Fisheries Research 181 (2016) 48–62

Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres

Bayesian state-space approach to biomass dynamic models with skewed and heavy-tailed error distributions

Carlos Montenegro[a,*], Márcia Branco[b]

[a] Instituto de Fomento Pesquero, Chile
[b] Instituto de Matemática e Estatística, Universidade de São Paulo, Brazil

ARTICLE INFO

Article history:
Received 25 August 2015
Received in revised form 20 March 2016
Accepted 21 March 2016
Handled by A.E. Punt
Available online 12 April 2016

Keywords:
Nonlinear dynamic models
Logistic growth model
Surplus production model
State-space models
Bayesian MCMC
Chilean shrimp

ABSTRACT

We use the state-space approach to the logistic population growth model to update our knowledge of a population of marine shrimp off the Chilean coast. The unobserved state is the annual shrimp biomass, and the observation is the mean catch per unit effort. The observation equation is linear, and the state equation is nonlinear. The models include normal, student-*t*, skew-normal, and skew-*t* distributions for additive observation errors; and log-normal, log-*t*, log-skew-normal, and log-skew-*t* distributions for multiplicative observation errors. We use Bayesian approach to obtain inference, and the posterior distributions are approximated using *Markov chain Monte Carlo* methods. Deviance Information Criteria are lower in models considering log-skew-normal and log-skew-*t* observation errors. Furthermore, considering the posterior predictive distributions of the autocorrelations of the observation errors, these two models work best for the analyzed data set.

1. Introduction

Biomass dynamic models (also known as production models or surplus production models) have a long history in fisheries science and have provided a key basis leading to the popularity of maximum sustainable yield (*MSY*) and its associated biomass (B_{MSY}) as biological reference points for fisheries management (Punt, 2003). Such models can be used in situations when the only available data are time series of abundance indices and landings (i.e., harvesting). Surplus production models relate to Russell's (1931) formulation of stock dynamics and, in a difference equation or discrete form, have the general structure:

$$X_{t+1} = X_t + g(X_t) - H_t, \quad X_t \geq 0, \quad H_t \geq 0, \quad t = 0, 1, \ldots, \tag{1}$$

where X_t is the biomass of the population at time t (unobservable state), $g(X_t)$ is the growth function, and H_t is the biomass caught during time t. The logistic growth model is the most used formulation of biomass dynamic models (Haddon, 2001). The logistic form of Schaefer (1954) assumes that the production curve is symmetrical around a biomass B_{MSY} that can produce a maximum sustainable yield. Other options of growth functions include Gompertz, Richards, Ricker, and Beverton-Holt, among others (Birch, 1999; De Lara and Doyen, 2008). For simplicity, in this paper, we consider the logistic growth model of Schaefer (1954), which is represented by the differential equation:

$$g(X_t) = \frac{dX_t}{dt} = rX_t\left(1 - \frac{X_t}{K}\right) \equiv rX_t - \frac{rX_t^2}{K}, \quad r > 0, \quad K > 0, \tag{2}$$

where r is the intrinsic rate of population growth and K is a parameter that corresponds to the unfished equilibrium population size, known as carrying capacity (Krebs, 1985). The difference equation version of the model, considering the presence of harvesting and additive process noise, is:

$$X_{t+1} = X_t + rX_t\left(1 - \frac{X_t}{K}\right) - H_t + u_t, \tag{3}$$

where u_t are independent and identically distributed random variables with zero mean and finite variance (i.e., a white noise process). Several useful management quantities can be derived from surplus production models (Hilborn and Walters, 1992; Brodziak and Ishimura, 2011). Three of them are *MSY*, biomass that maximizes surplus production (B_{MSY}), and harvest rate that maximizes surplus production (H_{MSY}):

$$MSY = \frac{rK}{4}, \quad B_{MSY} = \frac{K}{2}, \quad H_{MSY} = \frac{r}{2}. \tag{4}$$

* Corresponding author.
E-mail addresses: cmontene@ime.usp.br, carlos.montenegro@ifop.cl (C. Montenegro), mbranco@ime.usp.br (M. Branco).

http://dx.doi.org/10.1016/j.fishres.2016.03.021
0165-7836/

When process noise is assumed to be multiplicative, the state equation becomes:

$$X_{t+1} = \left(X_t + rX_t\left(1 - \frac{X_t}{K}\right) - H_t\right)e^{u_t}. \quad (5)$$

A common assumption is that an abundance index is proportional to the biomass, i.e.:

$$Y_t = qX_t + \nu_t, \quad Y_t \geq 0, \quad q > 0, \quad (6)$$

where Y_t is a relative biomass index, q is a constant of proportionality known as the catchability coefficient, and ν_t are independent and identically distributed random variables (observation error). In practice, Y_t is often the mean fishing yield, calculated as the total catch divided by the total fishing effort; this may be calculated from commercial fishing data or research surveys. If the observation error is assumed to be multiplicative, the observation equation becomes:

$$Y_t = qX_t e^{\nu_t}. \quad (7)$$

Several approaches can be used to estimate the unobservable states and parameters of such models. The main approaches are: (1) observation error models, (2) process noise models, and (3) process noise and observation error models (Polachek et al., 1993; de Valpine and Hastings, 2002; Punt, 2003). In the first approach, the only source of uncertainty is due to observation or measurement errors, i.e., the equation relating the observable with the unobservable states has a random component, whereas the process equation is exact. In the second approach, the only source of uncertainty is due to process noise, i.e., the observation equation is exact. The third approach admits uncertainty in both equations and can be formalized through state-space models (West and Harrison, 1997; Durbin and Koopman, 2001; Shumway and Stoffer, 2011). In this representation, the time evolution of the system under study is assumed to be determined by an unobservable vector, which is associated with a series of observable vectors. If the equations are linear and the errors are Gaussian, the model is a Linear Gaussian State-Space Model (LGSSM), and the estimation problem can be solved with the well-known Kalman filter (Petris et al., 2009).

A more flexible formulation for the state space model can be done considering a more general functional and random structure, as following:

Consider a state equation: $\mathbf{X}_{t+1} = \mathbf{f}_t(\mathbf{X}_t, \boldsymbol{\omega}_t)$, $\boldsymbol{\omega}_t \sim g_l(\boldsymbol{\omega}_t)$, $t = 0, 1, \ldots$,
an observation equation: $\mathbf{Y}_t = \mathbf{h}_t(\mathbf{X}_t, \boldsymbol{\nu}_t)$, $\boldsymbol{\nu}_t \sim g_r(\boldsymbol{\nu}_t)$, and
the initial condition: $(\mathbf{X}_1|\mathbf{D}_0) \sim g_l(\boldsymbol{\omega}_0)$,

where $\mathbf{X}_t$ is a p-dimensional state vector, $\boldsymbol{\omega}_t$ is the l-dimensional process noise, f_t: $\mathbb{R}^p \times \mathbb{R}^l \to \mathbb{R}^p$ is a transition function, $\mathbf{Y}_t$ is a m-dimensional observation vector, $\boldsymbol{\nu}_t$ is the r-dimensional vector of observation error and h_t: $\mathbb{R}^p \times \mathbb{R}^r \to \mathbb{R}^m$ is an observation function. Moreover, $\boldsymbol{\nu}_t$'s and $\boldsymbol{\omega}_t$'s are white noises and have known probabilities densities functions $g_l(\boldsymbol{\omega}_t)$ and $g_r(\boldsymbol{\nu}_t)$. Usually the vectorial functions $\mathbf{f}_t$ and $\mathbf{h}_t$ are known but can depend on unknown parameters.

In our particular case, $\mathbf{f}_t(\mathbf{X}_t, \boldsymbol{\omega}_t)$ is given by (3), and $\mathbf{h}_t(\mathbf{X}_t, \boldsymbol{\nu}_t)$ is given by (6) and (7). To complete the model specification, it is necessary to propose probability distributions for the errors terms. In marine populations, the standard specification for such models is the log-normal distribution for the observation errors and process noise terms (Polachek et al., 1993; Meyer and Millar, 1999; Punt, 2003). However, given the asymmetry of the ecological variables, we opt for alternatives that can better accommodate asymmetry. Outliers tend to have a large influence in frequentist and Bayesian models based on normal distributions (Berger, 1994). Chen and Fournier (1999) evaluate the impacts of outliers on the derivation of posterior distributions in a Bayesian analysis of von Bertalanffy's growth model. In a similar approach, Chen et al. (2003) review three approaches that can be used to develop robust frequentist or Bayesian stock assessment methods. Jiao and Chen (2004) couple a generalized linear model with the biomass dynamic model using an observation error estimator with normal, log-normal, Gamma, and Poisson distributions. All these approaches deal with either the presence of outliers or skewed data but not with both characteristics. More flexible classes of distributions have been proposed in the literature to deal with the issues of asymmetry and outliers (see Genton, 2004). One such class is the scale mixtures of skew-normal (SMSN) distribution (Branco and Dey, 2001). Contreras-Reyes and Arellano-Valle (2013) propose SMSN distributions for modeling the age-length relationship of cardinalfish. They consider a non-linear regression model following the statistical methodology proposed by Lachos et al. (2011, 2013), which uses the maximum likelihood approach. Our modeling approach is similar to theirs, but in the context of dynamic models.

This study considers the following alternatives for the observation errors: additive normal, student-*t*, skew-normal, and skew-*t* error components and the multiplicative versions of these models, i.e., log-normal, log-student-*t*, log-skew-normal, and log-skew-*t*. We conduct a Bayesian inference with the logistic growth model with those observational error terms and apply it to data from a population of marine shrimp off the Chilean coast. The objectives of this study are to present the Bayesian state-space representation of a biomass dynamic model in the context of the SMSN class of distributions and to compare the performance of the several families of this class in terms of the data set of Chilean shrimp (*Heterocarpus reedi*). In the case of multiplicative observation errors, we compare the performance using a log-transformation. This study is organized as follows. Section 2 presents the models used in the Bayesian approach, including the families of skew-normal, student-*t*, and skew-*t* distributions. Extensions for the log versions are introduced, and some hierarchical forms derived from their stochastic representations are described. We also show the Bayesian formulation of the logistic growth model in state-space form. The prior distribution and the likelihood equation for the proposed models and the MCMC algorithm used to obtain the posterior distribution of the quantities of interest are presented as well. Section 3 presents our approach to the study of the data on annual catch per unit of effort (CPUE) and landings from a shrimp fishery off south-central Chile. Finally, Section 4 contains a brief discussion about our modeling strategy and other possible approaches. The Appendix contains some characteristics of the random variables used in this article. There, we explain the criteria used to analyze the results of the inferential process, and we provide graphics of the trace iterations of the MCMC procedure for each model parameter.

2. Materials and methods

2.1. Asymmetrical and heavy-tailed distributions

The usual notations, ϕ and Φ, respectively, for the probability density function (pdf) and the cumulative distribution function (cdf) of a standard normal random variable are considered here. Also, t_ν and T_ν indicate the pdf and cdf of a standard Student-*t* distribution with ν degrees of freedom (df).

One flexible class of distributions is the skew-normal family of densities (*SN*). The first systematic treatment of the skew-normal class in the scalar case was given by Azzalini (1985, 1986). A random variable *Y* has a skew-normal distribution if its pdf is:

$$p(y) = \frac{2}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)\Phi\left(\lambda\frac{y-\mu}{\sigma}\right), \quad y \in \mathbb{R},\ \mu \in \mathbb{R},\ \lambda \in \mathbb{R},\ \sigma > 0 \quad (8)$$