Chapter 24

Statistical Models for Climate Change Studies

T.V. Sathianandan
ICAR-Central Marine Fisheries Research Institute, Kochi

The recruitment of most marine fish is highly variable and most often it may not be visible to have an evident relation with the abundance of the parental stock (spawning stock biomass). Among other factors a major reason for this is the high and variable rates of mortality of the young ones (juvenile mortality) in the early life stages, which are assumed to be highly influenced by environmental processes. In particular, the larvae of pelagic fishes are highly sensitive to environmental changes and larval mortality rate will affect recruitment of fish into the fishery thus affecting their biomass and availability in the catch.

Being environment driven, effect of climate change on fish stock as well as catch has to be examined through appropriate modelling techniques. Depending on the biological features (recruitment size/age, maturity) it is wise to include lagged effects environmental variables in the model describing the fishery. More than one environmental variable can be included to evaluate their relationship with the variables of interest such as fish catch, stock biomass and catch rates or Landings Per Unit Effort (LPUE). While modelling the fishery incorporating environmental variables it is good to start with simple models with only one environmental variable and when it is found significant more than one variable or its combinations can be included in the model. In order to examine regional or seasonal effects we may attempt by including seasonal and regional data in the modelling exercise. In the model, Landings per unit effort (LPUE) can be response variables while fishing effort and environmental variables are explanatory variables. When we attempt modelling with a set of variables both response as well as auxiliary care should be taken to avoid including variables having collinearity. All data series may be tested for normality (Quantile-Quantile plots -QQ-plots) and collinearity (pairs plots, Zuur et al. 2010). Separate model/analysis for regions is necessary in order to account for the different environmental conditions and spatial independence among regions.

Model for Recruitment

A method generally used in fisheries to relate mature populations (*S*) with the recruitment abundance (*R*) at any given time *t*, was suggested by Ricker (1975), who proposed an exponential functional relationship of the following form between them.

$$R_t = S_t e^{a - bS_t}$$

$$y_t = \log\left(\frac{R_t}{S_t}\right) = a - b\,S_t + \epsilon_t$$

where *a* and *b* are describe population productivity at low density and capacity limited by density-dependence, respectively. If we assume log-normally distributed recruitment, then

Chen & Irvine (2001) extended the Ricker model (extended Ricker model) to include environmental variables as follows:

$$y_t = \log\left(\frac{R_t}{S_t}\right) = a - b\,S_t + \sum_i c_i\,Z_{i,t} + \epsilon_t$$

Statistical models

To evaluate the influence of different environmental variables on the response variables some of the suitable statistical models are:

- Dynamic Factor Analysis (DFA)
- Generalised Least Squares (GLS)
- Time series models

Different analysis may reveal varying results and hence suitable model selection criteria have to be used before finalising the model. Suggested criteria are Akaike's information criterion (AIC), decision tree etc.

Generalised Least Square Method

In linear least square method the model for the observed data Y in terms of the explanatory variables **X** is

$$\boldsymbol{Y = X\,\beta + \varepsilon}$$

where **Y** is the response vector, **X** is the explanatory variables matrix , $\boldsymbol{\beta}$ is the vector of coefficients and ϵ is the vector of random errors having mean vector zero and dispersion matrix σ^2**I** where **I** is the identity matrix. That is the random errors are independent and identically distributed with constant variance. But in many situations this assumption is violated due to unequal variance of the error terms and also they may be correlated. This limits the use of linear least square method and the alternative approach in such situations is Generalized Least Square Method. In generalised least square the model for observed data Y in terms of the explanatory variables **X** is

$$\boldsymbol{Y = X\,\beta + \varepsilon}$$

The error vector $\boldsymbol{\epsilon}$ has mean vector zero and dispersion matrix σ^2**V** where **V** is not an identity matrix. When **V** is diagonal with unequal values the errors are uncorrelated and when **V** is not diagonal the errors are correlated. In climate change related studies, we are interested in evaluating the influence of different climatic variables on fish yield. In the above model, the response vector can be the fish resources yield and the explanatory variables are the climate variables. In generalised least square model the estimate of parameter vector $\boldsymbol{\beta}$ is given as follows.

$$\boldsymbol{\beta} = \ (\boldsymbol{X'V^{-1}X})^{-1}\boldsymbol{X\,V^{-1}y}$$

Dynamic Factor Analysis

Dynamic factor analysis is a recent development in multivariate time series analysis where the objective is to estimate underlying common trends in a set of time series data and also to extract the influence of another set of explanatory time series towards the common trends. The mathematical expression of the model is

$$y_t = A \times z_t + B \times x_t + \varepsilon_t$$

$$z_t = z_{t-1} + e_t$$

Where, y_t is a vector time series with *k* time series components, z_t is the vector time series with *m* components representing the common trends (*m* < *k*), A is a matrix of order $k \times m$ (termed as factor loadings), x_t represents the explanatory variables time series vector with r components, B is another regression parameter matrix of order $k \times r$ and ε_t *and* e_t are vectors representing the noise/error components. For climate change studies vector y_t is composed of LPUE time series of fishery resources yield and the explanatory variable time series vector x_t is composed of time series data on climate variables.

Multivariate time series models models.

Values observed over a period of time can be treated as a time series process generated by a mechanism and can be studied in time series context to see the trend, inter-relations, periodicity etc., using time series techniques. By exploiting the inter-relations between time series sequences models can be developed to forecast future values with more precision. The most popular class of stationary stochastic model used for time series modeling is ARIMA model (Auto Regressive Integrated Moving Average model) introduced by Box and Jenkins (1976). This includes, autoregressive models, moving average models, random walk models, autoregressive moving average models, integrated models and seasonal models. In autoregressive (AR) models, the current value of a process is expressed as a linear aggregate of past values along with a random shock. The AR model of order *p*, denoted by AR(*p*), is of the form

$$\tilde{z}_t = \phi_1 \tilde{z}_{t-1} + \phi_2 \tilde{z}_{t-2} + \cdots + \phi_p \tilde{z}_{t-p} + \varepsilon_t$$

Using the back shift operator B, the AR(p) model can be written as

$$\phi(B)\tilde{z}_t = \varepsilon_t$$

Where

$$\tilde{z}_t = \phi_1 \tilde{z}_{t-1} + \phi_2 \tilde{z}_{t-2} + \cdots + \phi_p \tilde{z}_{t-p} + \varepsilon_t$$
$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$$

is the characteristic polynomial of the AR(p) model. In moving average (MA) models, the current value of a process is expressed as a linear combination of a finite number of previous random shocks. The MA model of order *q*, denoted by MA(*q*), has the mathematical form

$$\begin{aligned}\tilde{z}_t &= \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots - \theta_q\varepsilon_{t-q} \\ &= (1-\theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)\varepsilon_t \\ &= \theta(B)\varepsilon_t\end{aligned}$$

The combined model is the Auto Regressive Moving Average model denoted by ARMA(p,q) having the following mathematical expression

$$\begin{aligned}&\tilde{z}_t - \phi_1\tilde{z}_{t-1} - \phi_2\tilde{z}_{t-2} - \cdots - \phi_p\tilde{z}_{t-p} = \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots - \theta_q\varepsilon_{t-q} \\ &(1-\phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)\tilde{z}_t = (1-\theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)\varepsilon_t \\ &\phi(B)\tilde{z}_t = \theta(B)\varepsilon_t\end{aligned}$$

Each of the above models has modified versions to incorporate explanatory time series variables into the model. Similarly there are multivariate versions for each the models where in more than one time series are modeled simultaneously known as vector time series models. For example the model with the addition of an explanatory variable is known as ARX model and its expression will be as follows.

$$\boldsymbol{\phi(B)\tilde{Z}_t = \beta_0 x_t + \beta_1 x_{t-1} + \cdots + \beta_r x_{t-r} + \varepsilon_t}$$

The above model is capable of extracting the influence of an explanatory variable for different lags such as climate on the response variable which can be the fish yield.