CHAPTER 2

METHODOLOGY

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

(a) study design and data selection

(b) graphical methods

(c) statistical methods

1. Study Design and Data Selection

1.1 Study Design

A cohort study design is used, based on the measurement of hydrographic data at specified locations and variables around the Bay. The period covered is 7 February 1995 to 29 August 1996. The determinants are stations and days, and the multivariate outcomes are temperature, salinity, transparency, alkalinity, oxygen, pH, nitrite, nitrate, phosphate and silicate concentrations.

1.2 Data Selection

The population comprises the levels of the specified hydrographic variables at specified locations in Pattani Bay on specified days during 1995 and 1996. The data for this analysis were obtained from 14 locations around the Bay and measured at intervals from two to six weeks during 1995 and 1996 (7 February 1995-29 August 1996) by the Pattani Coastal Aquaculture Station. The sampling locations are as follows, and are shown in Figure 1.

1. Dato

2. Yaring River mouth

3. Middle of the Pattani Bay

4. Parae

5. Cockle Bed

6. Tunyong Lulo

7. Prince of Songkla University

8. Talo Samilae

9. Laem Nok

10. Industry Estate

11. Bam Num

12. Budi
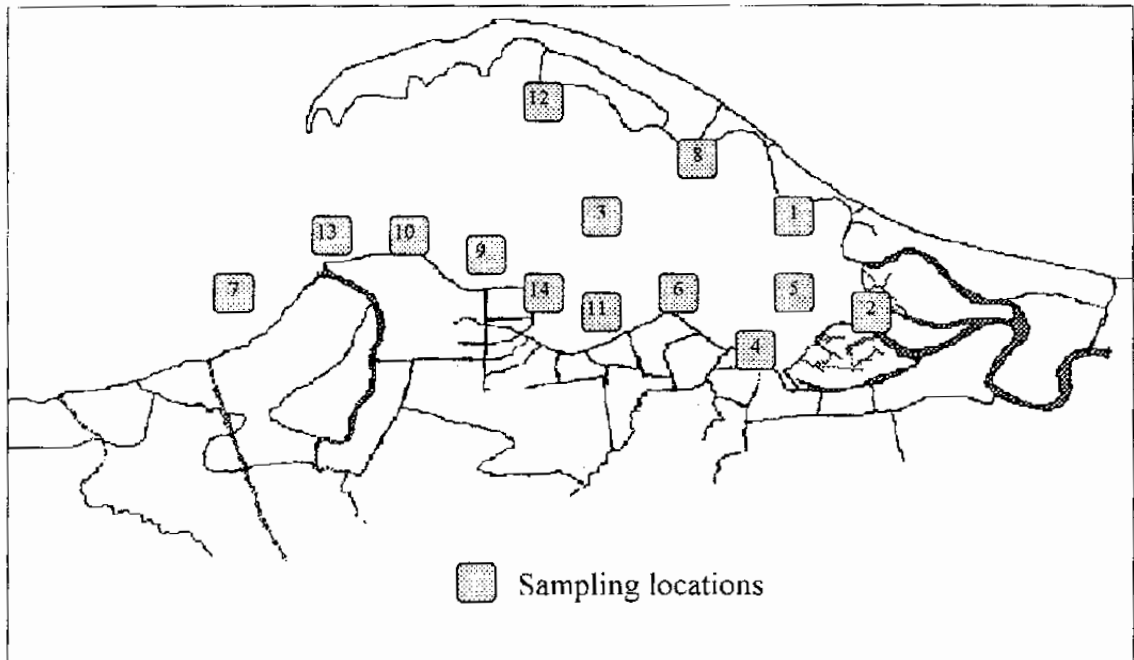
13. Pattani River mouth

14. Bana

Figure 1: A map of Pattani Bay showing the sampling locations

The data are stored in four tables in the database *pbfo.mdb*, as shown in the relationships diagram in Figure 2.
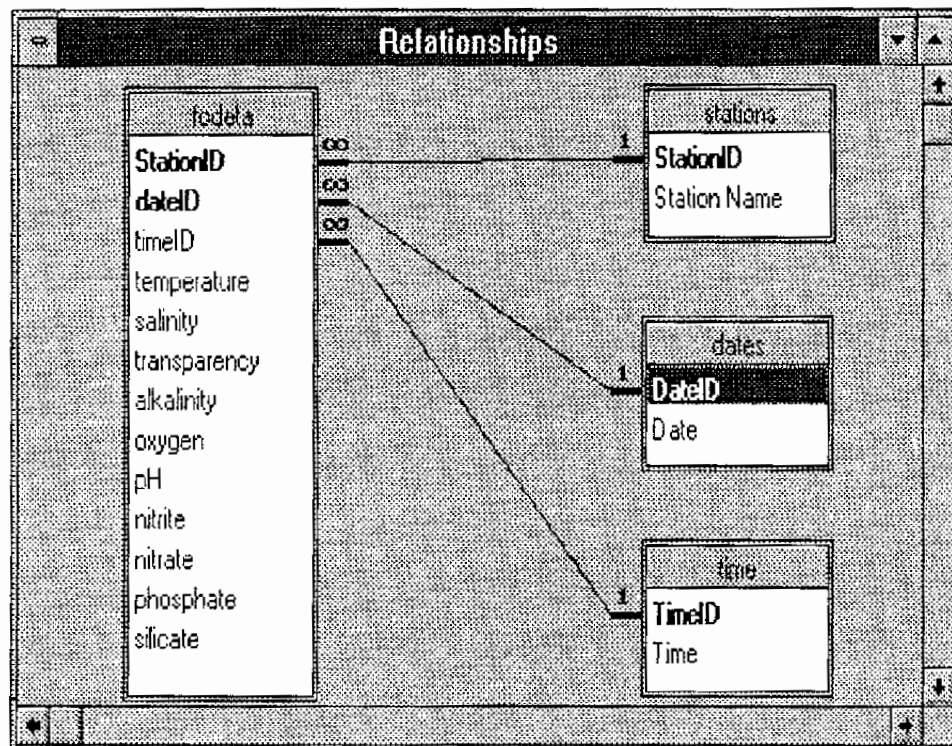

Figure 2: The relationships diagram in the database pbfo.mdb

2. Graphical methods

The graphical methods are presented in the following steps.

2.1 Histograms and statistics of raw data for all variables.

2.2 Histograms and statistics for variables requiring transformation using
Logarithms and square roots.

2.3 Time series plots of the variables.

2.4 Two-way analysis of variance of the variables.

2.5 Range plots of expected values.

3. Statistical methods

The statistical methods used for the data analysis are two-way analysis of variance and regression analysis. These methods may be described as follows.

3.1 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 between days, and to compare the means of variables between stations. The p-value is based on an F-statistic defined as follows (McNeil, 1996, page 73).

$$F = \frac{(s_1 - s_{12})/(c-1)}{s_{12}/(n-c-r+1)}$$

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 all of the data table. If this assumption is not met, some adjustment is needed, and the E-M algorithm (Dempster et. al, 1977, page 1-38) is used. 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.

3.2 Regression analysis

In regression analysis (see for example, Kleinbaum et al, 1998, page 111-135), a continuous outcome variable y is modeled as a linear combination of determinants x_1, x_2, x_3, ... x_p as follows

$$y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j \qquad (2.1)$$

Two-way analysis of variance can be carried out using regression analysis, by taking the βs as two sets of indicator variables corresponding to each factor (see, for example, Lunn and McNeil, 1992, page 171)

Linear regression analysis may be used to find the relationship between hydrographic data and stations and days. There are three assumptions for linear regression analysis, as follows.

(1) the association is linear

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

(3) these errors are normally distributed.

Since the determinants in Equation 2.1 are indicator variables, the linearity assumption is automatically satisfied.