Chapter 2

Materials and Methods

This chapter describes the research methodology used in the study. First we describe the study design and data source and management. Conceptual framework is also described. Finally we identify data analysis and the statistical methods used.

2.1 Study design

The study design is cross-sectional type, with the observations being monthly catch weight in kilograms, The outcome of interest was the quantities of fish landing on unit price less than, or equal to, 25 baht per kilogram.

Definition of low value fish

Low value fish were fish landing on unit price less than, or equal to, 25 baht per kilograms, it was referred to Khemakorn *et al*. (2005). They found that the price of low value food fish has increased continuously from 3.44 baht/kg in 1971 and 18.31 baht/kg in 1995. The price fluctuated but showed the increasing trend with 28.5 baht/kg in 2002.

2.2 Data source and management

Data were taken from ten major commercial fish landing sites around the Songkhla Lake: Khu Tao (1), Kuan Nieng (2), Pak Pa Yoon (3), Jong Ke (4), Lampam (5), Khu Kud (6), Ko Yai (7), Ra Nod (8), Thale Noi (9), and Hua Khao Daeng (10) (Chesoh, 2009) as shown in Figure 2.1.

Data on 127 aquatic species were collected monthly from January 2003 to December 2006 by the National Institute of Coastal Aquaculture (NICA) of the Department of Fisheries of Thailand. The main fishing gears are trap, set bag net and gill net.

There are five variables; four year, 12 months, 127 species, weight of the catch and three main gear types used. Year and month of catch variables were aggregated into the season term, with 48 seasons.

The catches were classified as low value fish if unit price was less than, or equal to, 25 baht per kilogram. Fish caught were also classified in term of fish group according to biological characteristics (vertebrate and invertebrate) and their living habitat (freshwater, brackish and marine) (Choonhapran, 1996).

There are 6 groups if fish species: freshwater invertebrates, estuarine invertebrates, marine invertebrates, freshwater vertebrates, estuarine vertebrates and marine vertebrates.

The outcome is the catch weight of low value fish, the determinants are year, monthly, gear and fish group. Then, to satisfy the statistical assumption of constant variance, the catch weight of the low value fish needed to be transformed using natural logarithms.

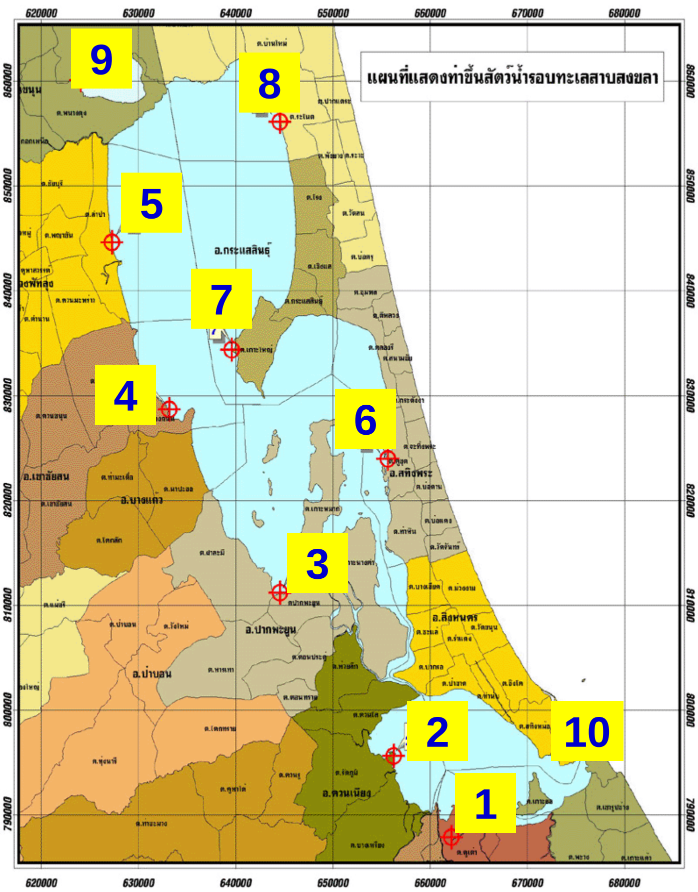

Figure 2.1: Fish landings in Songkhla Lake

2.3 Path diagram and variables

The path diagram of this study is shown in Figure 2.2. This study carried out statistical analyses for investigating the catch weight of low value fish with the determinant variables comprising year, month, gear, species and group.

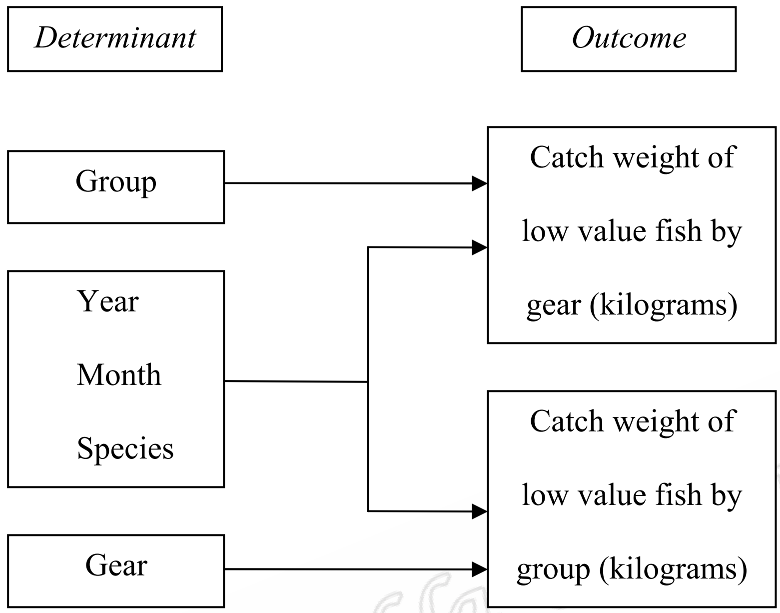

Figure 2.2: Path diagram of the study

2.4 Statistical methods

Multiple Linear Regressions Analysis

Linear regression is a method for the analysis of data in which the outcome variable is continuous, the determinant is categorical. If we have a continuous response y and more than one determinant variable, we use *multiple linear regressions*. The model is given by,

$$y = \beta_0 + \sum \beta_i x_i + \varepsilon \qquad (2.1)$$

where y is the outcome variable, β_0 is a constant, β_i is a set of parameters (i is the number of determinants), and x_i is a set of determinants. *The least square line* is the statistical analysis the line fitted, which minimizes the distances of the points to the line, measured in the vertical direction. This line is also called the *regression line*, and may be represented as

$$y = a + bx \tag{2.2}$$

where a is the *intercept* and b is the *slope* or *regression coefficient*.

Linear regression or *multiple linear regressions* have the same three assumptions as follows. First, the association is linear, the variability of the errors (in the outcome variable) is uniform and last, these errors are normally distributed. These assumptions may be assessed by examining the residuals. To assess the first two assumptions, the residuals should be plotted against *predicted values* given by the linear model. The normality assumption may be assessed by plotting the residuals against their normal scores, and tested using the Shapiro-Wilk test (McNeil, 1996).

Coefficient of determination

The *goodness-of-fit* of the least-squares line is defined as the square of the correlation coefficient or coefficient of determination or r^2 (McNeil, 2005).

The r^2 goodness-of-fit statistic may be expressed in terms of the sum of squares of the residuals and the sum of squares of the mean-corrected responses.

Suppose that y is the observed values and f is associated with the model predicted value. The variability of the data can be measured through different sums of squares:

$$SS_{tot} = \sum_i (y_i - \bar{y})^2 \tag{2.3}$$

$$SS_{err} = \sum_i (y_i - f_i)^2 \tag{2.4}$$

Where SS_{tot} is the total sum of squares and SS_{err} denoted as the sum of squared errors, also called the residual sum of squares. Thus coefficient of determination can be calculated as:

$$r^2 = 1 - \frac{SS_{err}}{SS_{tot}} \tag{2.5}$$

Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) is a way of selecting a model from a set of models. The chosen model is the one that minimizes the Kullback-Leibler distance between the model and the truth. It's based on information theory, but a heuristic way to think about it is as a criterion that seeks a model that has a good fit to the truth but few parameters. It is defined as

$$\text{AIC} = -2\ (\ln(\ \text{likelihood})) + 2K \tag{2.6}$$

where likelihood is the probability of the data given a model and K is the number of free parameters in the model. AIC scores are often shown as ΔAIC scores, or difference between the best model (smallest AIC) and each model (so the best model has a ΔAIC of zero).

The second order information criterion, often called AICc, takes into account sample size by, essentially, increasing the relative penalty for model complexity with small data sets. It is defined as

$$\text{AICc} = -2\ (\ln(\text{likelihood})) + 2K * (n/(n\text{-}K\text{-}1)) \tag{2.7}$$

where n is the sample size. As n gets larger, AICc converges to AIC (n - K -1 -> n as n gets much bigger than K, and so ($n/(n - K - 1)$) approaches 1), and so there's really no harm in always using AICc regardless of sample size. In phylogenetics, defining "sample size" isn't always obvious. In model selection for tree inference, sample size often refers to the number of sites. In model selection in comparative methods, sample size often refers to the number of taxa.

Akaike weights can be used in model averaging. They represent the relative likelihood of a model. To calculate them, for each model first calculate the relative likelihood of the model, which is just exp(-0.5 * ΔAIC score for that model). The Akaike weight for a model is this value divided by the sum of these values across all models (Burnham and Anderson, 2002).