Chapter 2

Methodology

In this chapter we describe the source of data and summaries the statistical methods used to analyze these data. These methods include those used both for preliminary data analysis and for modeling.

The preliminary data analysis methods involve summarizing the data, checking the need for a data transformation. The methods of modeling involve various linear regression models.

All of the graphical, statistical model fitting and assessing the goodness-of-fit were carried out using program written in R statistical system (Venables and Smith 2004).

2.1 Source of data

The fish catch distribution describes the past and current status of catch dynamics in the aquatic system. Thus, the distribution of fish catch is a method of determining the presence and stocking of the quantity of fish caught in any particular area regardless of its species, volume, fishing ground, weather, size, and age of maturity (Sparre and Venema 1998). A complete fish catch assessment always contains a vast array of information of annual landed catch, age distribution, index of relative abundance, recruitment, natural mortality and fishing mortality, and biological characteristics (Cooper 2006).

A fish cluster defines a group of species, regardless of taxonomic position, that uses environmental resources in a similar way (Clements et al 2006). Fish cluster analysis has

been proposed as a potential tool for species conservation purposes, but the methodology is still largely undeveloped.

Apart from skipper logbooks and research surveys, the most common sources of catch data are fishery landing records and port samples. Landing records, which result directly from the sale of caught fish, provide information only on landed commercial catches. Generally, fishery data are recorded by daily trip of fishing boat and hence this is used as a measure of catch weights in kilograms separating by species and by gear used (Isaac 2002).

In this study, we use the data from two difference sources. The first source is the total fish catch annual record from January 1977 to December 2006, routinely recorded by statistical officers of three regional fisheries offices within the Department of Fisheries, Ministry of Agriculture and Co-operatives, Thailand, namely the Songkhla Provincial Fisheries Office, the Phattalung Provincial Fisheries Office and the National Institute of Coastal Aquaculture (NICA). The second source comprises the commercial fish catch weight data collected from ten major fish landing sites around the entire Songkhla Lake. From January 2003 to September 2005, these data were collected by the National Institute of Coastal Aquaculture (NICA) of the Department of Fisheries of Thailand, and thereafter to December 2006 by the candidate. Species were identified and categorized following Choonhapran (1996). The data consist of 127 fish species commonly caught by fishermen, using three fishing gear types (set bag net, trap and gill net) for each month over a four year period from 2003–2006. The variables from the two sources are shown in Figure 2.1.

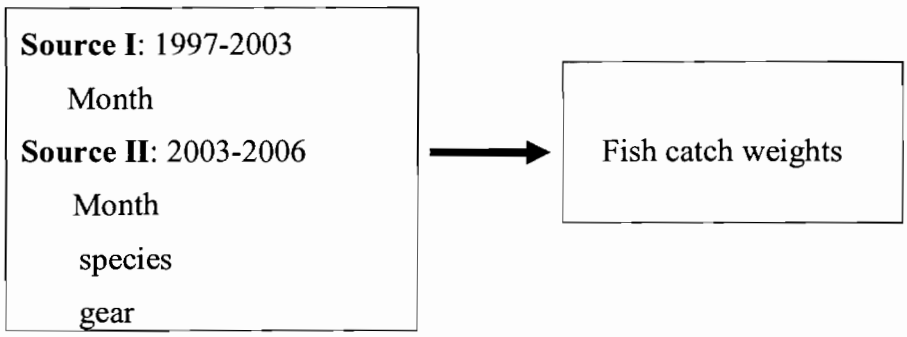

Figure 2.1: Path diagram for study

2.2 Statistical methods

Data transformation

A statistical linear model involves the assumptions that the errors from the model should be normally distributed with homogeneity of variance. If not, the data may need to be transformed. Assuming non-negative data, the common transformations are the logarithm, square root, or possibly other powers. The distribution of a single continuous variable can be displayed using a histogram. If data are heavily skewed, we might consider using logarithms since this is the extreme of the positive power transformations. The logarithm transform can be base 2 or base 10 or natural logarithms (base e = 2.7183..). The base does not affect the shape of the resulting distribution. It just affects the scale. It should be noted that if zeros occur in the data, some adjustment should be made before taking the logarithm, such as adding a small constant to all values or simply replacing the zeros by a small constant.

To decide on the transformation, we can group the data into sub-samples and use a scatter plot to display the relation between the logarithms of the standard deviation and the logarithms of the means of the sub-samples. Then we fit a straight line to find the slope. If a linear relation is apparent on the graph, the data should be transformed (Tukey 1977),

where the magnitude of the slope determines the transformation. If the slope is close to 1, taking logarithms is recommended.

In this study, the time series data of total fish catch weight tonnage (w_t) at time t, were transformed by using natural logarithms to meet statistical analysis requirements. The transformation is thus

$$y_t = \ln(w_t). \tag{2.1}$$

Linear regression model on time series

Linear regression is used to assess the relation between two continuous variables. If the determinant is time, the data constitute a time series. A time series is then a continuous set of data measured sequentially in time at regular intervals. The main objectives of analyzing time series data are to forecast future values of the series, to estimate the trend or overall character of the series, to model the dynamic relations between two or more time series, and to summarize these characteristic features.

The simplest regression model for $t = 1, 2, \ldots, n$ is

$$y_t = \alpha + \beta t + \varepsilon_t \tag{2.2}$$

where y_t is the outcome variable or the continuous set of data measured at times t, α is a constant, β is a slope or trend, and ε_t are independent random errors with mean zero and common variance σ^2.

If we take into account a seasonal effect, equation (2.2) can be extended to

$$y_t = \alpha + \beta t + \gamma_s + \varepsilon_t \tag{2.3}$$

where $s = 0, 1, \ldots, S$, and $\gamma_1, \gamma_2, \ldots, \gamma_S$ comprise a set of seasonal effects $\{\gamma_s\}$ with $\gamma_1 = 0$.

We also need to take into account correlation between successive observations, and this can be achieved by extending equation (2.3) to include previous observations as predictors in the model. For example, if the preceding two observations are used, the model is

$$y_t = \alpha + \beta t + \gamma_s + \delta_1 y_{t-1} + \delta_2 y_{t-2} + \varepsilon_t \qquad (2.4)$$

where y_t is the outcome in month t, and δ_1, δ_2 are the coefficients denoting the influence of this outcome in the previous two months. This model is thus an observation-driven model (Cox 1981).

If y_t is the log-transformed value of w_t, to obtain forecasts for w_{t+k}, where k is a future time period, equation (2.4) must be transformed back by exponentiation and the forecast is the mean of w_t. It may be shown using statistical theory that w_t has a log-normal distribution with expected value

$$E[w_t] = \exp\left(\mu + \tfrac{1}{2}\sigma^2\right) \qquad (2.5)$$

where μ and σ are the mean and standard deviation, respectively, of y_t. Thus if σ is the standard deviation of the errors in the fitted regression model, the forecast of w_{t+k} is

$$E[w_{t+k}] = \exp\left(\alpha + \beta(t+k) + \gamma_s + \delta_1 y_{t+k-1} + \delta_2 y_{t+k-2} + \tfrac{1}{2}\sigma^2\right) \qquad (2.6)$$

Additive model

In this study we are fitting models with the two factors as predictors. These are fish species (s) and observation period (t). The two-way ANOVA method is the simplest regression model for this situation, and is expressed as

$$y_{st} = \mu_s + \nu_t + \varepsilon_t \qquad (2.7)$$

where y_{st} is the outcome for $s = 1, 2, \ldots, S$ and $t = 1, 2, \ldots, T$. S and T denote the numbers of fish species and observation periods, respectively. This model has $S+T-1$ parameters. The set $\{\mu_s\}$ comprises the mean outcome of fish species s. The model assumes that the distribution of a species group over the period is the same for all species, only differing in level through the set $\{\nu_t\}$ the mean outcome for observation period t, so the model also assumes that the temporal pattern $\{\nu_t\}$ is the same for all fish species.

A more general model that overcomes these restrictions is

$$y_{st} = \mu_s + \alpha_s \beta_t + \varepsilon_t \qquad (2.8)$$

where one of the sets $\{\alpha_s\}$, $\{\beta_t\}$ needs to be scaled to avoid redundancy. The number of parameters in this model is $2S + T - 2$.

The model fitting is not simple because the model is non-linear. However, if it is assumed that $S > T$, Theil (1983) showed that the least squares solution is obtained by choosing, μ_s as the row mean of y_{st} ($\bar{y}_{s\cdot}$) and the set $\{\beta_t\}$ as the eigenvector corresponding to the largest eigenvalue of the matrix $Y_c^T Y_c$, where Y_c is the row mean adjusted data matrix with elements $y_{st} - \bar{y}_{s\cdot}$, and Y_c^T is its transpose. The estimated values of α_s are then given by $\alpha_s = \sum_{t=1}^{T} \beta_t y_{st}$. This model is also widely used in demography for mortality forecasts, where it is know as the Lee-Carter model (Lee and Carter 1992, Booth and Tickle 2007). More generally equation (2.8) can be extended to include additional sets of parameters, with the model taking the form

$$y_{st} = \mu_s + \sum_{k=1}^{m} \alpha_s^{(k)} \beta_t^{(k)} + \varepsilon_t \qquad (2.9)$$

In this model set $\{\beta_t^{(k)}\}$ is the eigenvector of $Y_c^T Y_c$ corresponding to k^{th} largest eigenvalue, and $\alpha_s^{(k)} = \sum_{t=1}^{T} \beta_t^{(k)} y_{st}$ as before.

Noted that eigenvectors are scaled to have sum of squares equal to 1 and pairs of eigenvectors have sum of products equal to 0.

Since the parameters $\{\beta_t^{(k)}\}$ are estimated and the $\{\alpha_s^{(k)}\}$ are assumed fixed, the $\{\alpha_s^{(k)}\}$ should be scaled to remove redundancy. The model may also be regarded as a linear regression model, provided it is assumed that the $\{\alpha_s^{(k)}\}$ parameters are fixed, rather than determined by the data. In this case the $\{\mu_s\}$ and $\{\beta_t^{(k)}\}$ parameters are estimated from the regression model and these estimates are essentially the same as those given by the method described above. The advantage of this approach is that standard errors are obtained.

However, given that the predictors $\{\beta_t^{(k)}\}$ are actually obtained from the data, the number of residual degrees of freedom is effectively reduced by the number of $\{\alpha_s^{(k)}\}$ parameters, that is, $m(T-2)$, allowing for the fact that these sets of parameters are scaled.

Goodness-of-fit

A residuals plot is one of the methods for assessing the adequacy of the normality assumption of the error. It is a scatter plot of the residuals against corresponding quintiles from the standardized normal distribution. If the error is normally distributed, the points should be approximately linear.