Chapter 2

Methodology

This chapter describes the source of data and summarizes the graphical and statistical methods used for analyzing the data.

2.1 Data Managements

The datasets were loaded into a Microsoft Excel spreadsheet file and exported to WebStat (a set of programs for graphical and statistical analysis of data stored in an SQL database, written in HTML and VBScript) for cleaning. Then they were restructured into a simple format suitable for using the R program (Murrell, 2006; R Development Core Team, 2008) for further graphical displays, map creations, and statistical analysis.

2.2 Phytoplankton abundance

2.2.1 Data source

Phytoplankton were measured at ten sampling stations located in the Na Thap River, labeled 1-10 (Figure 2.1), at bimonthly intervals from June 2005 to December 2007. Sample transects were located, with three replicates for each stations. Phytoplankton were collected by water filtration: 100 liters was filtered by 20 micrometers plankton nets, collected by 20 and 69 micrometers plankton nets which were set up with a flow meter in front of the plankton net. Then they were fixed with 5% formalin for later

identification and calculation of density of phytoplankton in cells m^{-3} (Smith, 1950; Prescott, 1962; Wongrat, 1995; Wongrat, 1998).

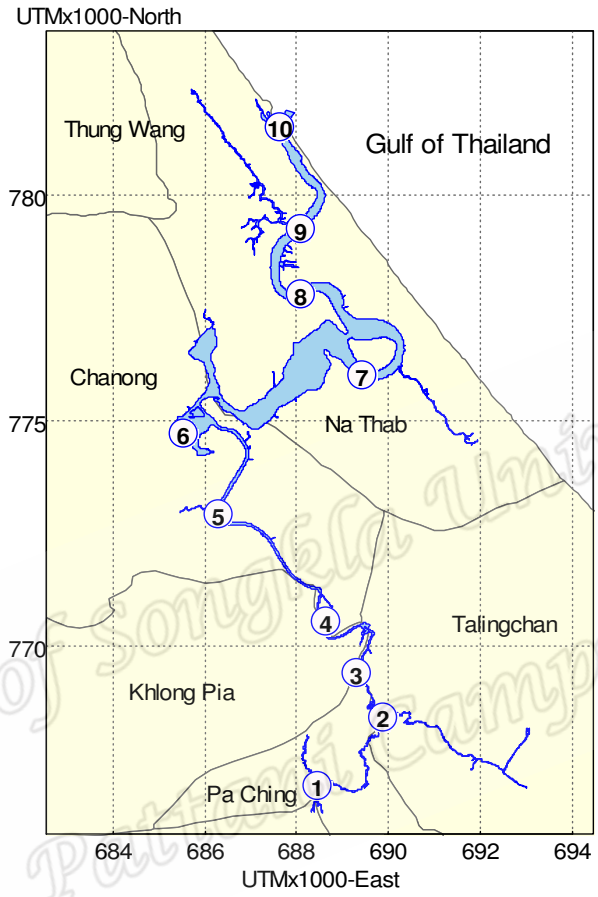

Figure 2.1: The Na Thap River and sampling stations (labeled 1-10)

Water quality parameters were investigated at similar intervals and sampling stations as were phytoplankton measured. Water transparency (Tran), salinity (Sal), turbidity (Turb), pH of water (pH) and dissolved oxygen (DO) were determined in the field (insitu determination) whereas measurement of biological oxygen demand (BOD), sulfate (SO_4), oil and grease (oilG), nitrate-nitrogen (NO_3-N), phosphate-phosphorus (PO_4-P), ammonia-nitrogen (NH_3-N), total iron (Iron), total coliform bacteria (coliT), total fecal coliform bacteria (coliF), cadmium, copper were exsitu determination.

Water samples were analyzed by applying the standard method (APHA, AWWA and WEF, 1998) both in field work and laboratory.

Table 2.1: Sampling stations, location and description of each station in Na Thap River

Station	Location	Description
1: Klong Pho Ma	Moo 1, Ban Pa Ching, Pa Ching sub-district.	Initial section, in Klong Pho Ma
2: Inflow pump	Moo 1, Ban Kok Muang, Pa Ching sub-district	Near inflow pumping house of Chana Thermal Power Plant in initial section
3: Outflow pump	Moo 6, Ban Kuan Hao Chang, Klong Pia sub-district	Near cooling water outflow of Chana Thermal Power Plant in initial section
4: Kuan Hao Chang	Moo 6, Ban Kuan Hao Chang, Klong Pia sub-district	At a concrete bridge, far from station in initial section 3 about 2.6 km.
5: Tha Klong Cha Nong	Moo 6, Ban Tha Klong, Cha Nong sub-district	At a concrete bridge, far from station 4 in middle section about 3.9 km.
6: Thung Kuad	Moo 7, Ban Thung Kuad, Cha Nong sub-district	At the wooden bridge for rafting, far from station 5 in middle section about 3 km.
7: Ma Ngon	Moo 5, Ban Ma Ngon, Na Thap sub-district	Directly opposite Ban Khu Nam Rob and Ban Na Sameian, Na Thap sub-district, far from station 6 in middle section about 5.5 km.
8: Tha Klong	Moo 4, Ban Tha Klong, Na Thap sub-district	At a concrete bridge, far from station 7 in estuarine section about 5 km.
9: Klong Kha	Moo 1, Ban Klong Kha, Na Thap sub-district	Mouth of Klong Kha stream, far from station 8 in estuarine section about 2.5 km.
10: Pak Bang	Moo 2, Ban Pak Bang Na Thap, Na Thap sub-district	Mouth of Na Thap River connected to the sea, far from station 9 in estuarine section about 4 km.

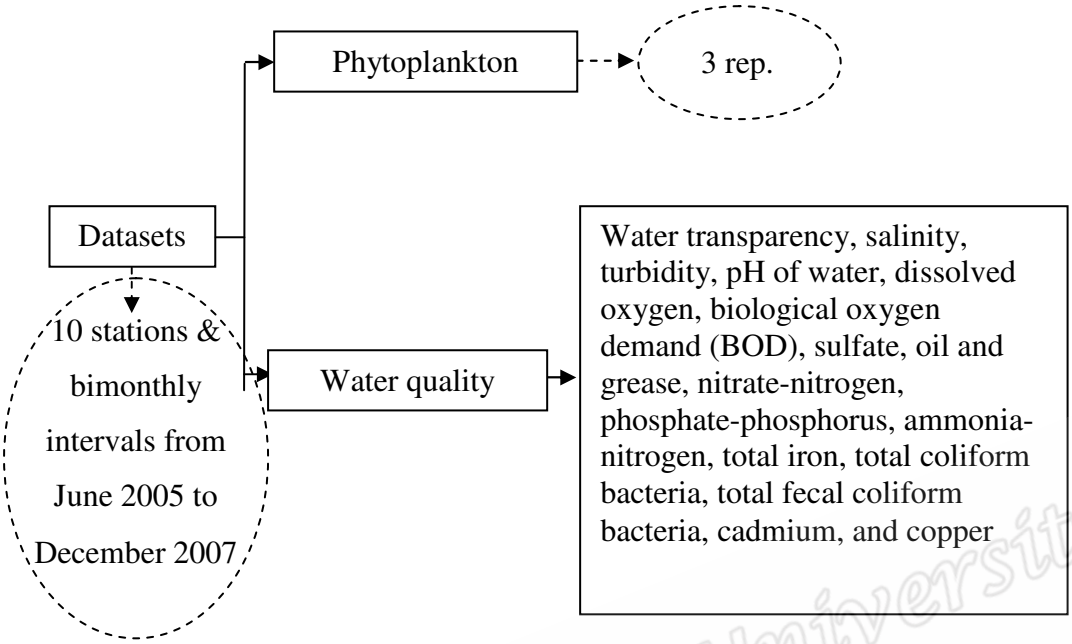

Figure 2.2: Dataset of the study of Phytoplankton abundance

2.2.2 Steps for model development

Figure 2.3 shows the steps for model development to assess phytoplankton distributions. The ten sampling stations and sixteen bimonthly periods were combined into one hundred forty four station-month combinations. The response variables were the log-transformed densities of 31 selected genera with greater than 24% occurrence. The predictors consisted of five environmental factors and four unique variable derived from factor analysis. Finally, to analyse and describe the relationships between these variables, a multivariate multiple regression model was used to relate these multiple responses to the multiple predictors.

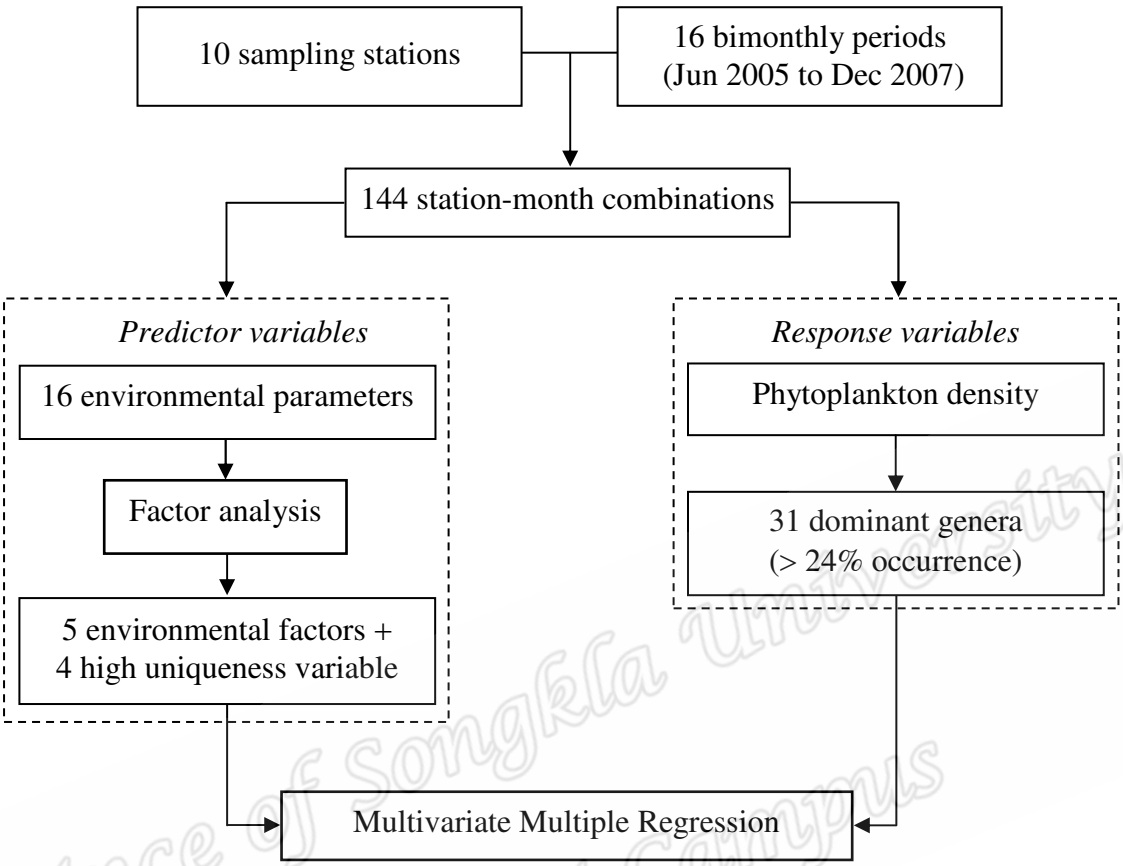

Figure 2.3: Steps for development of model to assess phytoplankton density

2.2.3 Statistical methods

The statistical methods for the analysis of phytoplankton abundances were based on a multivariate multiple regression model involving factor analysis, which was used to define the factors in the environment used to link phytoplankton community structure and environmental predictors.

2.2.3.1 Multivariate multiple regression

The multivariate multiple regression model was used to evaluate the effects of multiple predictor variables (the five environmental factors and the four high

uniqueness environmental parameters) on multiple response variables (the densities of the thirty one genera of phytoplankton), both the predictors and the response variables were observed at the one hundred forty four occasions. The multivariate multiple regression model (Mardia et al., 1979) is expressed in a matrix form, that is,

$$\mathbf{Y}_{(n \times p)} = \mathbf{X}_{(n \times q)} \mathbf{B}_{(q \times p)} + \mathbf{E}_{(n \times p)} \tag{2.1}$$

In this formulation $\mathbf{Y}_{(n \times p)}$ is an observed matrix of p response variables on each of n occasions, $\mathbf{X}_{(n \times q)}$ is the matrix of q predictors (including a vector of 1s) in columns and n occasions in rows, $\mathbf{B}_{(q \times p)}$ contains the regression coefficients (including the intercept terms), and $\mathbf{E}_{(n \times p)}$ is a matrix of unobserved random errors with mean zero and common covariance matrix $\mathbf{\Sigma}$. Ordinary (univariate) multiple regression arises as the special case when $p = 1$. If $q - 1$ environmental predictors $f_i^{(k)}$ ($k = 1, 2, \ldots, q - 1$) are available, the predict model for outcome j occasion i model may be expressed as

$$y_{ij} = \mu_j + \sum_{k=1}^{p} \beta_j^{(k)} f_i^{(k)} + z_{ij} \,, \tag{2.2}$$

where y_{ij} is the observed abundance for genera j on occasion i, μ_j is the mean abundance associated with genera j, $\beta_j^{(k)}$ is the effect of environmental variable k on genera j, and z_{ij} are the random errors.

The model fit may be assessed by plotting the residuals against normal quantiles (Venables and Ripley, 2002), and also by using the set of r-squared values for the response variables to see how much of the variation in each is accounted for by the model.

The method also provides standard errors for each of the $p \times q$ regression coefficients thus providing p-values for testing their statistical significance after appropriate allowance for multiple hypothesis testing. The multivariate analysis of variance (MANOVA) decomposition is also used to assess the overall association between each environmental predictor and the set of outcomes by the likelihood ratio, Pillai's trace criterion (Olson, 1976; Johnson and Wichern, 1998).

2.2.3.2 Factor analysis

Factor analysis is a mathematical model that tries to explain the correlation between a large set of variables in terms of a small number of underlying factors. A major assumption of the analysis is that it is not possible to observe these factors directly: the variables depend upon the factors but are also subject to random errors (Mardia et al., 1979).

In the first study, factor analysis is performed on the environmental variables with the aim of substantially reducing correlations between them that could mask their associations with the outcome variables. Each factor identifies correlated groups of variables. Ideally each group (which must contain at least two variables to contribute to the factor analysis) contains variables with small correlations with variables in other groups. To achieve this, any variable uncorrelated with all other variables is omitted from the factor analysis. Each factor comprises weighted linear combinations of the variables, and these factors are rotated to maximize the weights of variables within the factor group and minimize the weights of variables outside the group. The resulting weights are called "loadings". Variables omitted from the factor analysis due

to low correlation with all other variables (high "uniqueness") are treated as separate predictors, so predictors include single variables as well as factors.

The number of factors selected was based on obtaining an acceptable statistical fit using the chi-squared test, and these factors were fitted using maximum likelihood with promax rotation in preference to varimax, which requires the rotation to be orthogonal (Browne, 2001; Abdi, 2003).

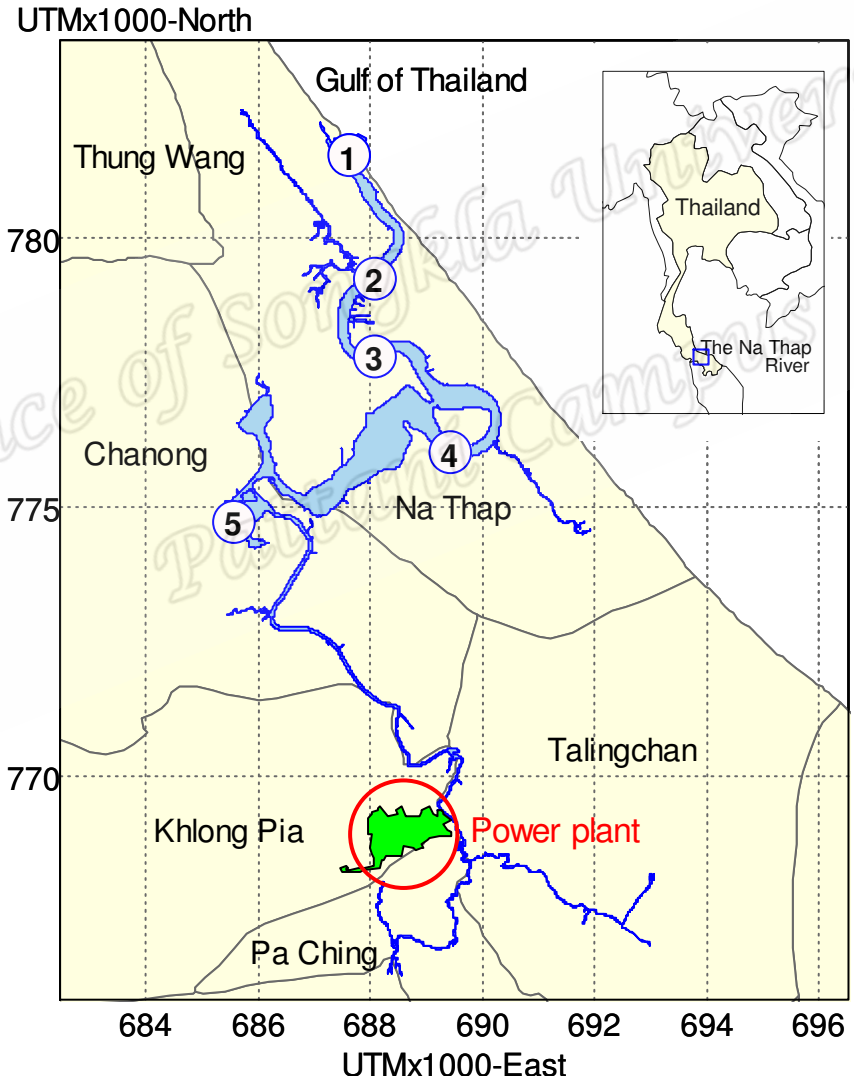

Figure 2.4: The Na Thap River and sampling stations (labeled 1-5)

2.3 Polychaeta abundance

2.3.1 Data source

Polychaetes were measured at five sampling stations located in the Na Thap River, on a bi-monthly basis during a period from June 2005 to May 2010 at 5 stations in the river, located in brackish and saline water (Figure 2.4). Sample transects were located, with three replicates for each stations. Polychaetes were collected by using Ekman dredge 15x15 cm^2 to sieve with 420 micrometre mesh size. The samples were stored in 5% formalin solution in the laboratory. Each species were counted and classified under a stereoscope (Fitter and Manuel, 1986).

Table 2.2: Sampling stations, location and description of each station in Na Thap River

Station	Location	Description
1: Pak Bang	Moo 2, Ban Pak Bang Na Thap, Na Thap sub-district	Mouth of Na Thap River connected to the sea, far from station 9 in estuarine section about 4 km.
2: Klong Kha	Moo 1, Ban Klong Kha, Na Thap sub-district	Mouth of Klong Kha stream, far from station 8 in estuarine section about 2.5 km.
3: Tha Klong	Moo 4, Ban Tha Klong, Na Thap sub-district	At a concrete bridge, far from station 7 in estuarine section about 5 km.
4: Ma Ngon	Moo 5, Ban Ma Ngon, Na Thap sub-district	Directly opposite Ban Khu Nam Rob and Ban Na Sameian, Na Thap sub-district, far from station 6 in middle section about 5.5 km.
5: Thung Kuad	Moo 7, Ban Thung Kuad, Cha Nong sub-district	At the wooden bridge for rafting, far from station 5 in middle section about 3 km.

Water quality parameters were investigated at similar intervals and sampling stations as were Polychaetes measured. Water samples from each site were collected by using

1 liter plastic bottles in size. Salinity was determined in the field (insitu determination). Water samples were analyzed by applying the standard method (APHA, AWWA and WEF, 1998).

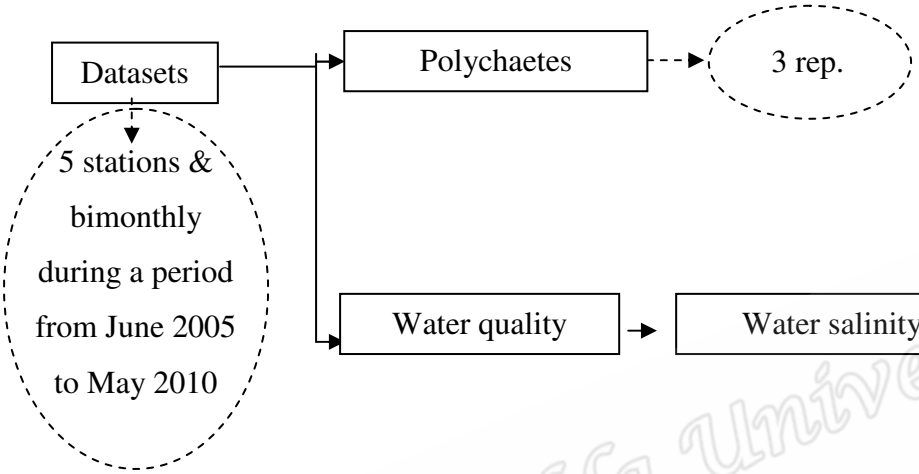

Figure 2.5: Dataset of the study of Polychaetes abundance

2.3.2 Path diagram and variables

The path diagram of this study is shown in Figure 2.6. This study carried out statistical analyses for investigating the Polychaeta organism densities with the determinant variables comprising month and water salinity.

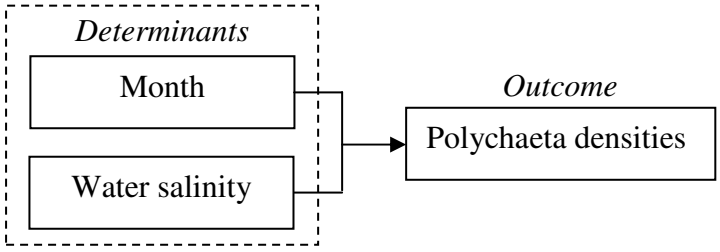

Figure 2.6: Path diagram showing roles of variables

2.3.3 Statistical methods

2.3.3.1 Multiple linear regression

Linear regression analysis was used to analyze data in which both the determinants and the outcome are continuous variables. In the simplest case involving a single determinant, it can summaries the data in the scatter plot by fitting a straight line. In conventional statistical analysis the line fitted is the least squares line, which minimizes the distances of the points to the line, measured in the vertical direction. If there is more than one determinant, the method generalizes to multiple linear regression, in which the regression line extends to the multiple linear relation represented as (McNeil, 1998).

$$Y = \beta_0 + \sum \beta_i x_i + \varepsilon \tag{2.3}$$

where Y is the outcome variable, β_0 is a constant, $\{\beta_i\}$ is a set of parameters ($i = 1$ to p, the number of determinants), and $\{x_i\}$ is a set of determinants ($i = 1$ to p). The model is fitted to data using least squares, which minimizes the sum of squares of the residuals.

Linear regression analysis resets on three assumptions as follows. First, the association is linear, the variability of the error (in the outcome variable) is uniform and last, these errors are normally distributed. If these assumptions are not met, a transformation of the data may be appropriate. Linear regression analysis may also be used when one or more of the determinants is categorical. In this case the categorical determinant is broken down into c-1 separate binary determinants, where c is the number of categories. The omitted category is taken as the baseline or referent category.

Equation (2.3) generalizes straightforwardly to any specified number of determinants, including categorical determinants that are treated as factors in the linear regression model. In such cases the parameters of interest are *differences* between parameters specifying the factor levels and their overall mean, and *sum* contrasts are needed to obtain appropriate confidence intervals for these differences.