

CMFRI

Winter School on
Impact of Climate Change
on Indian Marine Fisheries

Lecture Notes

Part 1

Compiled and Edited by

E. Vivekanandan and J. Jayasankar

Central Marine Fisheries Research Institute (CMFRI),
(Indian Council of Agricultural Research)
P.B. No. 1603, Cochin - 682 018, Kerala

(18.01.2008 - 07.02.2008)



TRENDS, SEASONALITY, CYCLICITY – A TIME SERIES VIEW POINT



J. Jayasankar

Central Marine Fisheries Research Institute, Kochi

(jjsankar@gmail.com)

Introduction

Analysis of climatic data unwaveringly involves two factors of tagging viz space and time. If our focus is on the upheavals in these factors in the long run temporally distributed data is the basic requirement. These types of observations on one or more variables that occur in a time sequence are called as time series (TS). Mostly these observations are collected at equally spaced, discrete time intervals. One of the enduring assumptions pertaining to the time series is that some aspects of the past pattern will continue to remain in the future. Also in this setup the past values of the main variable almost fully explains the realizations unlike typical regression setups wherein exclusive explanatory variables would have been drafted. One offshoot of this phenomenon is the fact that analyses may often lead to answers for the query “What next?” rather than “Why so?” Even the forecasts based on the time series should be considered more on the short-term context rather than the long-term perspective. Another rule of the thumb suggested by exponents is that there should be at least 50 observations for a stable TS analysis.

Components of time series and decomposition:

The first major step into analyzing time series data is to consider types of data patterns, so that appropriate models can be utilized. The basic components of TS are

- (i) Horizontal – when data values fluctuate around a constant value
- (ii) Trend- when there is long term increase or decrease in the data.
- (iii) Seasonal- when the series is influenced by seasonal factor and recurs on regular periodic basis
- (iv) Cyclical- when the data course exhibits rise and falls that are not for a fixed period

Many data series have a combination of these preceding patterns. Hence the major process is to separate out the contribution of these four major components to the data and whatever remains in an unidentifiable form is the “random” or “error” component. Time plot (data plotted over time) and seasonal plot (data plotted against individual seasons in which the data were observed) help in visualizing these patterns while exploring the data. A crude yet practical way of decomposing the original data (ignoring cyclical pattern) is to go for a seasonal decomposition either by assuming an additive or multiplicative model.

An additive model will be of type

$$Y_t = T_t + S_t + E_t$$

And a multiplicative model would resemble

$$Y_t = T_t \cdot S_t \cdot E_t$$

Where

Y_t – Original TS data

T_t – Trend component

S_t – Seasonal component

E_t – Error/ Irregular component

If the magnitude of TS varies with the level of the series then one has to go for a multiplicative model. This decomposition may enable one to study the TS components separately or will allow workers to de-trend or to do seasonal adjustments if needed for further analysis.

Moving Averages and Exponential Smoothing Methods

Moving averages as a smoothing tool is an exploratory mechanism which is used to test the waters and prepare for major analyses. Whereas towards the end of this document we may come across another component called moving average which is a model ingredient of time domain approach to estimating the TS. In this section moving average as a smoothing tool will be highlighted.

Simple Moving Averages

A Moving Average (MA) is simply a numerical average of the last N data points. In general, the MA at time t, taken over N periods, is given by

$$M_t^{[1]} = \frac{Y_t + Y_{t+1} + \dots + Y_{t-N+1}}{N}$$

Where Y_t is the observed response at time t. Another way of stating the above equation is

$$M_t^{[1]} = M_{t-1}^{[1]} + (Y_t - Y_{t-N})/N$$

At each successive time period the most recent observation is included and the farthest observation is excluded for computing the average. Hence the name 'moving' averages.

Double MA

Basically the simple MA with superscript [1] is intended for data of constant nature. But if the data has a linear or quadratic trend, the simple MA will not be conclusive. To circumvent this issue the simple MAs are subjected to another round of averaging as mentioned in the previous section.

Simple Exponential Smoothing (SES)

If we denote the time series data as Y_1, Y_2, \dots, Y_t and if it is desired to forecast the next value of the TS viz Y_{t+1} that is yet to be observed as a forecast, F_t . Then the forecast F_{t+1} is based on weighting the most recent observation Y_t with a weight value α and weighting the most recent forecast F_t with a weight of $(1-\alpha)$ where α is a smoothing constant/ weight between 0 and 1. Thus the forecast for the period t+1 is given by

$$F_{t+1} = F_t + \alpha(Y_t - F_t)$$

The choice of the smoothing coefficient α has huge impact on the forecast. A large value of α (near 0.9) gives very little smoothing in the forecast, whereas a small value of α (near 0.1) gives considerable smoothing. Alternatively, one can choose α from a grid of values (say $\alpha=0.1, 0.2, \dots, 0.9$) and choose the value that yields the smallest Mean Squared Error value.

Once these chain of values of F_{t+1} are presented as functions of α , past values Y_t and F_t , the main point of focus for completing the series happens to be the initial value F_1 . One method of initialization is to use first observed value Y_1 as the first forecast ($F_1=Y_1$) and then proceed. Another possibility would be to average the first four or five values in the data set and use this as the initial forecast. However, because the weight attached to this user-defined F_1 is minimal its effect on F_{t+1} is negligible.

Double Exponential Smoothing (Holt)

This is performed to allow forecasting data with trends. This linear exponential smoothing method propounded by Holt has two equations compared to the lone one of simple exponential smoothing model to deal with – one each for level and trend. There are two smoothing parameters (weights) α and β which can be chosen from a grid of values (say, each combination of $\alpha=0.1, 0.2, \dots, 0.9$ and $\beta=0.1, 0.2, \dots, 0.9$) and then

select the combination of α and β which correspond to the lowest Mean Squared Error.

Triple exponential Smoothing (Winters)

This method is recommended when seasonality exists in the time series data. This method is based on three smoothing equations- one for level, one for trend and one for the seasonality. It is similar to Holt's method, with one additional equation to deal with seasonality. In fact there are two different Winter's methods depending on whether seasonality is modeled in an additive or multiplicative way.

Stationarity of a TS process

A TS is said to be stationary if its underlying generating process is based on a constant mean and constant variance with its level of dependence occurrence of a given state's level on the previous states' levels being similar throughout. Thus if different sets of a realization are considered the different subsets will typically have means, variances and inter state similarity that do not differ significantly. Statistical theory provides a conclusive test for testing the stationarity, which is called as Dickey- Fuller test. It is multiple regression based test on differenced series ($Y_t - Y_{t-1}$) and their previous state expressions.

Autocorrelation functions

Autocorrelation

Autocorrelation is a measure of similarity of the current state's expression with previous state's expressions in a TS is measured by the simple correlation between current observation (Y_t) and observation from p periods before the current one (Y_{t-p}). That is for a given series Y_t , autocorrelation at lag p=correlation (Y_t, Y_{t-p}) and is given by

$$r_p = \frac{\sum_{t=1}^{n-p} (Y_t - \bar{Y})(Y_{t-p} - \bar{Y})}{\sum_{t=1}^n (Y_t - \bar{Y})^2}$$

Obviously the value of autocorrelation ranges from -1 to +1. As a rule of the thumb the maximum number of useful rp are roughly n/4 where n is the number of periods upon which information on Y_t is available.

Partial Autocorrelation

Partial autocorrelations are used to measure the degree of association between Y_t and Y_{t-p} when the Y-effects at other time lags 1,2,3,..., p-1 are removed.

Autocorrelation Function (ACF) and Partial autocorrelation Function (PACF)

Theoretical ACFs and PACFs, which can be termed as a measure of the autocorrelations versus lags, are available for the various models chosen. Thus comparing the plots of sample ACFs versus lags with these theoretical ACFs and PACFs is an important step in selecting time domain models.

The exploration of the TS invariably leads construction of models which can be specific use to pin pointing the underlying system and hence can play a major role in forecasting. These sets of regression equations (not necessarily linear in form) are broadly categorized as time domain models. The models can be subclassified into those purely based on autocorrelation (AR), those purely having correlated errors (not noises) (MA) and a combination of both (ARMA). The general characteristics of theoretical ACFs and PACFs are as follows: (here 'spike' represents the line at various lags in the plot with length equal to magnitude of autocorrelations)

Model	ACF	PACF
AR	Spikes decay towards zero	Spikes cutoff to zero
MA	Spikes cutoff to zero	Spikes decay to zero
ARMA	Spikes decay to zero	Spikes decay to zero

Time domain modeling

After exploration of the TS the precise setup of the cause and effect issues have to be finalized. Modeling is the best tool towards achieving this end, which once done satisfactorily can lead to more reliable forecasts.

ARIMA Modeling

In general, an ARIMA model is characterized by the notation ARIMA(p,d,q) where p,d and q denote orders of auto-regression, integration (which indirectly implies need for differencing before fitting the model) and moving average respectively. In ARIMA, TS is a linear function of past actual values and random shocks. For instance for a TS process {Y_t}, a first order auto-regressive process is denoted by ARIMA (1,0,0) or simply AR(1) and is given by

$$Y_t = \mu + \phi_1 Y_{t-1} + \varepsilon_t$$

and a first order moving average process is denoted by ARIMA (0,0,1) or simply MA(1) and is given by

$$Y_t = \mu - \theta_1 \varepsilon_{t-1} + \varepsilon_t$$

Alternatively, the model ultimately derived may be a mixture of these processes and of higher orders as well. Thus a stationary ARIMA (p,q) process is defined by the equation

$$Y_t = \mu + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p} - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q} + \varepsilon_t$$

where ε_t 's are independently and normally distributed with zero mean and constant variance σ^2 for $t=1,2,\dots,n$. The values of p and q in practice lie between 0 and 3.

Seasonal ARIMA modeling

Identification of relevant models and inclusion of suitable seasonal variables are necessary for seasonal modeling and their applications, say forecasting SST for the coming decade. Seasonal factors of production of principal crops are of greater utility for planners, administrators and researchers alike. Agricultural seasons vary significantly among the states of India. Similarly seasonal catch and composition of fish landed across the coast of India has a definitive pattern. The seasonal issues which relate to the reproductive biology of commercially tapped species have a big say in the type, quantity and quality of the landed resources, that again at different points of a year repeatedly. Thus seasonal forecasts of crop production can also be made using seasonal ARIMA models.

The seasonal ARIMA model is defined by a shorter notation i.e. ARIMA (p,d,q) (P,D,Q) where the p denotes the order of (lags to be considered) autoregression, q that of moving average, both for the non-seasonal components. The P,Q denote the respective retrospection on seasonal autoregression and moving average. d and D denote the differencing (for achieving stationarity) required in both non-seasonal and seasonal terms. The equation commonly represented in the back-shift operator notation (a convenient

notation which indicated the extent to which the operation has to be carried out towards the back values, for eg. $B(Y_t)$ means Y_{t-1} and $B^2(Y_t)$ means Y_{t-2} and so on) has left hand side indicating the four components viz non-seasonal autoregression, seasonal autoregression, non-seasonal differencing and seasonal differencing operating over the present realization, Y_t , and the right hand side containing the non-seasonal and seasonal moving averages followed by the white noise. The seasonal autoregression has lags made up of standard time gaps which are obviously more than one state unit. For example quarterly effects are denoted by lags each of three months each, i.e. present state having regressed upon the state three months back and so on.

The Art of ARIMA modeling

Though the very nature of the model albeit expressed in simple notations, has a lot of computational rigours, software have reduced that part of the burden enormously. Hence a typical construction of an ARIMA setup to a time series data can be conveniently termed as an art which has three essential stages.

Identification

The foremost step in the process of modeling is to check for the stationarity of the series as the estimation procedures are available only for stationary series. Though in a typical ARIMA integration is mentioned as the part and parcel of modeling, the computation is limited to fitting autoregression and moving average coefficients after subjecting the data to required differencing. There are two types of stationarity, viz. stationarity in mean and stationarity in variance. A cursory look at the time series plot coupled with the study of AR and PAR spikes will provide clues for the presence of stationarity. Another way to check the same is to fit a first order autoregressive model for the raw data and test whether the coefficient ' ϕ_1 ' is less than one. Another full fledged statistical test provided by most of the analytical software is the Dickey Fuller test. Though the mean stationarity could be achieved by differencing, variance based stationarity may require transformation of data like log transformation.

Thus if ' X_t ' denotes the original series, the non-seasonal difference of the first order is

$$Y_t = X_t - X_{t-1}$$

Followed by the seasonal differencing (if required)

$$Z_t = Y_t - Y_{t-s} = (X_t - X_{t-1}) - (X_{t-s} - X_{t-s-1})$$

The next step in the identification process is to find the initial values for the orders of seasonal and non-seasonal parameters, p,q and P,Q. The could be obtained by looking for significant autocorrelation and partial autocorrelation coefficients . Say, if second order autocorrelation coefficient is significant , then AR(2) or MA(2) or ARMA(2,2) model could be tried to start with. This however is not a hard and fast rule, as sample autocorrelation coefficients are poor estimates of population autocorrelation coefficients. Still they can be used as initial values while the final models are achieved after going through stages repeatedly.

Estimation

At the identification stage, one or more models are tentatively chosen that seem to provide statistically adequate representations of the available data. Then precise estimates of parameters of the model are obtained by least squares as advocated by Box and Jenkins. Standard computer packages like SAS, SPSS etc. are available for finding the estimates of relevant parameters using iterative procedures.

Diagnostics

Different models can be obtained for various combinations of AR and MA individually and collectively. The best model is obtained with the following diagnostics:

Low Akaike Information Criteria (AIC)/ Bayesian Information Criteria (BIC)/ Schwarz-Bayesian Information Criteria (SBC)

AIC is given by $AIC = (-2 \log L + 2m)$ where $m = p + q + P + Q$ and L is the likelihood function. Since $-2 \log L$ is approximately equal to $\{n(1 + \log 2\pi) + n \log \sigma^2\}$ where σ^2 is the model mean squared error (MSE), AIC can be written as $AIC = \{n(1 + \log 2\pi) + n \log \sigma^2 + 2m\}$ and because first term in this equation is a constant, it is usually omitted while comparing models. As an alternative to AIC, sometimes SBC is also used which is given by $SBC = \log \sigma^2 + (m \log(n))/n$.

Non-significance of auto correlations of residuals via tests

After tentative model has been fitted to the data, it is important to perform diagnostic checks to test the adequacy of the model and if need be to suggest potential improvements. One way to accomplish this is through the analysis of residuals. It has been found that it is effective to measure the overall adequacy of the chosen model by examining a quantity Q known as Box- Pierce statistic (a function of autocorrelations of residuals) whose approximate distribution is chi-square and is computed as:

$$Q = n \sum r^2(j)$$

Where summation extends from 1 to k with k as the maximum lag considered, n is the number of observations in the series, $r(j)$ is the estimated autocorrelation at lag j ; k can be any positive integer and is usually around 20. Q follows Chi-square with $(k - m_1)$ degrees of freedom with m_1 as the number of parameters estimated in the model. A modified Q statistic is the Ljung-Box statistic which is given by

$$Q = n(n+2) \sum r^2(j) / (n-j)$$

The Q statistic is compared to critical values from chi-square distribution. If model is correctly specified, residuals should be uncorrelated and Q should be small. A significant value indicates that the chosen model does not fit well.

As a cautionary note while performing all these steps, considerable care is required and the approach should always be to denounce any zeroed - in model so that the final selection will be the one that has come through an exhaustive search.

References and suggested reading

- Blank, D.S 1986 SAS system for forecasting time series, SAS Institute Inc, USA
- Box, GEP, Jenkins, G.M. and Reinsel, G.C. 1994 Time series analysis: Forecasting and control, Pearson Education, Delhi
- Makridakis, S., Wheelwright, S.C. and Hyndman, R.J. 1998. Forecasting: Methods and Applications, John Wiley, New York.
- Pankratz, A. 1983. Forecasting with univariate Box- Jenkins models: concepts and cases, New York: John Wiley & Sons.