)} - 1/\hat{N}_{b(i)} - 1/\hat{N}_{b(j)}$. $\Delta_{i,j}$ thus quantifies the difference between the pairwise estimate $\hat{N}_{b(i,j)}$ and the harmonic mean of the individual estimates $\hat{N}_{b(i)}$ and $\hat{N}_{b(j)}$. Now let Φ represent the sum of the squared $\Delta_{i,j}$: $\Phi = \sum_{j>i} \Delta_{i,j}^2$. To obtain a least-squares solution for the $\hat{N}_{b(i)}$, we first take partial derivatives of Φ and set them to zero:

$$\frac{\partial \Phi}{\partial \hat{N}_{b(1)}} = 2\Delta_{1,2}\left(\frac{1}{\hat{N}_{b(1)}}\right)^2 + 2\Delta_{1,3}\left(\frac{1}{\hat{N}_{b(1)}}\right)^2 + \cdots + 2\Delta_{1,K}\left(\frac{1}{\hat{N}_{b(1)}}\right)^2 = 0$$

$$\vdots$$

$$\frac{\partial \Phi}{\partial \hat{N}_{b(K)}} = 2\Delta_{1,K}\left(\frac{1}{\hat{N}_{b(K)}}\right)^2 + 2\Delta_{2,K}\left(\frac{1}{\hat{N}_{b(K)}}\right)^2 + \cdots + 2\Delta_{K-1,K}\left(\frac{1}{\hat{N}_{b(K)}}\right)^2 = 0.$$

Dividing the ith equation by $2/\hat{N}_{b(i)}^2$ eads to equations of the form $\Delta_{1,2} + \Delta_{1,3} + \cdots + \Delta_{1,K} = 0$. Finally, expanding the $\Delta_{i,j}$ and gathering terms leads to a system of K linear equations:

$$2\left(\frac{1}{\hat{N}_{b(1,2)}} + \frac{1}{\hat{N}_{b(1,3)}} \cdots + \frac{1}{\hat{N}_{b(1,K)}}\right) - \frac{K-1}{\hat{N}_{b(1)}} - \frac{1}{\hat{N}_{b(2)}} - \frac{1}{\hat{N}_{b(3)}} \cdots - \frac{1}{\hat{N}_{b(K)}} = 0$$

$$\vdots \quad (2)$$

$$2\left(\frac{1}{\hat{N}_{b(1,K)}} + \frac{1}{\hat{N}_{b(2,K)}} \cdots + \frac{1}{\hat{N}_{b(K-1,K)}}\right) - \frac{1}{\hat{N}_{b(1)}} - \frac{1}{\hat{N}_{b(2)}} - \cdots - \frac{1}{\hat{N}_{b(K-1)}} - \frac{K-1}{\hat{N}_{b(K)}} = 0$$

Since the $\hat{N}_{b(i,j)}$ are obtained from the data, these K equations can be solved explicitly for the K unknowns $\hat{N}_{b(i)}$ — the estimates of effective size in the individual brood years that we desire.

The various pairwise estimates $\hat{N}_{b(i,j)}$ typically will not have the same information content. Sample size (S) often differs among years, and the number of independent alleles upon which a particular estimate is based (n) can also vary over time. To account for these differences, we weight each pairwise estimate by the reciprocal of its variance. An approximate expression for the variance of $\hat{N}_e$ was given by Pollak (1983); it can be modified for the salmon model as follows:

$$V_{\hat{N}_b} \approx \frac{8N_b^4}{n}[1/(4N_b^2) + 1/(bN_b\tilde{S}) + 1/(b^2\tilde{S}^2)], \quad (3)$$

where $\tilde{S}$ is the harmonic mean S in the two years being compared, and b is an analogue for elapsed time in generations in the Pacific salmon model that depends on age structure and number of years between samples (Waples 1990; Tajima 1992). A difficulty arises in that the variance is a function of true N_b, which is unknown. Since the goal is to compute relative weights for individual pairwise estimates $\hat{N}_{b(i,j)}$, it seems best to use a single, global estimate of N_b ($\tilde{N}_b$) for all pairwise comparisons, which can be computed as the harmonic mean of all the pairwise $\hat{N}_{b(i,j)}$ values.

The algorithm described above requires estimates $\hat{N}_{b(i,j)}$ for each pair of samples. SALMONNb provides two options: (i) Read the pairwise estimates from a file; or (ii) Calculate them from raw genetic data. Corresponding input files for the two options are:

1 The first line is the number of years of samples. Subsequent lines, one for each pair of samples, are formatted in six columns. 1–2: sample labels or years compared in chronological order; 3: n; 4: $\tilde{S}$; 5: b; and 6: $\hat{N}_{b(i,j)}$.
2 Genotypic data in the format for FSTAT (Goudet 2001) or GENEPOP (Raymond & Rousset 1995) and maturity at age data are provided in two separate files. The file of age data includes sample labels indicating year of collection, followed by a column of ages and a column giving age-specific probabilities of maturity. Pollak's (1983) method is used to calculate $\hat{F}$, Tajima's (1992) algorithm is used to calculate b from the age at maturity data, and the $\hat{N}_{b(i,j)}$ are calculated according to Waples (1990).

Table 1 shows output of SALMONNb using Option 1. This dataset has samples from four years. The pairwise values n, $\tilde{S}$, b and $\hat{N}_{b(i,j)}$ are supplied by the user; SALMONNb calculates $\tilde{N}_b$, Var [$\hat{N}_{b(i,j)}$] and the $\hat{N}_{b(i)}$ as described above. Note that $\hat{N}_{b(3)} = 89.1$ is lower than any of the pairwise $\hat{N}_{b(i,j)}$ involving year 3 and $\hat{N}_{b(4)} = 626.6$ is higher than any of the pairwise $\hat{N}_{b(i,j)}$ involving year 4. This demonstrates that simply taking the mean (or even harmonic mean) of the $\hat{N}_{b(i,j)}$ estimates involving a particular year can give a

Table 1 Summary of output from program SALMONNb using Option 1 and data for four consecutive years of samples. For each pairwise comparison of samples i and j, $\tilde{S}$ is the harmonic mean sample size, n is the number of independent alleles used in the comparison, $\hat{N}_{b(i,j)}$ are the pairwise estimates of N_b (all supplied by the user in Option 1), and Var[$\hat{N}_{b(i,j)}$] is the variance of $\hat{N}_{b(i,j)}$ computed from Equation 3. In this example, the coefficients b were 2.32, 2.71 and 2.09 for comparisons 1, 2 and 3 years apart, respectively. $\tilde{N}_b$ is the harmonic mean of the $\hat{N}_{b(i,j)}$ and the $\hat{N}_{b(i)}$ are the estimates of N_b in the individual years

Year	1	2	3	4
Pairwise $\tilde{S}$ (above diagonal) and n (below diagonal):				
1		76.3	80.0	65.9
2	21		78.8	62.4
3	17	18		67.9
4	15	16	13	
Pairwise $\hat{N}_{b(i,j)}$ (above diagonal) and Var[$\hat{N}_{b(i,j)}$] (below diagonal):				
1		181.9	111.8	345.2
2	21852		124.7	221.7
3	20914	24449		161.3
4	43045	30459	41230	
$\tilde{N}_b$	= 166.4			
Yearly estimates $\hat{N}_{b(i)}$	178.5	175.2	89.1	566.5

misleading impression of effective size in that year. In years with unusually low N_b, the pairwise $\hat{N}_{b(i,j)}$ will consistently overestimate N_b for that year because it is being compared with years with larger effective size; the converse is true for years with unusually large N_b. The algorithm described here jointly considers the information for all pairs of samples to arrive at an estimate of $\hat{N}_{b(i)}$ for each year that minimizes squared deviations from the true value.

Option 2 allows the user to exclude alleles below a critical frequency to reduce potential bias in $\hat{N}_b$. As a default, SALMONNb reports separate estimates after excluding alleles with frequencies less than 0.05, 0.02 and 0.01; users can pick a different critical value as an option. In Option 2, if the age file indicates 100% maturity at age 1, the program will return a temporal estimate of N_e for each pair of samples using the standard, discrete generation temporal model (Waples 1989).

To evaluate performance of SALMONNB, we considered three different scenarios of true $N_{b(i)}$ values in four consecutive years: $N_{b(1–4)}$ = (100, 100, 100, 100); (50, 100, 50, 100); (50, 100, 200, 400). Four annual samples produce six pairwise estimates $\hat{N}_{b(i,j)}$. We used Monte-Carlo methods to generate random variation in the $\hat{N}_{b(i,j)}$ around the true values, with the magnitude of variation comparable to that expected to arise from sampling a finite number of individuals (S = 25 or 100) and independent alleles (n = 25 or 100). This variation was generated independently for each of the six $\hat{N}_{b(i,j)}$ values within a replicate, and the process was repeated to generate 1000 replicate datasets that were analysed using Option 1 in SALMONNb.

Results (Table 2) show that, provided the pairwise estimates $\hat{N}_{b(i,j)}$ are unbiased, SALMONNb produces unbiased estimates of effective size in individual years: the harmonic mean $\hat{N}_{b(i)}$ was essentially identical to the true $N_{b(i)}$ in every case. Importantly, the lack of bias was consistent regardless of the magnitude of $N_{b(i)}$, the sequence of the $N_{b(i)}$ values, or the values of n or S. Precision was evaluated by comparing empirical 95% confidence intervals (CIs) for $\hat{N}_{b(i)}$ with the parametric CIs that would apply to a temporal estimate of effective size based on a pair of samples of S individuals and n degrees of freedom (see Equation 16 in Waples 1989). Results in Table 2 suggest that the relative precision of $N_{b(i)}$ compared to the theoretical expectation depends primarily on the true effective size. For years with $N_{b(i)}$ = 50 (as might occur for many populations of conservation interest), there was little difference between the empirical and theoretical CIs. When true $N_{b(i)}$ was 100 or larger, the empirical CIs were wider than the theoretical CIs, particularly with respect to the upper bound. For true $N_{b(i)}$ = 100, therefore the point estimate $\hat{N}_{b(i)}$ from SALMONNb is unbiased but precision is somewhat lower than would be obtained from analysis of two temporal samples from a population of constant size. Presumably this reduction in precision reflects the difficulty in estimating $N_{b(i)}$ indirectly from pairwise estimates $\hat{N}_{b(i,j)}$, a difficulty that can be largely overcome if effective size is small enough (and hence the drift signal strong enough).

SALMONNb is a FORTRAN 90 program written for a personal computer. The FORTRAN code was compiled with the LAHEY FORTRAN 95 COMPILER, release 5.00f (Lahey 1998). SALMONNb uses FORTRAN routines from Numerical Recipes (Press *et al*. 2002). The SALMONNb program, User's Manual, and example data sets can be downloaded from the anonymous ftp site, ftp://ftp.afsc.noaa.gov/Sida/SalmonNb/.

Table 2 Results of the analysis of five simulated datasets, which differed in the true effective size in each of four consecutive years [$N_{b(i)}$], sample size (S) and number of independent alleles used (n). For each dataset, 1000 replicate series of $\hat{N}_{b(i,j)}$ values were generated, randomly distributed around the true $N_{b(i,j)}$, and SALMONNb was used to estimate $N_{b(i)}$ in each year. $\tilde{\hat{N}}_{b(i)}$ is the harmonic mean of the $\hat{N}_{b(i)}$ estimates over all replicates. Empirical 95% confidence intervals (CIs) are based on the distribution of the $\hat{N}_{b(i)}$ across replicates; theoretical CIs were calculated according to Waples (1989); '*inf*' indicates the upper bound of a CI that extends to infinity

			True		Empirical CI		Theoretical CI	
S	n	Year (i)	$N_{b(i)}$	$\tilde{\hat{N}}_{b(i)}$	Lower	Upper	Lower	Upper
100	100	1	100	100.0	58	266	62	170
		2	100	100.0	60	276	62	170
		3	100	100.0	62	248	62	170
		4	100	100.0	60	292	62	170
100	100	1	50	50.0	34	86	34	73
		2	100	100.0	55	428	62	170
		3	50	50.0	35	85	34	73
		4	100	100.1	53	453	62	170
100	100	1	50	50.0	36	75	34	73
		2	100	100.0	60	229	62	170
		3	200	199.8	90	*inf*	105	492
		4	400	399.9	115	*inf*	163	9609
25	100	1	50	50.0	28	170	26	123
		2	100	100.0	41	*inf*	41	2402
		3	200	199.8	52	*inf*	56	*inf*
		4	400	399.9	59	*inf*	69	*inf*
100	25	1	50	50.0	26	145	22	108
		2	100	100.0	40	*inf*	38	326
		3	200	199.8	54	*inf*	60	*inf*
		4	400	399.9	67	*inf*	83	*inf*

Acknowledgements

Vincent Castric provided useful comments on a draft of the manuscript.

References

Goudet J (2001) FSTAT: *A program to estimate and test gene diversities and fixation indices* (version 2.9.3). Available from http://www2.unil.ch/popgen/softwares/fstat.htm.

Krimbas CB, Tsakas S (1971) The genetics of *Dacus oleae*. V. Changes of esterase polymorphism in a natural population following insecticide control — selection or drift? *Evolution*, **25**, 454–460.

Lahey Computer Systems Inc. (1998) *LAHEY FORTRAN 95 language reference*. Lahey Computer Systems, Inc, Nevada.

Leberg P (2005) Genetic approaches for estimating the effective size of populations. *Journal of Wildlife Management*, **69**, 1385–1399.

Nei M, Tajima F (1981) Genetic drift and estimation of effective population size. *Genetics*, **98**, 625–640.

Pollak E (1983) A new method for estimating the effective population size from allele frequency changes. *Genetics*, **104**, 531–548.

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (2002) *Numerical Recipes in FORTRAN 90: the art of parallel scientific computing*. Cambridge University Press, New York.

Raymond M, Rousset F (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity*, **86**, 248–249.

Tajima F (1992) Statistical method for estimating the effective population size in Pacific salmon. *Journal of Heredity*, **83**, 309–311.

Wang J (2005) Estimation of effective population sizes from data on genetic markers. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **360**, 1395–1409.

Waples RS (1989) A generalized approach for estimating effective population size from temporal changes in allele frequency. *Genetics*, **121**, 379–391.

Waples RS (1990) Conservation genetics of Pacific salmon. III. Estimating effective population size. *Journal of Heredity*, **81**, 277–289.

Waples RS (2005) Genetic estimates of contemporary effective population size: to what time periods do the estimates apply? *Molecular Ecology*, **14**, 3335–3352.

Waples RS (2006) Seed banks, salmon, and sleeping genes: effective population size in semelparous, age-structured species with fluctuating abundance. *American Naturalist*, **167**, 118–135.