Latest Formulation of SSM
a portion of a manuscript under review
K. Newman, Oct. 16, 1998

A state-space model is a model for two parallel time series, one unobservable series, the state process, $\mathbf{n}_t$, and one observable series, the observation process, $\mathbf{c}_t$ (West and Harrison 1997). It is usually assumed that given the past state $\mathbf{n}_{t-1}$, $\mathbf{n}_t$ is independent of all other previous states (the state process is first order Markovian) and given $\mathbf{n}_t$, $\mathbf{c}_t$ is independent of all other states.

The observation process is a length L vector of estimated recoveries of tagged fish by fisheries in each of L non-overlapping areas during the weeks t=1,2,...,T. The state process is a corresponding vector of unobservable abundances of the cohort in each of the L areas during the weeks t=0,1,...,T. The time period 0 represents the beginning of the harvest season (generally late June to early July) and T=16 (early to mid-October). The spatial framework, i.e., the L areas, is a line along the west coast of North America from southern Oregon to northern British Columbia that is partitioned into L non-overlapping segments (L=12 in the application). The particular form of the SSM is

$$\mathbf{n}_t|\mathbf{n}_{t-1} = M_t S_t \mathbf{n}_{t-1} + \mathbf{w}_t \tag{1}$$
$$\mathbf{c}_t|\mathbf{n}_t = H_t \mathbf{n}_t + \mathbf{v}_t \tag{2}$$

S_t, M_t, and H_t are L by L survival, movement, and harvest matrices. Both S_t and H_t are diagonal matrices— survival and harvest in one area has no effect on survival and harvest in another area during the same time period. The state process error component, $\mathbf{w}_t$, is a L by 1 vector following a multivariate normal distribution with mean zero and covariance matrix $\mathbf{\Sigma}_{w,t}$. Similarly the observation process error component, $\mathbf{v}_t$, is L by 1 and multivariate normal with mean zero and covariance matrix $\mathbf{\Sigma}_{v,t}$. The covariance matrices $\mathbf{\Sigma}_{w,t}$ and $\mathbf{\Sigma}_{v,t}$ are functions of the proportions in S_t, M_t, H_t as well as the abundance and are based on approximations to the variances and covariances for sums of multinomial random variables.

The general structure of the SSM can be concisely partitioned into three modules, initial distribution, survival/mortality, and movement. It is this flexibility that has led to considerable experimentation and generated suggestions by biologists to evaluate different theories.

Initial distribution

To begin the process, the initial state vector, $\mathbf{n}_0$, is found by first calculating the expected number of fish alive. A known number, R, of tagged salmon leave the freshwater natal area (e.g., a hatchery) and, assuming independence between the fish, the expected number surviving to the beginning of the harvest season is $R\theta_{i,s}$, where $\theta_{i,s}$ is the unknown survival rate. The survivors are assumed to distribute themselves along the line segment according to a (scaled) beta($\theta_{i,\alpha}$, $\theta_{i,\beta}$). The expected numbers per area are used to construct $\mathbf{n}_0$.

Survival and mortality

For the survival matrix S_t (and relatedly the harvest matrix H_t) there are three parameters, $\theta_{s,q_{U.S.}}$, $\theta_{s,q_{Canada}}$, and $\theta_{s,n}$. The first two are catchability coefficients found in the Baranov catch equations (Ricker 1975). The diagonal elements of S_t are the probabilities of surviving fishing related and natural forces of mortality. In particular, for probabilities of survival from time t to $t+1$ for an area a in U.S. waters and an area b in Canadian waters are

$$\begin{aligned} S_{t+1}[a,a](U.S.) &= \exp\left[-N - F_{a,t}(U.S.)\right] \\ S_{t+1}[b,b](Canada) &= \exp\left[-N - F_{b,t}(Canada)\right] \end{aligned}$$

where

$$\begin{aligned} F_{a,t}(U.S.) &= \theta_{s,q_{U.S.}} \frac{\text{Effort}_{a,t}}{c_1(a_R - a_L)} \\ F_{b,t}(Canada) &= \theta_{s,q_{Canada}} \frac{\text{Effort}_{b,t}}{c_1(b_R - b_L)} \\ N &= c_2 \theta_{s,n} \end{aligned}$$

The terms c_1 and c_2 are scaling constants that keep the parameter estimates of a roughly similar magnitude as other estimates; in the application c_1=50,000 and c_2=0.001. The terms $a_R - a_L$ and $b_R - b_L$ are the lengths of fishing areas a and b, thus scaling the effort measure to effort per unit length; e.g. 500 units of effort in a short segment should have greater impact that 500 units spread across a longer segment. The parameter $\theta_{s,n}$ is the natural mortality, assumed constant between time periods. For the first time period, $t = 1$, 100% survival is assumed (S_1=I, an identity matrix).

The harvest in the final time period is treated differently. There is no or minimal ocean fishing taking place at this time, thus no information about abundance. To artificially force the fish into the terminal area by or before this final period, the final period harvest rate is fixed at 100% for all areas. Since the negative log likelihood function includes a term of squared differences between observed and estimated catches, this complete harvest rate penalizes model parameters that result in very many fish left at sea.

Movement

The jth column of the movement matrix at time t, M_t, is a vector of probabilities (summing to 1.0) for movement from the j area to any of the L areas. The probability for moving from area i to area j is based on an individual fish movement model, which assumes that the probability distribution for the location at time $t+1$ of a fish at location p_t at time t is a beta($\alpha(p_t, t)$, $\beta(t)$). The α parameter is a function of time and location while the β parameter depends on time alone. The formulation is made in terms of the expected value and the parameter β:

$$\mu = (p_t + 0.01)\left[\frac{\exp\left(c_3 - \theta_{m,\mu}\, t/c_4\right)}{1.0 + \exp\left(c_3 - \theta_{m,\mu}\, t/c_4\right)}\right] \tag{3}$$

$$\beta = c_5\theta_{m,\beta}\frac{\exp(c_3 - t^2/T)}{1+\exp(c_3 - t^2/T)} \tag{4}$$

The constant 0.01 has the effect of allowing μ to move away from the natal area early in the fishing season. The constants c_4 and c_5 are for scaling parameter estimates. The parameter α is then found by:

$$\alpha = \mu\beta/(1-\mu)$$

The expected value of the next location, μ, is a product of the current location and a multiplier ≤ 1.0 (a logit function), that early in the time period (t near 0) is made arbitrarily near to 1.0 by choosing c_3 appropriately; in the application c_3=4.6, with $t = 0$, the multiplier is 0.99. As time increases the multiplier shrinks towards zero, which is the location on the line of the natal area. Movement beyond the natal area, whether north or south of the natal area, is not allowed, assuming the freshwater influence will attract the fish back to spawn. Nor is movement outside the line segment (to the far north or south) allowed— thus the system is closed.

The variance of the next location, σ^2, equals $\mu(1-\mu)/(\beta+1-\mu)$. It tends to increase in time, but is bounded above by $\mu(1-\mu)$ and can therefore decrease as μ changes with time.

Given the beta distribution for the individual fish at p_t, the probability of movement from an area a to an area b is found by double integration. The outer integral is over the area a, assuming a uniform probability distribution for fish in the area. Conditional on a given location in a, the inner integral evaluates the beta distribution over the the area b.

Parameter estimation

The total number of unknown parameters is eight; i.e., $\Theta = (\theta_{i,s}$,$\theta_{i,\alpha}$, $\theta_{i,\beta}$, $\theta_{s,q_{U.S.}}$, $\theta_{s,q_{Canada}}$, $\theta_{s,n}$, $\theta_{m,\mu}$, $\theta_{m,\beta})$, where three are related to the initial abundance and spatial distribution, three to survival, and two to movement. The Kalman filter is used to evaluate the likelihood with respect to Θ and maximum likelihood estimates can be calculated. Conditional on maximum likelihood estimates of Θ and the observation process, smoothed estimates of abundance, $\mathrm{E}(\mathbf{n}_t|\hat{\Theta}, \mathbf{c}_1, \ldots, \mathbf{c}_T)$, may be calculated using the recursive Kalman smoothing algorithm (Shumway (1988)).

In the application six of the eight parameters are estimated and two are fixed, namely $\theta_{i,\beta}$=2.0 and $\theta_{i,n}$=0.0. Setting $\theta_{i,\beta}$ to 2.0 still allows considerable flexibility in the initial distribution. The natural mortality parameter, $\theta_{i,n}$, is at times highly correlated with the initial survival parameter, $\theta_{i,s}$ and thus difficult to estimate. With T=16 weeks, the natural mortality is assumed to be relatively slight for these maturing adult fish. The survival matrix in this case is simply related to the previous period harvest matrix in this case: S_{t+1}=$I - H_t$.

Table 1: MLEs and Empirical Bayes (EB) estimates of SSM parameters for six cohorts of coho salmon. The estimated standard errors for the mles and the standard deviations of the posterior distributions of $\Theta|\hat{\Omega}, C_k$ are in the smaller type.

	$\theta_{i,s}$		$\theta_{i,\alpha}$		$\theta_{s,q_{US}}$		$\theta_{s,q_{Canada}}$		$\theta_{m,\mu}$		θ_{m,σ^2}	
Year	MLE	EB	MLE	EB	MLE	EB	MLE	EB	MLE	EB	MLE	EB
1986	5.37	5.36	2.26	2.26	7.12	7.06	3.88	3.89	30.83	30.80	2.47	2.57
	0.04	0.04	0.03	0.03	0.29	0.33	0.06	0.07	0.24	0.23	0.27	0.29
1987	1.42	1.42	1.97	1.99	3.51	3.65	4.03	4.02	42.79	42.55	10.01	10.88
	0.03	0.03	0.09	0.08	0.42	0.43	0.23	0.23	0.71	0.88	20.24	8.60
1988	6.53	6.53	2.41	2.38	4.71	4.75	3.14	3.15	27.37	27.40	1.40	1.41
	0.05	0.05	0.04	0.04	0.24	0.26	0.07	0.07	0.25	0.26	0.12	0.12
1989	4.28	4.28	2.12	2.13	12.75	12.74	10.20	10.18	37.35	37.33	0.95	0.98
	0.05	0.05	0.03	0.03	0.46	0.47	0.21	0.20	0.39	0.42	0.11	0.12
1990	2.42	2.42	2.72	2.72	19.87	20.06	3.46	3.47	31.20	30.99	55.97	24.01
	0.04	0.04	0.06	0.06	1.02	1.02	0.09	0.09	0.33	0.45	24.54	14.11
1991	5.74	5.74	2.94	2.92	16.77	16.67	2.96	2.97	31.07	31.07	22.95	21.85
	0.05	0.05	0.05	0.05	0.83	0.91	0.06	0.06	0.21	0.21	4.85	4.16

References

Newman, K.B. (1998.) State-space modeling of animal movement and mortality with application to salmon. *Biometrics* **54**, 274–297.

Ricker, W.E. (1975.) *Computation and interpretation of biological statistics of fish populations.* Bulletion 191 of the Fisheries Research Board of Canada. Department of Fisheries and Oceans.

Shumway, R.H. (1988.) *Applied statistical time series analysis.* Prentice Hall, New Jersey.

West, M., and Harrison, J. (1990.) *Bayesian forecasting and dynamic models.* New York: Springer-Verlag.