CHAPTER 2

METHODOLOGY

This chapter describes the methods used in the study. These methods include the following components.

(a) the study design;

(b) the data management;

(c) the methods used for the statistical analysis.

1. Study Design

A cross-sectional study design is used, based on (a) rainfall as recorded daily at three stations around the Bay, and (b) salinity as recorded from fourteen stations around the Bay. The period of interest is February 1995 to 30 August 1996 which is the latest recorded data. The outcome is salinity and the determinant is rainfall. The population may be taken as the totality of surface salinity and accumulated rainfall in the region enclosed by Pattani Bay during this period. The sample comprises the daily rainfall collected at three stations and salinity measurements collected at fourteen stations around the Bay on 25 selected days during this period.

The data on the rainfall were past records between January 1995 and September 1996 at 3 stations: (1) the Science Building of Prince of Songkla University at Pattani Campus (by Haroom Heamsuri in the Department of Science), (2) Yaring River mouth (by postman), and (3) Laem Tachi (by Lighthouse, Laem Tachi, Naval Hydrographic).

The data on salinity over a period between February 1995 and September 1996 were collected at 14 stations by the Pattani Coastal Aquaculture Station at Yaring as follows (see Figure 2).

(1) Dato (2) Yaring River mouth

(3) Middle of the Pattani Bay (4) Parae

(5) Cockle bed
(6) Tanyong Lulo
(7) Prince of Songkla University
(8) Talo Samilae
(9) Laem Nok
(10) Industry Zone
(11) Barn Num
(12) Budi
(13) Pattani River mouth
(14) Bana.

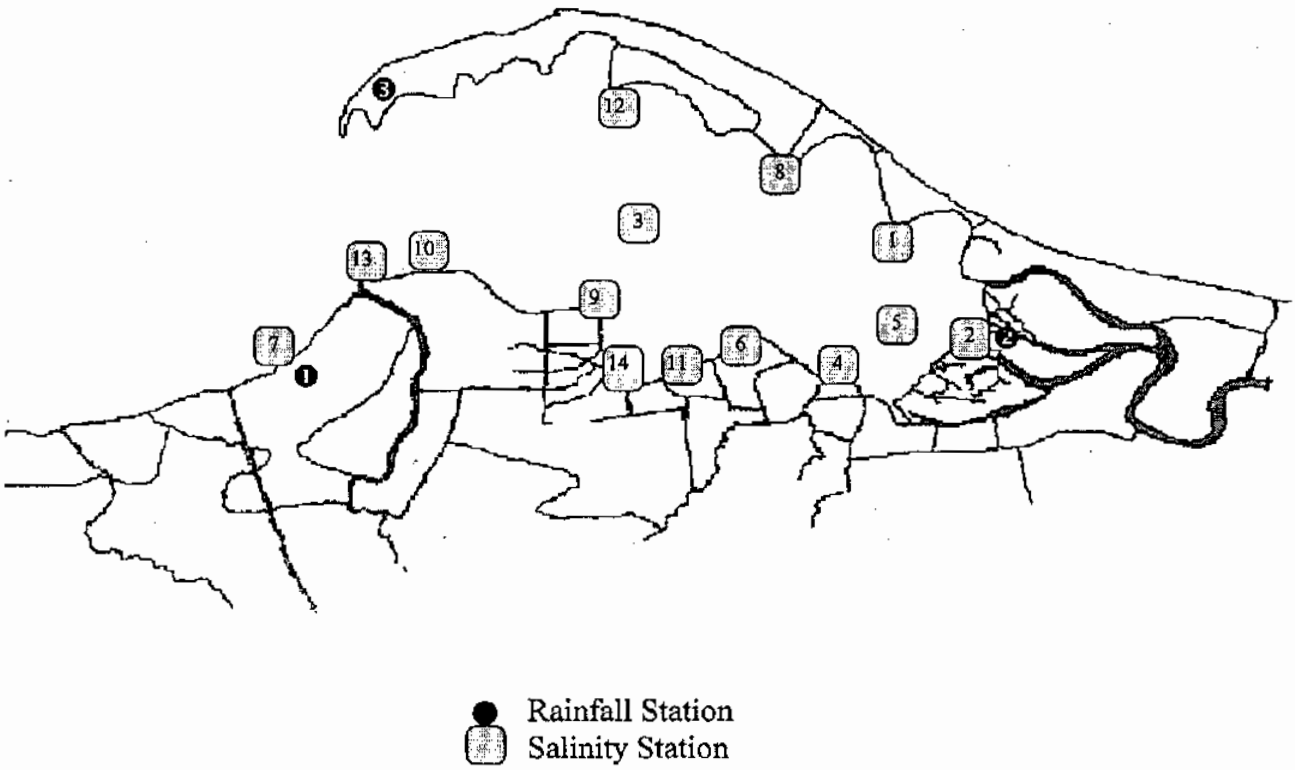

Figure 2: Locations of Stations Recording Rainfall and Salinity

2. Data Management

The raw data of rainfall and salinity are kept in Microsoft Access files (see Appendix), with one database containing the rainfall data and another containing the salinity.

2.1 Rainfall 1995-1996

Rainfall are recorded daily at three locations around the Bay: Prince of

Songkla University (by Haroom Heamsuri in the Department of Science), Yaring (by postman), and the meteorological station at Laem Tachi (by Lighthouse, Laem Tachi, Naval Hydrographic) on the Pattani Bay spit. These data are stored in the database pbrf.mdb, (see Choonpradub and McNeil, 1997), which has three base tables, station, dates and rainfall, containing the rainfall at each station on all days when rainfall occurred at one or more station. The relationships between stations, collecting dates and rainfall data are shown in Figure 3.

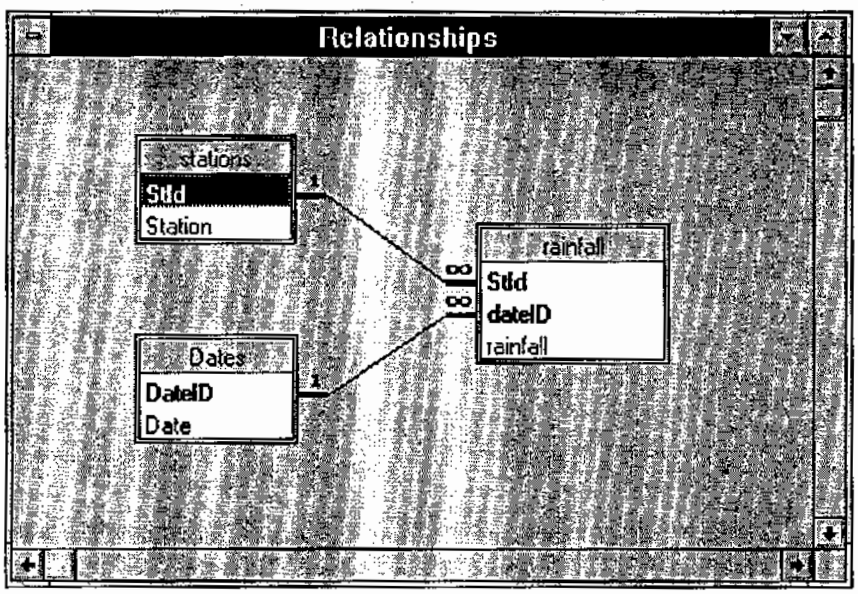

Figure 3: Relationships between Tables in pbrf.mdb Database

2.2 Salinity Collected by Pattani Coastal Aquaculture Station

Water samples from 14 locations around the Bay were measured at intervals from two to six weeks during 1995 and 1996 by the Pattani Coastal Aquaculture Station at Yaring. These data are stored in four tables in the database pbfo1.mdb, as shown in the relationships diagram in Figure 4.

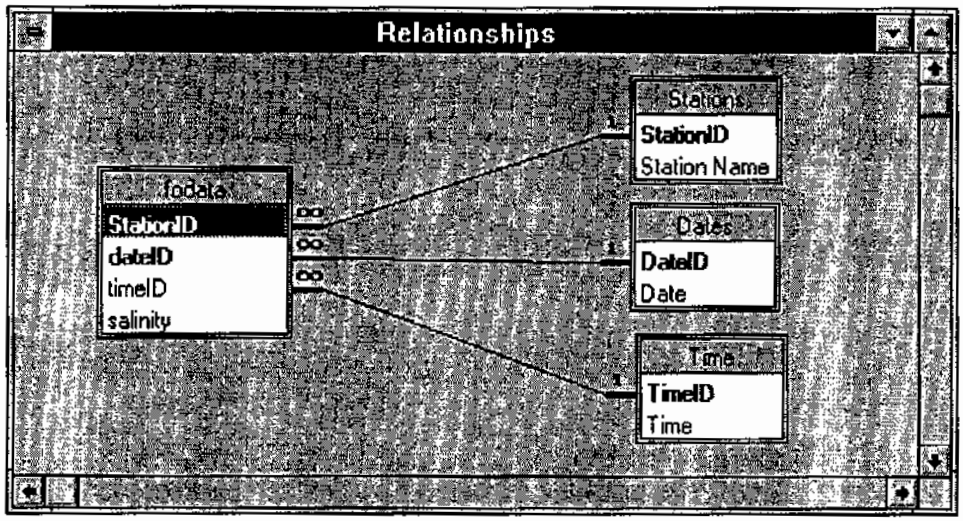

Figure 4: Relationships between Tables in pbfo1.mdb Database

3. Method of Data Analysis

The analysis is presented in the following steps.

3.1 Smoothing Technique

The rainfall data may be analyzed using smoothing techniques such as exponentially weighted averaging. An exponentially weighted accumulation of rainfall at times when salinity is recorded was calculated using a discount factor ρ.

$$sr_t = \sum_{j=0}^{50} \rho^j r_{t-j} \qquad (2)$$

where sr_t is an exponentially weighted accumulation of rainfall

t is a date when salinity is recorded

r_t is the rainfall on day t

j is range of time that rainfall is recorded corresponding to the data when salinity is recorded

The statistical methods used for the data analysis are based on regression analysis. The constant ρ is determined to maximise the correlation between the salinity and the accumulated rainfall (see Appendix, the computer program in Matlab for smoothed technique).

3.2 Analysis of Variance

3.2.1 One-way Analysis of Variance

One-way Analysis of Variance is the method used for the analysis of data in which the outcome is continuous and the determinant is categorical. The null hypothesis states that the samples have arisen from the same population. This null hypothesis can be tested by computing a statistic called the *F*-statistic and comparing it with an appropriate distribution to get a *p*-value. Supposethat there are n_j observation in sample *j*, denoted by y_{ij} for I=1,2, ... , n_j. The *F*-statistic is defined as follows (McNeil, 1996, page 67).

$$F = \frac{(s_0 - s_1)/(c-1)}{s_1/(n-c)} \qquad (3)$$

where

$$S_0 = \sum_{j=1}^{c} \sum_{i=1}^{nj} (y_{ij} - \bar{y})^2$$

$$S_1 = \sum_{j=1}^{c} \sum_{i=1}^{nj} (y_{ij} - \bar{y}_j)^2$$

and

$$\bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{nj} y_{ij}$$

$$\bar{y} = \frac{1}{n} \sum_{j=1}^{c} \sum_{i=1}^{nj} y_{ij}$$

$$n = \sum_{j=1}^{c} n_j$$

where S_0 is the sum of squares of the data after subtracting their overall mean, while S_1 is the sum of squares of the residuals obtained by subtracting each sample mean. If the population means are the same the numerator and the denominator in the *F*-

statistic are independent estimates of square of the population standard deviation and the *p*-value is the area in the tail of the F-distribution with *c*-1 and *n-c* degrees of freedom. The statistical assumptions are that the data arise from normally distributed population.

The aim of one-way analysis for salinity is to compare the mean of salinity in each station: n is the number of data collected for 25 days and c is the number of station such as Dato, Yaring River mouth, Middle of the Pattani Bay, Parae, Cockle Bed, Tanyong Lulo, Prince of Songkla University, Talo Samilae, Laem Nok, Industry Zone, Barn Num, Budi, Pattani River Mouth and Bana.

The aim of one-way analysis for rainfall is to compare the mean of rainfall in each station: n is the number of data collected for 25 days and c is the number of station such as Prince of Songkla University, Yaring River mouth and Laem Tachi.

3.2.2 Two-way Analysis of Variance

The response variable is classified by each of the two factors (station and day). There is at most one observation for each combination of these factors. Thus y_{ij} denotes an observation at station i on day j.

Two-way analysis of variance is used to compare the means of variables (rainfall and salinity) between days, and to compare the means of variables (rainfall and salinity) between stations. The p-value is based on an *F*-statistic defined as follows (McNeil, 1996, page 73).

$$F = \frac{(s_2 - s_{12})/(c-1)}{s_{12}/(n-c-r+1)} \qquad (4)$$

where

$$S_2 = \sum_{j=1}^{c}\sum_{i=1}^{r}(y_{ij} - \bar{y}_i)^2$$

$$S_{12} = \sum_{j=1}^{c}\sum_{i=1}^{r}(y_{ij} - \bar{y}_i - \bar{y}_j + \bar{y})^2$$

and

$$\bar{y}_i = \frac{1}{c}\sum_{j=1}^{c} y_{ij}$$

$$\bar{y}_j = \frac{1}{r}\sum_{i=1}^{r} y_{ij}$$

$$\bar{y} = \frac{1}{rc}\sum_{j=1}^{c}\sum_{i=1}^{r} y_{ij}$$

These formulas assume that there is exactly one observation in each of the data table. If this assumption is not met, some adjustment is needed, and the E-M algorithm is used (Dempster et. l, 1977, page 1-38). The assumptions required for two-way analysis of variance are as follows:

(1) the errors have constant variation

(2) the errors are independent and normally distributed

If assumption (1) is not satisfied it may be necessary to transform the data before the analysis. If assumption (2) is not met it may be necessary to remove outliers.

3.3 Regression Analysis

The multiple regression model is

$$y = \beta_0 + \beta_1 x_1 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i \ldots + \beta_p x_p \qquad (5)$$

where y is the outcome

β_0 is a constant

β_i is a set of parameters (i = 1 to p)

x_i is a set of determinants (i = 1 to p)

The model is fitted to data using least squares (see, for example, Kleinbaum et al, 1998). Linear regression analysis is used to find the relationship between rainfall and salinity, giving estimated values $b_0, b_1, \ldots, b_p$ for the coefficients. There are three assumptions for linear regression analysis, namely

(1) the association is linear,

(2) the variability of the errors (in the outcome variable) is uniform, and

(3) these errors are normally distributed.

If assumption (1) and/or (2) are not met, a transformation of the data may be appropriate.

The multiple regression model may be used to *adjust* an outcome variable for the effects of one or more determinants, using the formula

$$y^* = y - \{b_1(x_1 - \bar{x}_1) + b_2(x_2 - \bar{x}_2) + \ldots + b_p(x_p - \bar{x}_p) \qquad (6)$$

In this formula y^* is the adjusted value of y. This method is often used by econometricians for seasonally adjusting interest rates and unemployment statistics.