Chapter 2

Methodology

The studies examined two datasets from coastal and estuarine environments:
(1) imposex prevalence in female gastropods in the Gulf of Thailand in 2006; and
(2) density of macrobenthic fauna in the Middle Songkhla Lake during 1998-1999.
Section 2.1 describes the methodology details for the prevalence of imposex (study 1). Section 2.2 describes the methodology details for the macrobenthic fauna abundances (study 2). Section 2.3 describes data management. Section 2.4 details the statistical methods used for the two studies.

All the graphical displays, map creations, statistical model fitting, and goodness-of-fit assessments were carried out using the R program (R development core team 2009, Venables and Ripley 2002, Murrell 2006).

2.1 Prevalence of imposex

The prevalence of imposex is studied by examining its distribution in various species of gastropods with the objective of determining its association with TBT pollution in the Gulf of Thailand. Gastropod samples were obtained from 56 sampling sites located in the Gulf over two periods in 2006 (June 8 to June 27; and September 19 to October 10).

2.1.1 Data source

Some specimens were hand-collected in the intertidal zone during low tide, but most were obtained from the by-catch of small commercial fishing boats and classified into

species immediately. A small sample of shells belonging to species that were not immediately recognised was preserved in alcohol for later identification. A part of the shell was removed by hammering to identify sex.

Males of gastropods can easily be identified by the presence of a penis and vas deferens. These organs are on the right side of the body near the right rhinophore and cannot be withdrawn into the body.

Females have an inconspicuous vaginal pore next to the anus. Thus the presence of a prominent penis is the crucial factor. However, due to TBT pollution females may currently show a small penis and a small vas deferens at the same external body position as in the males. Therefore, a clear internal characteristic has to be used and that was found in a preliminary study to be the presence or absence of a good recognizable capsule gland in females. In practice the internal examination was only necessary in the beginning of a series of a species. The female penis reaches lengths that are 10% or less of that of males of the same shell height. Presence or absence of a penis whether or not with a vas deferens in females was conclusive for the qualification imposex or no imposex. To avoid errors, we did not use an imposex index such as used by Mensink et al (2002), since sizes of the specimens and sizes and shapes of the male penis among the several species studied varied too much for correctly using the index on all species under local field conditions. Thus females with stage 1 and 2 were considered as showing no imposex.

2.1.2 Path diagram and research questions

The effect of TBT on the prevalence of imposex in the Gulf of Thailand could be addressed simply by examining the prevalence of imposex at different locations in the

Gulf, and seeing if higher prevalence was associated with TBT contamination. However, if the effect of TBT in causing imposex differs between gastropods species, any such association could be distorted by variation of species distribution with location, so it is necessary to take species into account when developing a contamination model.

Figure 2.1 shows the relevant variables as a path diagram. A total of 13 areas comprised the determinant while 16 species groups comprised the covariate. The outcome of interest is imposex (normal female or imposex female).

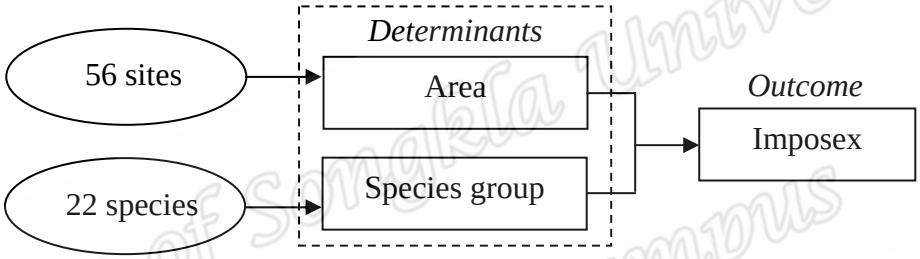

Figure 2.1: Path diagram showing roles of variables

Two research questions were considered, indicated by the dotted rectangles in the path diagram. In order of scientific importance, they are as follows.

(1) How does the prevalence of imposex in gastropods vary with area?

(2) How does the prevalence of imposex vary with species?

2.2 Macrobenthic fauna abundance

2.2.1 Data source

The datasets that were used for this work (Figure 2.2) consisted of biotic samples (macrobenthic fauna densities per square meter for each identifiable family) and other datasets containing information on environmental variables (water quality parameters

and sediment characteristics). All were collected from nine sampling stations located in the Middle Songkhla Lake, labeled 1-9 (Figure 2.3) at bimonthly intervals from April 1998 to February 1999. The details of each dataset are as follows:

(1) Macrobenthic fauna were collected via a Tamura's grab (0.05 m^2). The assemblage was conducted with eleven replications for each station. The samples were sieved consecutively through three orders of screen residue (5, 1, and 0.5 mm of sieve mesh size, respectively) and fixed in 10% Rose Bengal formalin solution for later taxa identification.

(2) Environmental variables (water depth, salinity, temperature, water pH, dissolved oxygen, total suspended solid, organic matter contents, organic carbon contents and total nitrogen contents, together with pH of sediment, and soil structure as sand, silt and clay percentages) were obtained in triplicate on the same occasions as the data of macrobenthic fauna.

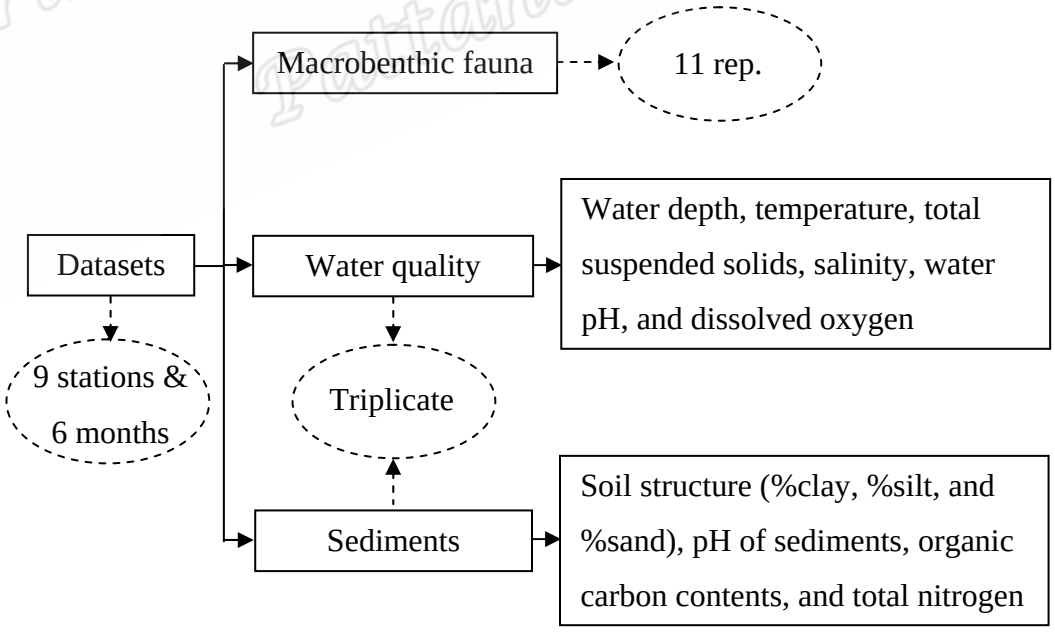

Figure 2.2: Dataset of the study of macrobenthic fauna variation

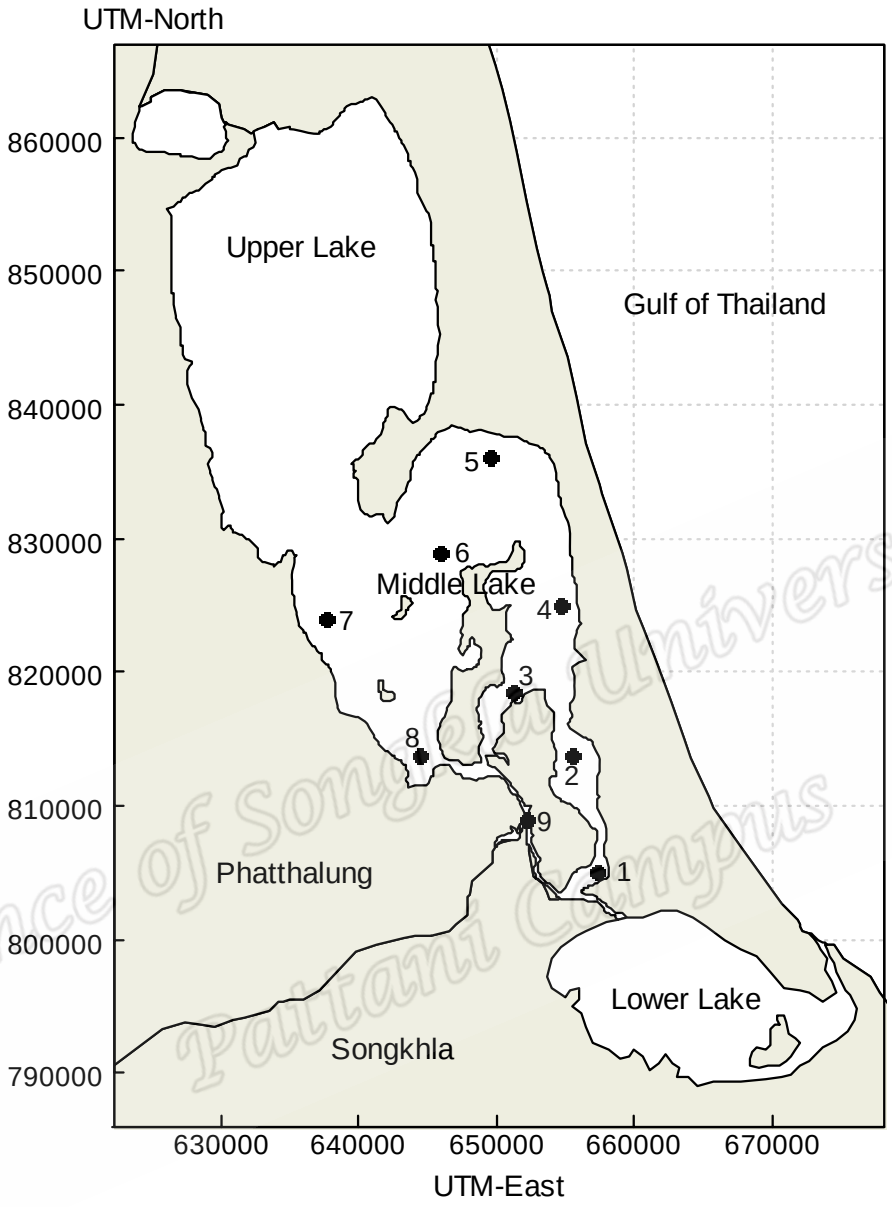

Figure 2.3: The Songkhla Lake and nine sampling stations (1: Ban Laem Chak, 2: Ban Koh Nang Khum, 3: Ban Nak Ka Rat, 4: Ban Koh Khob, 5: Ban Ta Kura, 6: Ban Laem Kruad, 7: Ban Hat Kai Tao, 8: Ban Ta Wa, and 9: Ban Bang Tan) within the Middle Songkhla Lake

2.2.2 Steps for model development

Figure 2.4 shows the steps for model development to assess macrobenthos distributions. The nine sampling stations and six bimonthly periods were combined into fifty four station-month combinations. The response variables were the log-transformed densities of 24 selected families with greater than 35% occurrence. The predictors consisted of three environmental factors and one 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.

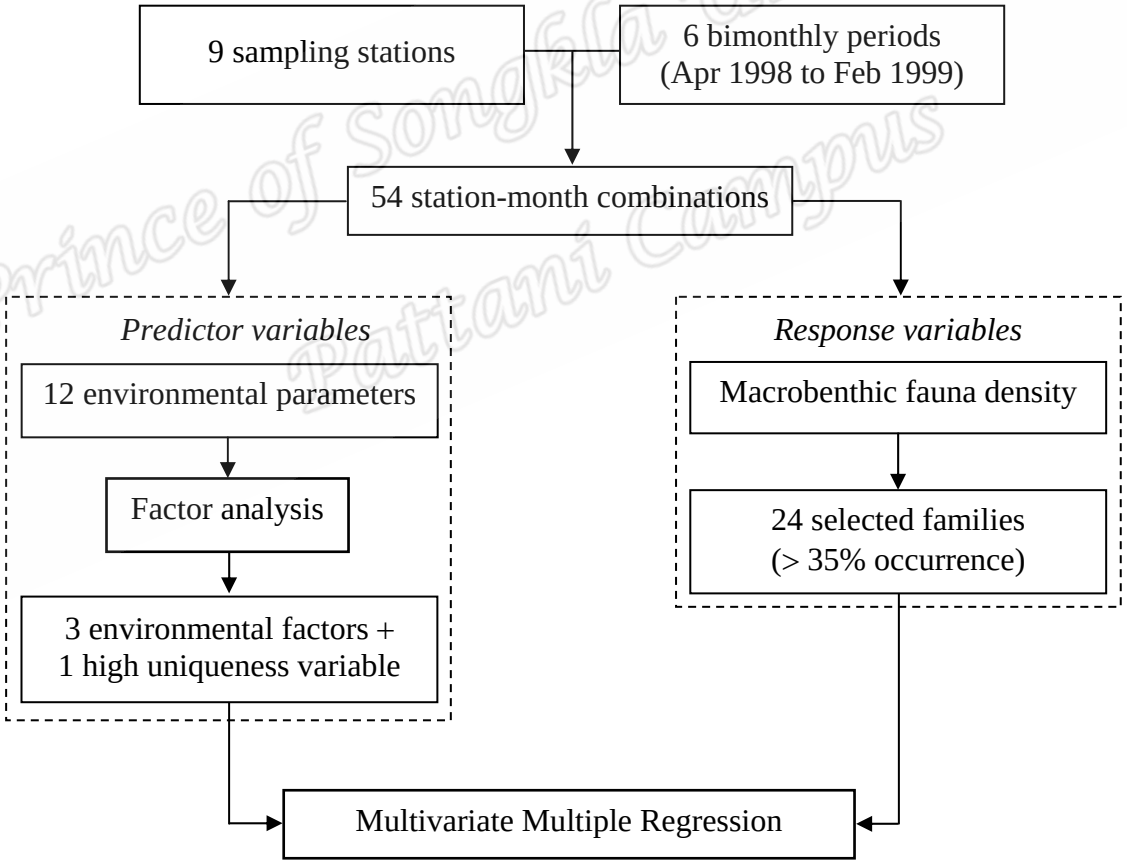

Figure 2.4: Steps for development of model to assess macrobenthic fauna variations

2.3 Data managements

Both 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 for further graphical displays, map creations, and statistical analysis.

2.4 Statistical methods

Several statistical methods were used in this thesis. Logistic regression modeling is well-suited to the analysis of the binary outcome in the gastropod dataset. The statistical methods for the analysis of macrobenthic fauna 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 benthic fauna community structure and environmental predictors.

2.4.1 Logistic regression

Logistic regression (Hosmer and Lemeshow 2000, Kleinbaum and Klein 2002) is a statistical method widely used to model the association between a binary outcome probability - the probability of a specific adverse outcome - and a set of fixed determinants. When the determinants are categorical factors, these factors can be structured as a multi-way contingency table of counts and the data for analysis comprise the proportions of adverse outcomes in the cells of this table.

In our first study (Publication 1), we investigated how imposex in female gastropods can be predicted by one or more predictor variables. A sample of 8,757 gastropods

belonging to 16 species groups that were sampled from 13 areas in the Gulf of Thailand in 2006 was considered. We employed logistic regression to model the effects of multiple determinants on the prevalence of imposex. If *p* is the prevalence of outcomes with a specific characteristic in a sample of size *n*, an asymptotically valid (for large *n*) formula for its standard error is

$$SE = \sqrt{\frac{p(1-p)}{n}}. \tag{2.1}$$

An asymptotically valid 95% confidence interval (95% CI) for the prevalence is thus given by

$$p - 1.96 \times SE,\ p + 1.96 \times SE\,. \tag{2.2}$$

When p_{ij} denotes the probability of outcome in taxa *j* on area *i*, the simplest such model takes the additive form

$$\ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = a_i + b_j\,. \tag{2.3}$$

Formula 2.3 can be inverted to give an expression for the prevalence *p* as

$$p_{ij} = \frac{1}{1+\exp(-a_i - b_j)}. \tag{2.4}$$

To avoid over specification of the parameters, one of the categories of the second determinant is taken as the reference and the corresponding parameter is zero. The model is fitted by maximizing the likelihood of the observed data given the parameters, providing standard errors for the fitted parameters. For each determinant, the method gives a *p*-value based on a chi-squared statistic for testing the null hypothesis that the outcome prevalence is the same for each of its component

categories. The method also gives separate *p*-values for comparing each parameter in the model with the mean (for the determinant of interest) or reference (for the covariate determinant).

Logistic regression provides a straightforward method for adjusting a prevalence that varies with a determinant of interest for the effect of a covariate determinant. To calculate the adjusted prevalence for category *i* of the determinant of interest, the term b_j in (2.4) is replaced by a constant b^*, that is,

$$p_i^* = \frac{1}{1+\exp\left(-a_i - b^*\right)}. \qquad (2.5)$$

The value of b^* in (2.5) is chosen to ensure that sum of the expected number of adverse outcomes is equal to the sum of the observed number, that is,

$$\sum p_i^* n_i = \sum p_i n_i, \qquad (2.6)$$

where n_i is the sample size in category *i* of the determinant of interest. This method extends straightforwardly to additional covariates.

2.4.2 Multivariate multiple regression

In the second study, the multivariate multiple regression model was used to evaluate the effects of multiple predictor variables (the three environmental factors and the unique environmental parameter) on multiple response variables (the densities of the twenty four families of macrobenthic fauna), both the predictors and the response variables were observed at the fifty four occasions. The multivariate multiple regression model (Mardia et al 1979) is expressed in a matrix form, that is,

$$\mathbf{Y}_{(n \times p)} = \mathbf{X}_{(n \times q)} \mathbf{B}_{(q \times p)} + \mathbf{E}_{(n \times p)} \qquad (2.7)$$

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

$$y_{ij} = \mu_j + \sum_{k=1}^{p} \beta_j^{(k)} f_i^{(k)} + z_{ij} \,, \tag{2.8}$$

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

2.4.3 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 our second 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).