Working Paper F.3

A state-space approach for modeling maturity as a function of time and age

Tim Miller, Larry Jacobson, Loretta O'Brien, Chris Legault and Paul Rago
Northeast Fisheries Science Center
National Marine Fisheries Service
166 Water Street, Woods Hole, MA 02543

This information is distributed solely for the purpose of predissemination peer review. It has not been formally disseminated by NOAA. It does not represent any final agency determination or policy.

Introduction

In age-structured stock assessment models, an estimate of the vector of probabilities of maturity at age (for females in particular) is a necessary component of an estimate of spawning stock biomass. When maturity at age changes over time, it is also necessary to account for these changes.

State-space models have proved useful for inferring parameters driving dynamics of many fish populations (e.g., Sullivan 1992; de Valpine and Hilborn 2005). The state-space model is appealing because it directly accounts for changes in population attributes through time that are imperfectly observed from various ongoing data collection activities. More generally, the state-space model is useful for framing questions about any time series of data where the attribute of interest changes through time (Durbin and Koopman 2001). Other aspects of state-space models that are appealing from a more practical prespective, include that predicted attributes at times where observations have large uncertainty due to either lower sampling or poorer information in the collected data can be improved by observations that neighbor the point in time with less uncertainty. Also, predictions can be made within the time series where observations may be missing or backward or forward in time beyond the available observations (i.e., hind- and fore-casting). As such properties are useful for projection of fish populations into the future, we propose the use of a state-space model for estimating maturatity at age over time.

The models

Maturity

We propose the use of generalized linear models (GLMs) to estimate probability of maturity as a function of age, sex and year. We will assume that the true coefficients of the GLM for the population are unbiasedly observed through the GLM coefficients estimated from the survey data, but that these observations have error due to sampling. We also will rely on the asymptotic normality of the estimated coefficients for year y and so that

$$\widehat{\boldsymbol{\beta}}_y \sim \mathrm{N}\left(\boldsymbol{\beta}_y, (\mathbf{X}_y^T \mathbf{W}_y \mathbf{X}_y)^{-1}\right)$$

where $\mathbf{X}_y$ is the $n_y \times p$ matrix of covariates, $\mathbf{W}_y = \mathrm{diag}\{p_{y,i}(1-p_{y,i})\}$ and $p_{y,i}$ is the expected values for the ith datum in year y (McCullagh and Nelder 1989, pgs. 115-119). When separate GLMs are fit by sex and year, the MLEs of the elements of $\widehat{\boldsymbol{\beta}}_y = (\widehat{\beta}_{y,0,m}, \widehat{\beta}_{y,1,m}, \widehat{\beta}_{y,0,f}, \widehat{\beta}_{y,1,f})^T$ are conditionally independent across sex (i.e, $Cov(\widehat{\beta}_{y,1,m}, \widehat{\beta}_{y,1,f}|\boldsymbol{\beta}) = 0$) and the yearly MLEs are also independent ($Cov\left(\widehat{\boldsymbol{\beta}}_y, \widehat{\boldsymbol{\beta}}_{y+1}\right) = 0$).

We will assume a structural time series model for the yearly coefficients of the maturity relationship (Harvey and Shephard 1993). In particular we will consider the local linear trend model

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_y|\boldsymbol{\beta}_y &\sim \mathrm{N}\left(\boldsymbol{\beta}_y, (\mathbf{X}_y^T \mathbf{W}_y \mathbf{X}_y)^{-1}\right) \\ \boldsymbol{\beta}_{y+1}|\boldsymbol{\beta}_y, \boldsymbol{\gamma}_y &\sim \mathrm{N}\left(\boldsymbol{\beta}_y + \boldsymbol{\gamma}_y, \boldsymbol{\Sigma}\right) \\ \boldsymbol{\gamma}_{y+1}|\boldsymbol{\gamma}_y &\sim \mathrm{N}\left(\boldsymbol{\gamma}_y, \boldsymbol{\Phi}\right). \end{aligned} \tag{1}$$

The first equation in eq. 1 reflects that the observed coefficients will be unbiased estimates of the true coefficients each year, the second equation reflects how the true coefficients each year are related to the true coefficients from the previous year and the third equation reflects that trend parameter behaves according to a random walk.

Let $\boldsymbol{\alpha}_y$ be the stacked vector of $\boldsymbol{\beta}_y$ and $\boldsymbol{\gamma}_y$. In the state-space framework, the observation and transition equations are

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_y &= \mathbf{Z}\boldsymbol{\alpha}_y + \boldsymbol{\epsilon}_y \\ \boldsymbol{\alpha}_{y+1} &= \mathbf{T}\boldsymbol{\alpha}_y + \boldsymbol{\delta}_y \end{aligned}$$

where the relational matrices are

$$\mathbf{Z} = \begin{pmatrix} \mathbf{I}_4 & \mathbf{0}_4 \end{pmatrix}$$

and

$$\mathbf{T} = \begin{pmatrix} \mathbf{I}_4 & \mathbf{I}_4 \\ \mathbf{0}_4 & \mathbf{I}_4 \end{pmatrix}$$

where $\mathbf{I}_k$ is a $k \times k$ identity matrix and $\mathbf{0}_k$ is a $k \times k$ matrix of zeros (following Durbin and Koopman 2001, Section 3.2). The observation error disturbance terms are $\boldsymbol{\epsilon}_y \sim \mathrm{N}\left(\mathbf{0}, (\mathbf{X}_y^T \mathbf{W}_y \mathbf{X}_y)^{-1}\right)$ and the process error disturbance terms are

$$\boldsymbol{\delta}_y \sim \mathrm{N}\left(\mathbf{0}, \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{0}_4 \\ \mathbf{0}_4 & \boldsymbol{\Phi} \end{pmatrix}\right)$$

which are assumed to be serially independent. The vector $\boldsymbol{\gamma}_y$ describes the relationships of the maturity coefficients $\boldsymbol{\beta}$ to time, but the trend is allowed to vary through time in the structural time series model. As the diagonal elements of $\boldsymbol{\Phi}$ increase the trend in the corresponding coefficients may fluctuate more from one year to the next and as the diagonal elements of $\boldsymbol{\Sigma}$ increase there can be more drift of the corresponding coefficients. Note that the deterministic linear model relating the coefficients to time is a special case where $\boldsymbol{\Phi} = \boldsymbol{\Sigma} = 0$ so that there is no process error or autocorrelation of observations.

The only other parameter is an initial state vector $\boldsymbol{\alpha}_1$ which is assumed to have the same process error variance as subsequent state vectors (i.e., that of $\boldsymbol{\delta}_y$). The parameters of the state-space model that we assume unknown (and estimate) are the components of the initial state vector and the process error matrices ($\boldsymbol{\Phi}$ and $\boldsymbol{\Sigma}$). Various hypotheses may be tested with likelihood ratio tests where null models have subsets of unknown parameters equal zero. For example, we may test whether the coefficients are trending rather than just randomly changing through time through a likelihood ratio test of the model where we estimate the elements of $\boldsymbol{\gamma}_1$ and $\boldsymbol{\Phi}$ and the reduced model where we assume those elements are zero.

Parameter estimation

Given the observations and variance estimates from the year- and sex-specific GLM fits, the unknown parameters are the elements of the initial state vector $\boldsymbol{\alpha}_1$ and the elements of the process error variance matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Phi}$. We assume that the process errors for each maturity coefficient are independent so that the process error matrices are diagonal. The

likelihood to be maximized with respect to the parameters $\boldsymbol{\theta}$ is

$$L\left(\boldsymbol{\theta}|\widehat{\boldsymbol{\beta}}_1,\ldots,\widehat{\boldsymbol{\beta}}_Y\right) = p(\widehat{\boldsymbol{\beta}}_1|\boldsymbol{\theta})\prod_{y=2}^{Y} p(\widehat{\boldsymbol{\beta}}_y|\widehat{\boldsymbol{\beta}}_{y-1},\boldsymbol{\theta})$$

where the conditional densities $p(\widehat{\boldsymbol{\beta}}_y|\widehat{\boldsymbol{\beta}}_{y-1},\boldsymbol{\theta})$ are provided by recursions of the well-known Kalman filter (e.g., Durbin and Koopman 2001, Chapters 4 and 7).

Given the smooth prediction of the coefficients at time y, $\widetilde{\boldsymbol{\beta}}_y$ and prediction error variance $\widehat{V}\left(\widetilde{\boldsymbol{\beta}}_y\right)$ we form the predicted logit of maturity at age as

$$\widetilde{l}(y,a,s) = \widetilde{\beta}_{0,y,s} + \widetilde{\beta}_{1,y,s}a,$$

proportion mature at age for sex s as

$$\widetilde{p}(y,a,s) = (1+\exp(-\widetilde{l}(y,a,s)))^{-1}$$

and corresponding 95% prediction interval as

$$CI\left(\widetilde{p}(y,a,s)\right) = \left\{1+\exp\left[-\left(\widetilde{l}(y,a,s) \pm z_{0.975}\widehat{SE}\left(\widetilde{l}(y,a,s)\right)\right)\right]\right\}^{-1}$$

where $\widehat{SE}\left(\widetilde{l}(y,a,s)\right) = \sqrt{\widehat{V}\left(\widetilde{l}(y,a,s)\right)}$ and

$$\widehat{V}\left(\widetilde{l}(y,a,s)\right) = \widehat{V}\left(\widetilde{\beta}_{0,y,s}\right) + a^2\widehat{V}\left(\widetilde{\beta}_{1,y,s}\right) + 2a\widehat{Cov}\left(\widetilde{\beta}_{0,y,s},\widetilde{\beta}_{1,y,s}\right).$$

Example: Georges Bank haddock

To demonstrate the structural time series approach, we fit models to maturity data for Georges Bank haddock collected during the spring survey between 1970 and 2004. The models we fit include a time varying linear trend model as the full model (M_f), a model with no yearly variation in the linear trend (M_2), a model further reduced with no linear trend (M_1) and a model further reduced with no process error (M_0) as the null model (Table 1).

From the loglikelihood ratio tests (see Table 1), we observe that the "best" model is M_1 with no linear trend detectable in any of the maturity coefficients. There was no difference in the log-likelihoods between models M_f and M_2, suggesting that the parameters composing $\boldsymbol{\Phi}$ are completely confounded with other parameters. The "best" model M_1 equates to a random walk of the coefficients through time. The estimates of the elements of the initial state vector and the process variance matrix assuming model M_2 are provided in Table 2. The resulting proportions mature at ages 1, 2 and 3 predicted by the Kalman smooth follow the yearly variation in the observed proportions at age well, especially those with better precision (Figures 1 through 6). Furthermore, the process variance parameters are so large that precision of forecasted proportions at age become poor very rapidly. The proportions of females mature at ages 1, 2, and 3 used in the 2005 assessment compare well with the smooth predictions in the corresponding years (Tables 3 and 4 and Figure 7).

The "best" predicted state variables in forecasts under this structured time series model (M_1) will remain at the same prediction as the last year where observations are available (i.e., 2004). This is an attribute of this particular model and hypotheses for other behaviors could be proposed. One might consider a drift back to some overall mean across all or a certain set of years where observations are available a better alternative. Perhaps these alternative hypotheses could be compared objectively using the same framework.

Other applications

Given yearly estimates of any other population attributes such as mean weight at age or length at age and corresponding estimates of observation errors, the structural time series approach applied here for maturity could also be used to detect whether trends exist in these other population attributes.

The structural time series approach can easily incorporate measured environmental covariates to improve predictions of the state through time and similar to GLMs the usefulness of various environmental covariates in prediction of the state can be assessed.

References

de Valpine, P. and Hilborn, R. 2005. State-space likelihoods for nonlinear fisheries time-series. Can. J. Fish. Aquat. Sci. **62**: 1937–1952.

Durbin, J. and Koopman, S. J. 2001. Time series analysis by state space methods. Oxford University Press, New York, 253pp.

Harvey, A. C. and Shephard, N. 1993. Structural time series models. *In* Handbook of Statistics, volume 11, pp. 261–302, Elsevier, Amsterdam.

McCullagh, P. and Nelder, J. A. 1989. Generalized Linear Models. Chapman & Hall, New York.

Sullivan, P. J. 1992. A Kalman filter approach to catch-at-length analysis. Biometrics **48**: 237–257.

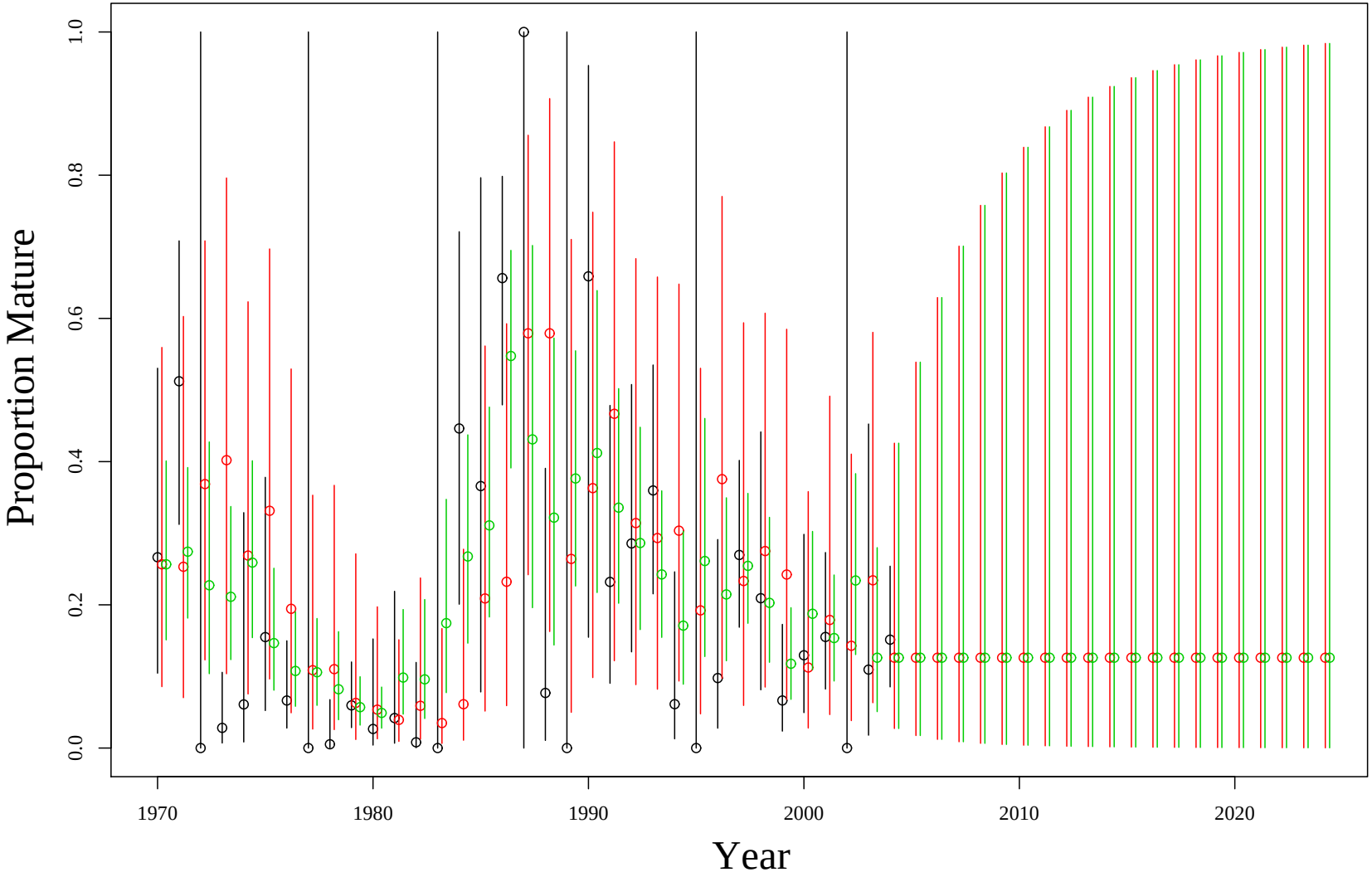

Figure 1. Estimated proportion mature for age 1 males by year. The observed proportions (black) are provided by yearly generalized linear models and Kalman filtered (red) and smoothed (green) proportions are predicted from a random walk model. Vertical lines represent corresponding 95% confidence intervals.

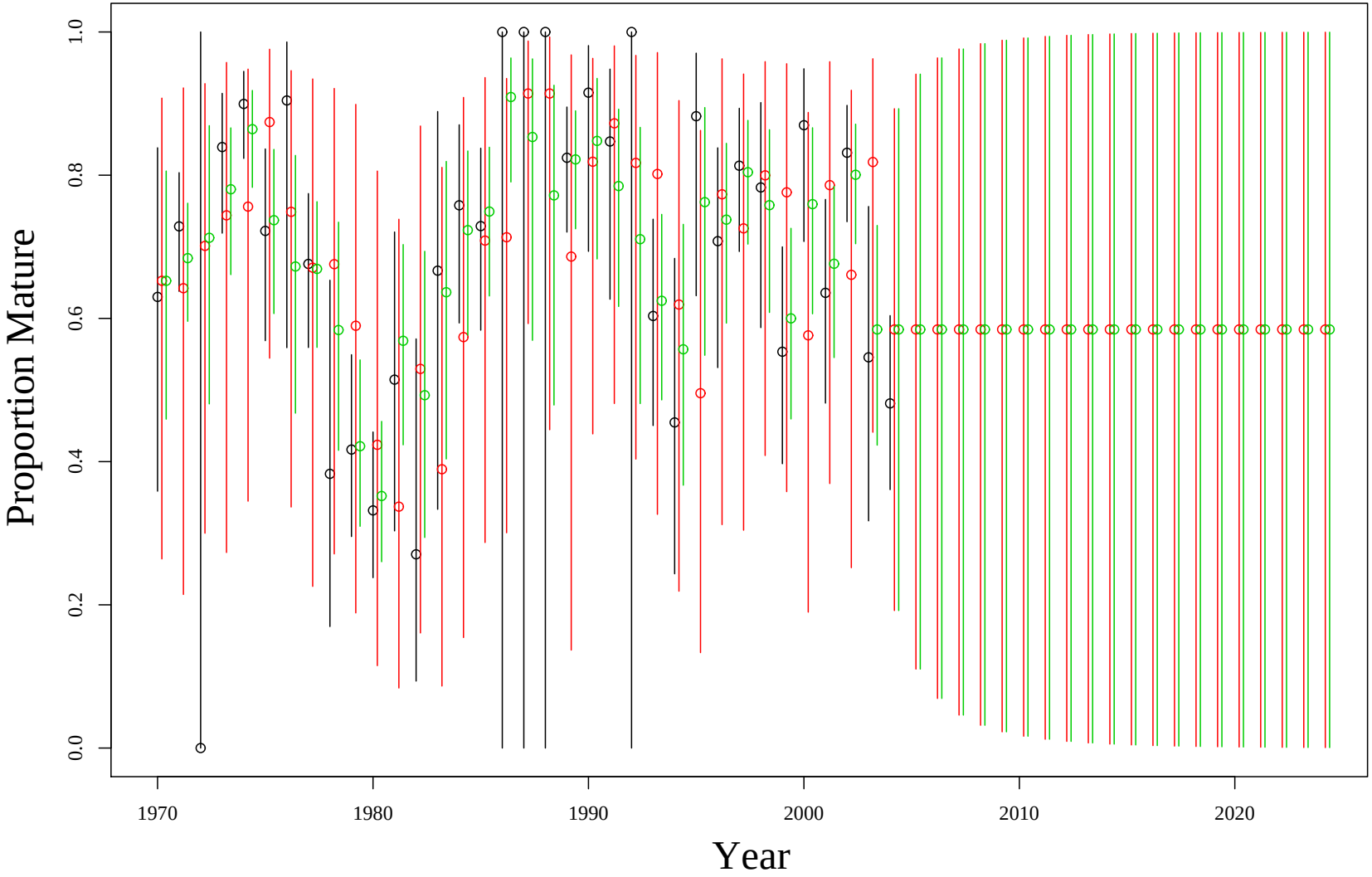

Figure 2. Estimated proportion mature for age 2 males by year. The observed proportions (black) are provided by yearly generalized linear models and Kalman filtered (red) and smoothed (green) proportions are predicted from a random walk model. Vertical lines represent corresponding 95% confidence intervals.

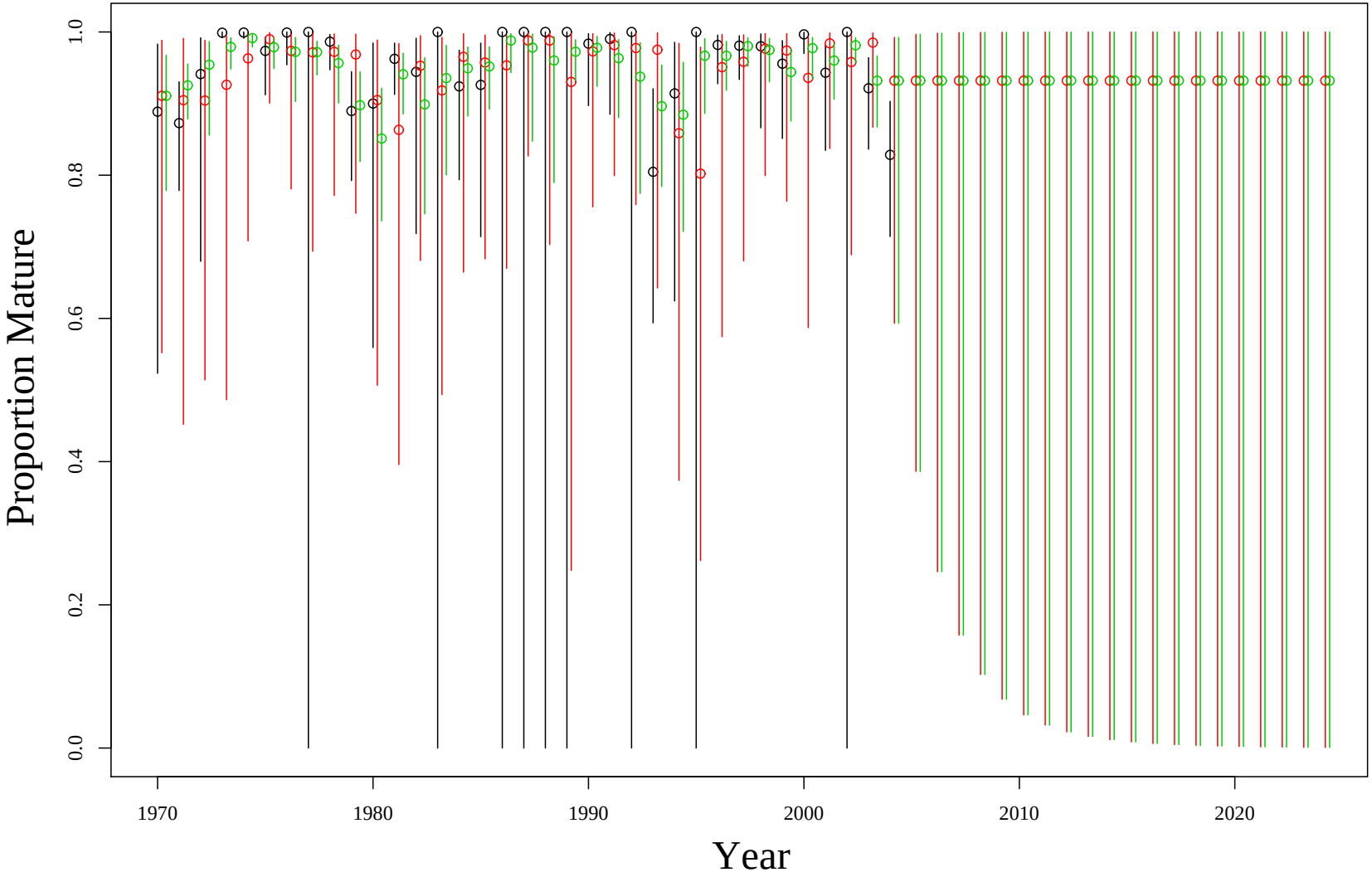

Figure 3. Estimated proportion mature for age 3 males by year. The observed proportions (black) are provided by yearly generalized linear models and Kalman filtered (red) and smoothed (green) proportions are predicted from a random walk model. Vertical lines represent corresponding 95% confidence intervals.

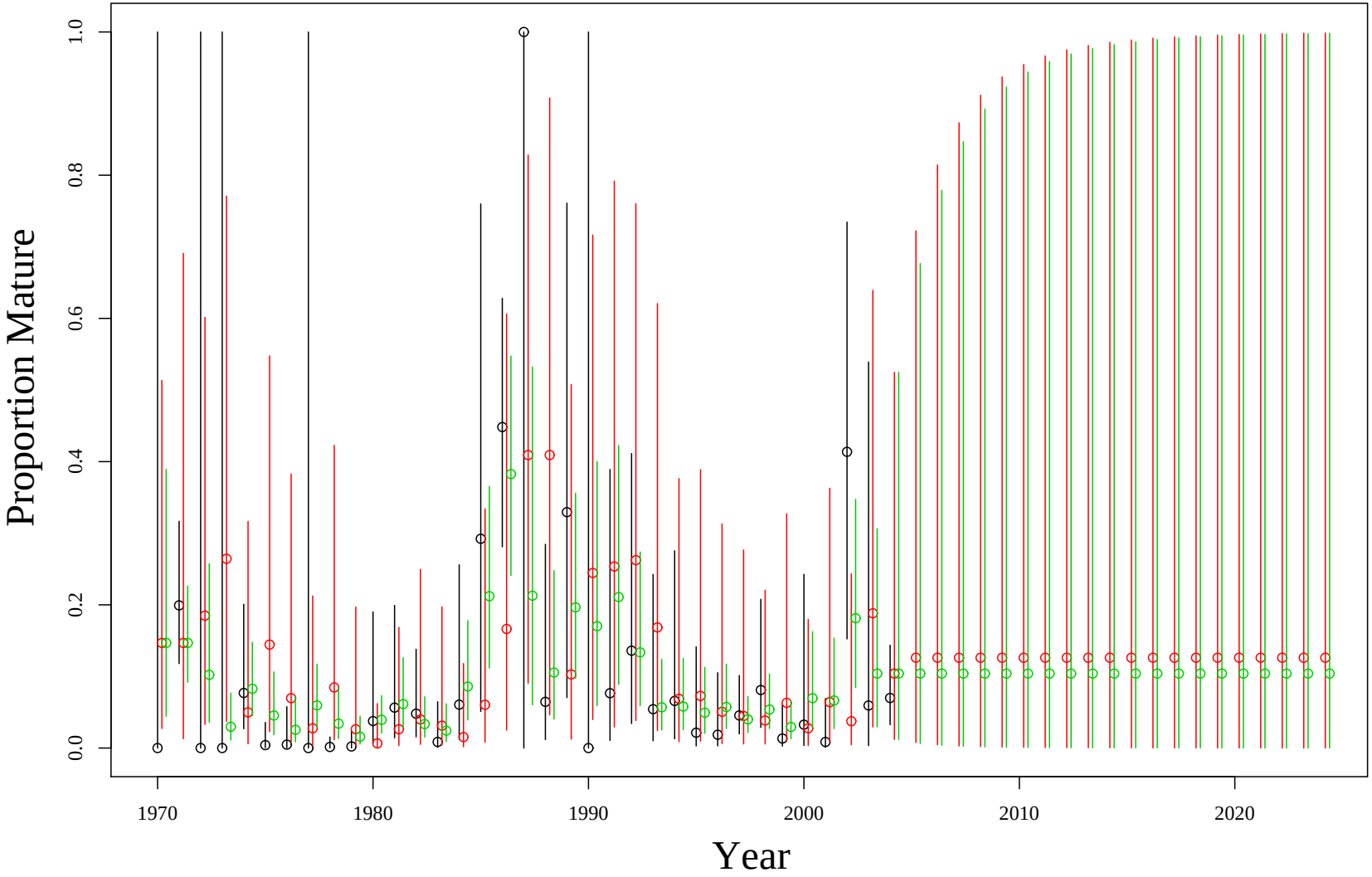

Figure 4. Estimated proportion mature for age 1 females by year. The observed proportions (black) are provided by yearly generalized linear models and Kalman filtered (red) and smoothed (green) proportions are predicted from a random walk model. Vertical lines represent corresponding 95% confidence intervals.

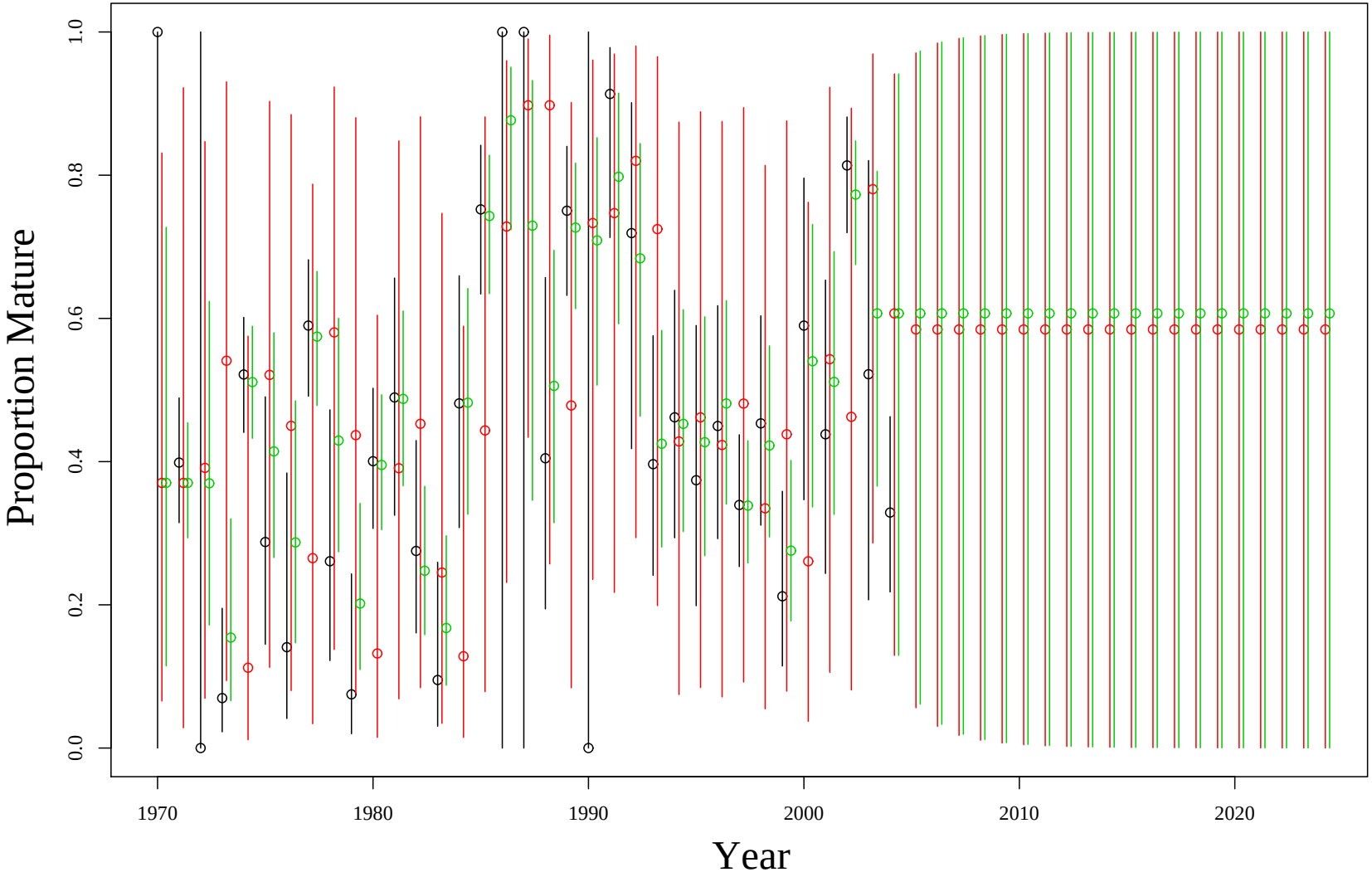

Figure 5. Estimated proportion mature for age 2 females by year. The observed proportions (black) are provided by yearly generalized linear models and Kalman filtered (red) and smoothed (green) proportions are predicted from a random walk model. Vertical lines represent corresponding 95% confidence intervals.

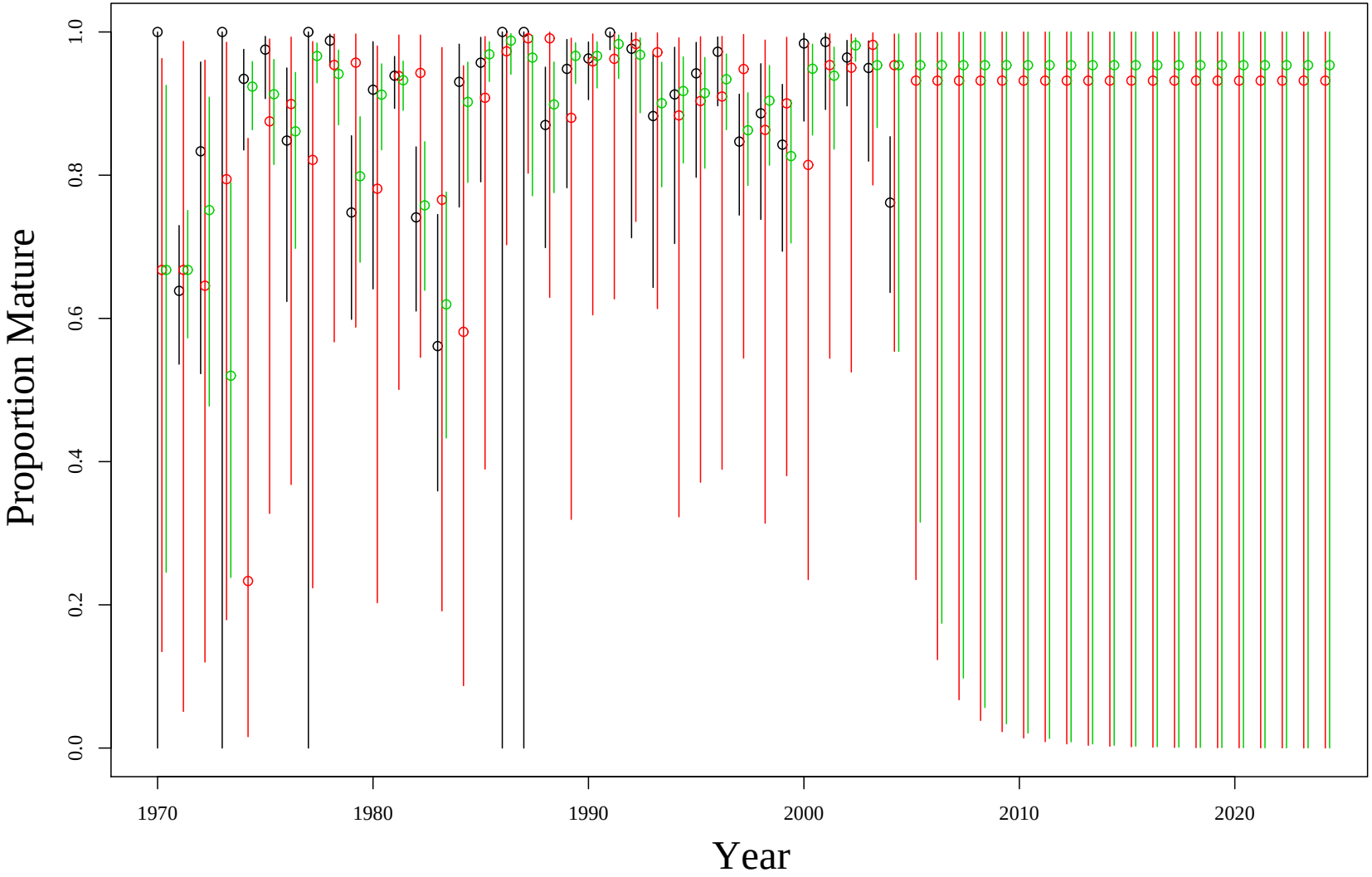

Figure 6. Estimated proportion mature for age 3 females by year. The observed proportions (black) are provided by yearly generalized linear models and Kalman filtered (red) and smoothed (green) proportions are predicted from a random walk model. Vertical lines represent corresponding 95% confidence intervals.

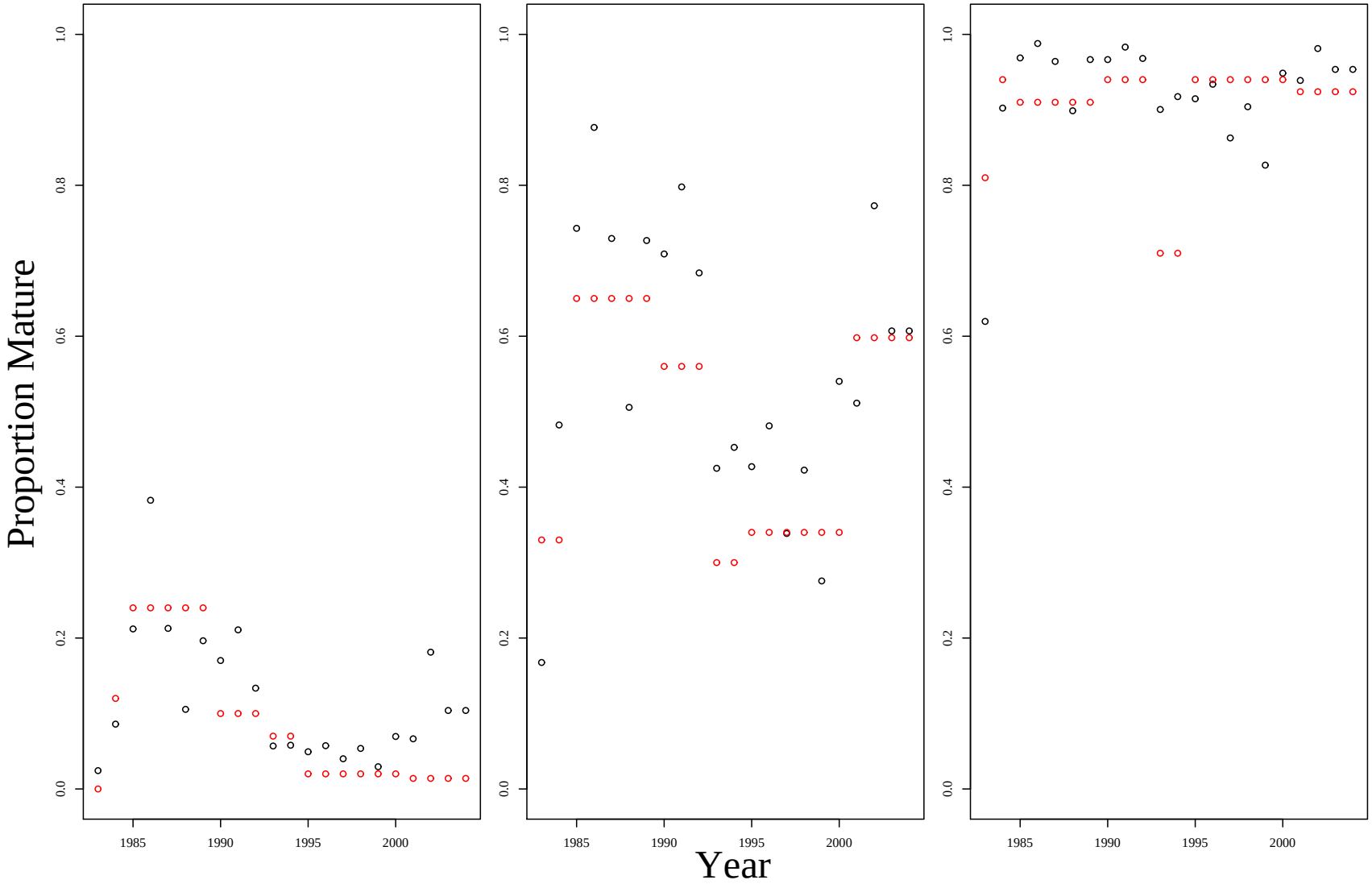

Figure 7. Kalman smoothed (black) proportion mature at age 1 (left) 2 (middle) and 3 (right) female Georges Bank haddock by year with corresponding estimates used in the 2005 assessment (red).

Table 1. Models fit with maximized log-likelihood, number of estimated parameters, χ^2 statistics for log-likelihood ratio tests and corresponding p-values.

Model	Parameter reduction	$\log(L)$	n_p	χ^2	P
M_f		-378.85	16	< 0.00	> 0.999
M_2	$\mathbf{\Phi} = 0$	-378.85	12	0.35	> 0.999
M_1	$\boldsymbol{\gamma}_1 = 0$	-379.03	8	220.82	< 0.001
M_0	$\mathbf{\Sigma} = 0$	-489.43	4		

Table 2. Estimated parameters and standard errors under model M_1. The coefficients $\boldsymbol{\beta}$ are for the initial state vector.

Parameter	Estimate	SE
β_1	-2.758	0.855
β_2	1.694	0.451
β_3	-2.989	1.315
β_4	1.229	0.499
$\Sigma_{1,1}$	0.351	0.170
$\Sigma_{2,2}$	0.090	0.048
$\Sigma_{3,3}$	0.750	0.344
$\Sigma_{4,4}$	0.106	0.055

Table 3. Proportions mature at age for females used in 2005 assessment for Georges Bank Haddock.

Year	1	2	3	4	5	6	7	8	9
1983	0.000	0.330	0.810	1.000	1	1	1	1	1
1984	0.120	0.330	0.940	1.000	1	1	1	1	1
1985	0.240	0.650	0.910	0.980	1	1	1	1	1
1986	0.240	0.650	0.910	0.980	1	1	1	1	1
1987	0.240	0.650	0.910	0.980	1	1	1	1	1
1988	0.240	0.650	0.910	0.980	1	1	1	1	1
1989	0.240	0.650	0.910	0.980	1	1	1	1	1
1990	0.100	0.560	0.940	0.990	1	1	1	1	1
1991	0.100	0.560	0.940	0.990	1	1	1	1	1
1992	0.100	0.560	0.940	0.990	1	1	1	1	1
1993	0.070	0.300	0.710	0.940	1	1	1	1	1
1994	0.070	0.300	0.710	0.940	1	1	1	1	1
1995	0.020	0.340	0.940	1.000	1	1	1	1	1
1996	0.020	0.340	0.940	1.000	1	1	1	1	1
1997	0.020	0.340	0.940	1.000	1	1	1	1	1
1998	0.020	0.340	0.940	1.000	1	1	1	1	1
1999	0.020	0.340	0.940	1.000	1	1	1	1	1
2000	0.020	0.340	0.940	1.000	1	1	1	1	1
2001	0.014	0.598	0.924	0.985	1	1	1	1	1
2002	0.014	0.598	0.924	0.985	1	1	1	1	1
2003	0.014	0.598	0.924	0.985	1	1	1	1	1
2004	0.014	0.598	0.924	0.985	1	1	1	1	1

Table 4. Smooth predictions of proportions mature at age for Georges Bank Haddock females.

Year	1	2	3	4	5	6	7	8	9
1983	0.024	0.168	0.620	0.929	0.991	0.999	1	1	1
1984	0.086	0.482	0.902	0.989	0.999	1.000	1	1	1
1985	0.212	0.743	0.969	0.997	1.000	1.000	1	1	1
1986	0.383	0.877	0.988	0.999	1.000	1.000	1	1	1
1987	0.213	0.730	0.964	0.996	1.000	1.000	1	1	1
1988	0.105	0.506	0.899	0.987	0.999	1.000	1	1	1
1989	0.196	0.727	0.967	0.997	1.000	1.000	1	1	1
1990	0.170	0.709	0.967	0.997	1.000	1.000	1	1	1
1991	0.211	0.798	0.983	0.999	1.000	1.000	1	1	1
1992	0.134	0.684	0.968	0.998	1.000	1.000	1	1	1
1993	0.057	0.425	0.900	0.991	0.999	1.000	1	1	1
1994	0.058	0.453	0.918	0.993	1.000	1.000	1	1	1
1995	0.049	0.427	0.915	0.994	1.000	1.000	1	1	1
1996	0.057	0.481	0.934	0.995	1.000	1.000	1	1	1
1997	0.040	0.339	0.863	0.987	0.999	1.000	1	1	1
1998	0.054	0.423	0.904	0.992	0.999	1.000	1	1	1
1999	0.029	0.276	0.827	0.984	0.999	1.000	1	1	1
2000	0.070	0.540	0.949	0.997	1.000	1.000	1	1	1
2001	0.066	0.511	0.939	0.996	1.000	1.000	1	1	1
2002	0.181	0.773	0.981	0.999	1.000	1.000	1	1	1
2003	0.104	0.607	0.954	0.996	1.000	1.000	1	1	1
2004	0.104	0.607	0.954	0.996	1.000	1.000	1	1	1