Chapter 1

Introduction

1.1 Rationale for study

This thesis investigates and compares statistical methods for analyzing associations between biological outcomes and their environmental or demographic determinants, using data from recent studies undertaken in southern Thailand.

Scientific studies, particularly in biology, are often concerned with understanding the basis for associations between specified determinants and outcomes of interest. For example, the outcome in a study might be the occurrence of a disease such as pneumonia in a human population, for which the risk factors include demographic factors such as the subject's age, gender and location of residence and the period of data collection. As a further example, the outcome might be the abundance of specified families of organisms (such as small crustaceans) in the bed of a lake, for which the determinants of interest include the characteristics of the surrounding environment including water and sediment characteristics.

Studies such as these embrace a wide spectrum of scientific disciplines. The first example is of interest to epidemiologists, who are concerned with the causes, prevention, and treatment of diseases in human populations, whereas the second example is of interest to ecologists, who focus on the interactions between organisms and their environment. As a result, the statistical methods favoured in the literature for analyzing data from such

studies tend to be quite different. Biostatisticians usually start with the Poisson distribution to model incidence rates based on the observed numbers of disease cases, and arrive at more complex generalized linear models to incorporate demographic and other (including genetic, environmental, climatic, occupational, and behavioural) risk factors, and to cope with departures from the assumptions of independence, linearity, variance homogeneity and the specific distributions required by the simpler models. In contrast, quantitative ecologists have developed quite different statistical methods, including so-called *ordination* procedures aimed at clustering taxa in space and time, often based on their own measures of association (such as the *Bray-Curtis* similarity index). For more complex analyses, the dominant method used by ecological researchers for examining associations between taxa abundances and their environmental determinants is canonical correspondence analysis, but this method is rarely used by scientists in other disciplines. On the other hand, regression analysis and its various extensions to generalized linear models, generalized additive models, and generalized linear mixed models, although used extensively by biostatisticians, is not commonly used in ecological research, although there are exceptions such as studies by White and Bennetts 1996, Liang et al 2002, Warton and Hudson 2004, Venables and Dichmont 2004a, 2004b, Gray 2005, Sileshi 2008, Kundbuy et al 2010.

These considerations prompted us to ask why medical scientists investigating human diseases and ecological scientists investigating animal and plant abundances use such different methods, given that the corresponding data structures are essentially the same.

This similarity of data structure is based on the fact that in each situation the outcome data are observed in *cells* defined by specified locations and time periods, and the determinants of these outcomes are characteristics of the cells. Note that a "determinant" is often called a "risk factor" in medical studies.

In our first study the cell locations ranged over the 19 districts in Surat Thani province of Southern Thailand, their periods comprised the 36 quarters ranging from January-March 1999 to October-December 2007, and the other cell characteristics were taken as the age-group (less than 1 year or 1-4 years) and gender of the study subjects. The corresponding outcome of interest was the proportion of subjects in the districts diagnosed with pneumonia after visiting a hospital. Because pneumonia is a serious problem among young children the study was restricted to children under 5 years of age. The objective of the study was to examine how these determinants – district location, period (including the season of the year as well as any annual trend), age group, and gender – affect the pneumonia incidence rate. Such studies are useful because their conclusions form the basis for health planning. For example, if the study had found that male babies in urban areas were at substantially higher risk of contracting pneumonia, that group could be targeted for preventative strategies.

Our second study is a similar example of a medical study. The cells comprised 32 regions covering all 14 provinces of southern Thailand and the periods comprised the 24 quarters ranging from January-March 1999 to October-December 2004. The remaining cell characteristics were again taken as age-group and gender, but because the outcome was

the incidence of tuberculosis and this disease is a problem for all age groups, these groups were defined more widely as 0-24 years, 25-39 years, 40-59 years, and 60 or more.

The third study was a typical one in the field of ecology. In this case the cells comprised nine fixed locations in Middle of Songkhla Lake in southern Thailand and the periods were specified days in the months of April, June, August, October and December 1998, and February 1999. The remaining cell characteristics were 12 environmental variables measured on each of the six occasions at the nine stations, namely the depth, temperature, salinity, pH, dissolved oxygen, and suspended solids concentration of the water, and the pH, total nitrogen content, organic carbon content, and percentages of sand, silt and clay in the sediment. The outcome variables of interest for this study were the sediment densities of the 24 most commonly occurring macrobenthic organism families.

Note that the ecological study is more complicated than the disease incidence studies because its outcome is multivariate, comprising 24 variables corresponding to the families, whereas there is a single outcome variable – the disease incidence rate – for the medical studies. Thus the appropriate methodology for the medical studies is essentially a special case of that for the ecological study. In any case, it is possible to use the same general method for each study, and also for a very wide range of studies in other disciplines, and this is the methodological focus for our thesis.

It often happens that general methods are simpler and more elegant than methods that have developed with the confines of a single discipline. We also investigated the possibility of using simpler methods based on models for transformed incidence rates

with normally distributed errors, and this is another focus for our thesis. We thus report on the comparison of these simpler methods with those based on more complex generalized linear models, using data from the three studies.

1.2 Literature review

This review covers selected studies where the data structure is as outlined in the preceding section. These studies cover a wider range of application than diseases in humans and abundances of biological organisms, but are of interest to us because many of them involve data from Thailand. The methods include generalized linear models with logistic, Poisson and more general negative binomial distributions, zero-inflated extensions of these models, generalized estimating equation, simpler methods for log-transformed data, and canonical correspondence analysis. Given that the outcome variable was multivariate (comprising 24 macro-benthic fauna families), we also used *multivariate* multiple linear regression analysis for the ecological study. However, although this method has been available in the statistical literature for many years it is rarely used in epidemiological or ecological studies. The details of these methods are covered in Chapter 2.

Logistic regression

Logistic regression is the appropriate method for analyzing data from studies where the outcome variable of interest is a proportion.

Laeheem et al (2008, 2009) examined determinants that affect the propensity of students in Pattani primary schools to bully other children. These risk factors included school

rural/urban location, age group, gender, religion, family physical abuse history and television cartoon type preference of the subject. In this study logistic regression was used to model the risk factors, and it was found that children witnessing or experiencing physical abuse by a parent was associated with a fourfold risk increase after adjusting for the other factors, and watching action cartoons was also associated with a substantial increased risk.

Other applications of the logistic model in social sciences are a study of the trends of completion of secondary school education in districts of Pattani province (Thongchumnum and Choonpradub 2008), and a study of the effects of demographic factors (including secondary school completion) on employment in districts of Pattani and Songkhla provinces (Thongchumnum et al 2008). Each of these studies used data from the 2000 Population Census of Thailand. The first study found that educational completion rates had increased substantially in all areas over the last 30 years, but were still very low (less than 30%) among Muslim men and women in rural areas. However, the second study found that persons who had completed secondary education in Songkhla province were no better equipped for employment than those who had only completed primary school, and were actually less able to find jobs in Pattani province.

Anuntaseree et al (2008) used logistic regression to study risk factors for passive smoking in 4,245 Thai one-year-old infants sampled from all four regions of Thailand as well as Bangkok City. They found that environmental tobacco smoke exposure is common and is mainly due to the presence of a smoking father.

Based on a database of 25,829 singleton maternal births at Pattani Hospital during the period from October 1996 to September 2005, Rachatapantanakorn and Tongkumchum (2009a) applied logistic regression analysis to analyse the effects of demographic determinants on the incidence of caesarean births. They found that Islamic women were less likely to give birth by caesarean section and older mothers were more likely to do so. There was also an association between higher education and caesarian section. A further study Rachatapantanakorn and Tongkumchum (2009b) investigated risk factors for neonatal morbidity based on a database of 19,268 singleton maternal deliveries at Pattani Hospital during the same period. This study used logistic regression to adjust for demographic and pregnancy-history factors, and found that the Muslim women had higher neonatal morbidity risks, particularly those associated with severe pregnancy-induced hypertension, eclampsia and thick mecomium stain.

Prompted by a small-scale descriptive study carried out in 1996, Swennen et al (2009) further investigated the prevalence of imposex (a serious disease caused by tributyltin coating to prevent barnacles from attaching to ship hulls) among gastropods in the Gulf of Thailand ten years later. A total of 8,757 specimens from 22 species of five families were collected from 56 sites within 13 sampling areas. Since sensitivity to imposex infection varies substantially among different species and thus seriously biases the results, logistic regression was used to obtain unbiased estimates of the imposex incidence over the 13 areas of the Gulf and thus more accurately determined the effects due to shipping routes. The incidence was found to be higher near the shipping routes in

the east side of the Bight of Bangkok off Si Racha and Pattaya and in the southern part around Pattani.

Poisson regression

Poisson regression is the appropriate method for analyzing data from studies where the outcome variable of interest is a non-negative count.

Parodi and Bottarelli (2006) reported that Poisson regression has been widely used for analyzing the incidence and mortality of chronic diseases, particularly in cohort studies comparing exposed and unexposed individuals after adjusting for other factors.

Søyseth et al (2007) compared the incidence of lung cancer between 7,044 patients hospitalized for pneumonia and a reference population of size 81,373 in Akershus University Hospital in Norway. Poisson regression was used to fit a model to lung cancer incidence with smoking habits, emigration, and death status as predictors. They found that hospitalization for pneumonia was associated with the diagnosis of lung cancer.

Negative binomial regression

Negative binomial regression is often used to model count response data when Poisson regression is inappropriate due to the presence of greater variability in the data (i.e.overdispersion).

Lim and Choonpradub (2007) studied the regional and temporal pattern of death reported from HIV/AIDS and other infectious diseases, based on reported hospital cases from 14 provinces of southern Thailand during 1999-2004. A negative binomial regression model was used to fit mortality incidence. The in-hospital mortality rate for HIV/AIDS for

males was twice that rate for females and had a lower rate in the four border provinces of southern Thailand. For other infectious diseases, the in-hospital mortality rate for those aged 40 and over tended to increase over the period of study.

Sriwattanapongse (2008) reported the association between malaria incidence and a set of determinants comprising seasons, districts, and age groups from data collected in the North-western area of Thailand from 1999 to 2004, using generalized linear regression models. Negative binomial regression was appropriate to analyse such data with an application to forecasting districts and age groups in which malaria outbreaks are likely to occur.

Famoye et al (2004) considered a generalized Poisson regression model that, in contrast to the conventional negative binomial model, allows for under-dispersion as well as over-dispersion. They used this model to examine the effect of medication on accident risk in Alabama based on a sample of 901 drivers aged sixty-five or more, and found that drivers even when over-dispersion exists the conventional negative binomial model may not accurately portray the relationship between the variance and the mean. Jansakul and Hinde (2004) considered both the standard form of the negative binomial model with variance-mean ratio of the form $1+\lambda/\theta$ and an alternative model with ratio $1+\alpha$ for $\alpha \geq 0$ (which requires use of the Newton-Raphson algorithm to obtain maximum likelihood estimates.)

Zero-inflated count models

When count data have a large number of zeros it has been argued that special models are needed to account for such "excess" zeros. Lambert (1992) introduced the zero-inflated count model, which consists of two parts: binary and count models. Counts are estimated using either Poisson or negative binomial regression. Many applications of these models occur in epidemiology, ecology and related fields.

For example, Sileshi (2008) studied twelve soil animal count datasets from woodland and agro-forestry systems in eastern Zambia, using zero-inflated models and other generalized linear models, and concluded that the negative binomial distribution, zero-inflated Poisson and negative binomial models performed better than log-normal and Poisson models for these data.

In contrast, Warton (2005) concluded, after comparing the fits of different models to 20 published datasets, that many zeros do not necessarily mean zero-inflation. He found that the negative binomial model is a good approach for analyzing datasets with large numbers of zero counts, and even simple Gaussian models with log-transformed abundances fitted such datasets well. Therefore, in practice, it is desirable to try several statistical methods to avoid violation of relevant assumptions.

Log-transformed linear model

Linear regression based on a Gaussian distribution is the appropriate method for analyzing data with a continuous outcome of interest. If assumptions do not hold, it is possible to transform either the independent or dependent variables in the regression

model. Many forms of transformation are available to ensure that the data more closely meet the assumptions. A log-transformation of the outcome variable, with an appropriate modification to handle zero counts or organism abundance densities, is commonly used in applications in epidemiology, ecology, biology and other fields.

Lim and Tongkumchum (2009) focused on the study of length of stay for 40,498 mortality cases in hospitals in southern Thailand from 2000 to 2003, with respect to age, gender, principal diagnosis, provinces and hospital size. Linear regression was used to fit log-transformed length of stay. For all cases, older female patients stayed in the hospitals longer than other patients, and patients of different age groups encountered different diseases and injury categories.

Ardkeaw and Tongkumchum (2010) used a log-transformed linear regression model to fit the incidence rates of four infectious diseases including acute diarrhea, pyrexia of unknown origin, hemorrhagic conjunctivitis, and pneumonia in seven provinces of northeastern Thailand from 1999 to 2004. Districts, seasons, and age groups were used as predictors. Results showed that the highest incidence rates of each infectious disease were found among children aged below five years, and several peaks of incidence rates of pyrexia were observed over the period of study.

Further applications of the linear model with log-transformed outcomes include fish catch weights in Lake Songkhla (Chesoh and Lim 2008) and identifications of patterns of malaria incidence rates in northwestern Thailand (Sriwattanapongse 2009).

Generalized estimating equations

A method known as generalized estimating equations (GEE) is used to account for correlations between outcomes with respect to both time and space.

Ardkeaw and Tongkumchum (2009) used this method in their study of patterns of acute diarrhea incidence in children aged less than five years in five provinces of northeastern Thailand from 1999 to 2004. The linear regression model was used to investigate disease incidences in relation to districts, seasons and years, and the GEE method was used to handle spatial correlation between districts. They concluded that the incidences were higher in the first six months. Most districts in Loei and Amnat Charoen provinces had greater incidence rates. Similarly, a study by Yotthanoo and Choonpradub (2010) focusing on under-reporting used the GEE method to detect an increasing trend of acute diarrhea incidence rates in six provinces of Thailand bordering Cambodia over the study period of 1999-2004.

Canonical correspondence analysis

Canonical correspondence analysis is a popular ordination technique for displaying associations between species abundance outcomes and environmental predictor variables, and is used extensively in the biological literature (von Wehrden et al 2009).

Hajisamae and Chou (2003) investigated 23 species of fish in shallow sublittoral zones of eastern Johor Strait, Singapore. Using Canonical correspondence analysis to examine the effects of five environmental variables on fish assemblages, they found that some species had a positive correlation with distance from the strait and temperature. However, most

of the species were negatively correlated with the two factors indicating that they disappeared or appeared in low numbers both during high temperature and at the inner part of the strait.

Glockzin and Zettler (2008) examined the marine environment influencing spatial distributions of macro-zoo-benthic communities. Benthic and environmental data were collected from 2003 to 2006 at 191 sampling stations in the Pomeranian Bay (southwest Baltic Sea). Based on species abundances, distinctive macrobenthic community patterns were identified and evaluated via univariate correlation methods, multivariate numerical classification, principal components analysis, and canonical correspondence analysis. These patterns were caused by clear responses of several benthic species to certain prevailing environmental conditions. The observed distribution of selected species followed a strong gradient of depth and was explained best by the sediment parameters including total organic carbon, median grain size and sorting.

Hettrich and Rosenzweig (2003) modeled the spatial distribution of grassland vegetation, mollusks and carabids in a study area at the Middle Elbe, Germany. Canonical correspondence analysis detected clear dependencies between the occurrence of biotic objects and mainly hydrological parameters. The predicted spatial distribution of species for nearly all field studies examined could be depicted with these instruments. Comparisons between investigated and predicted distribution of species showed high correspondence.

Multivariate multiple linear regression

Multivariate multiple linear regression is the extension of multiple linear regression to allow for several correlated outcome variables.

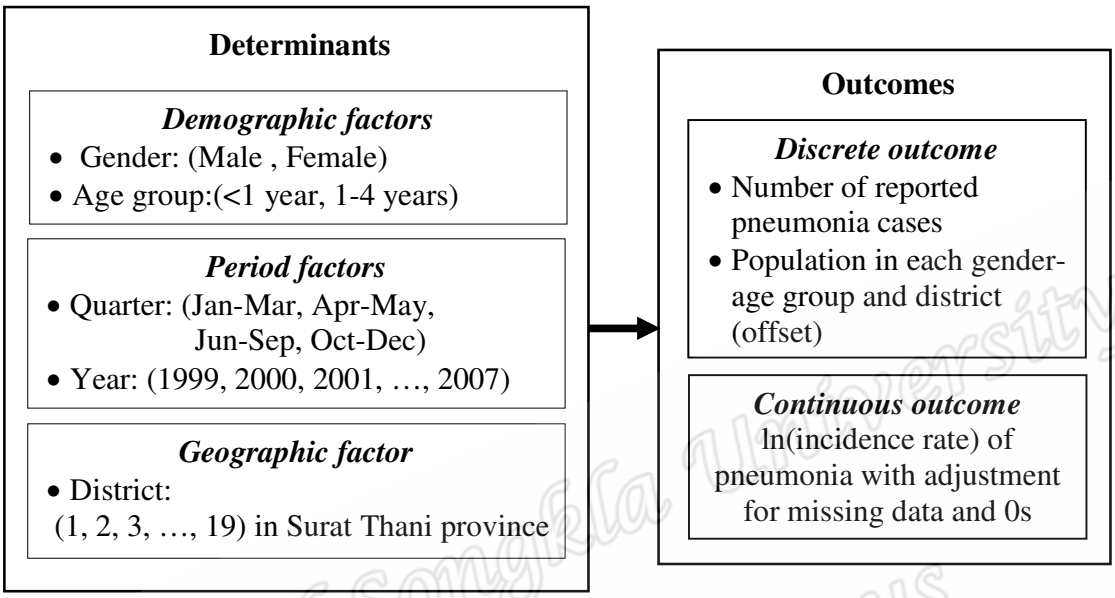
Based on nineteen data sets taken from the ecology literature, Warton and Hudson (2004) showed that multivariate regression methods gave different results, but were just as powerful as other methods based on different similarity indexes used in the ecological literature. Since the regression methods give more easily interpretable results and can be generalized to more complex designs, they recommended that these methods be used in preference to the existing methods in the ecological literature.

Liang et al (2002) used structural equation modeling (an extension of the multivariate regression model where observed outcomes are reduced to a smaller number of latent variables) to investigate the dependence of two levels of organism densities on physiochemical water characteristics.

1.3 The studies

Although the data for our studies arose from three quite difference sources, the data structure was essentially the same in each case, with the outcome defined as an incidence rate or a set of continuous variables based on cells classified by categorical determinant variables, and other continuous or categorical determinants defined on the same cells. The path diagrams for the two epidemiological studies and for the ecological study are shown in Figure 1.1and Figure 1.2 respectively.

Study 1: Pneumonia incidence in children aged under five years in Surat Thani, Thailand 1999-2007



Study 2: Tuberculosis (TB) in southern province of Thailand, 1999-2004

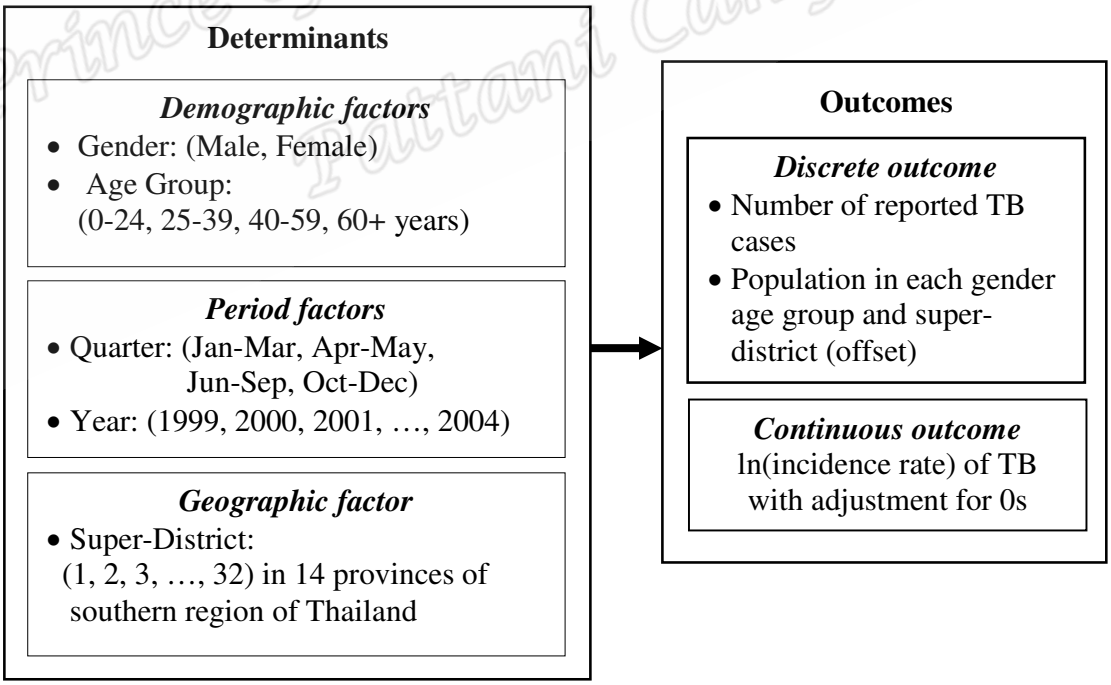


Figure 1.1: Path diagrams for variables used in two epidemiological studies

Study 3: Macrobenthic fauna densities in the Middle Songkhla Lake

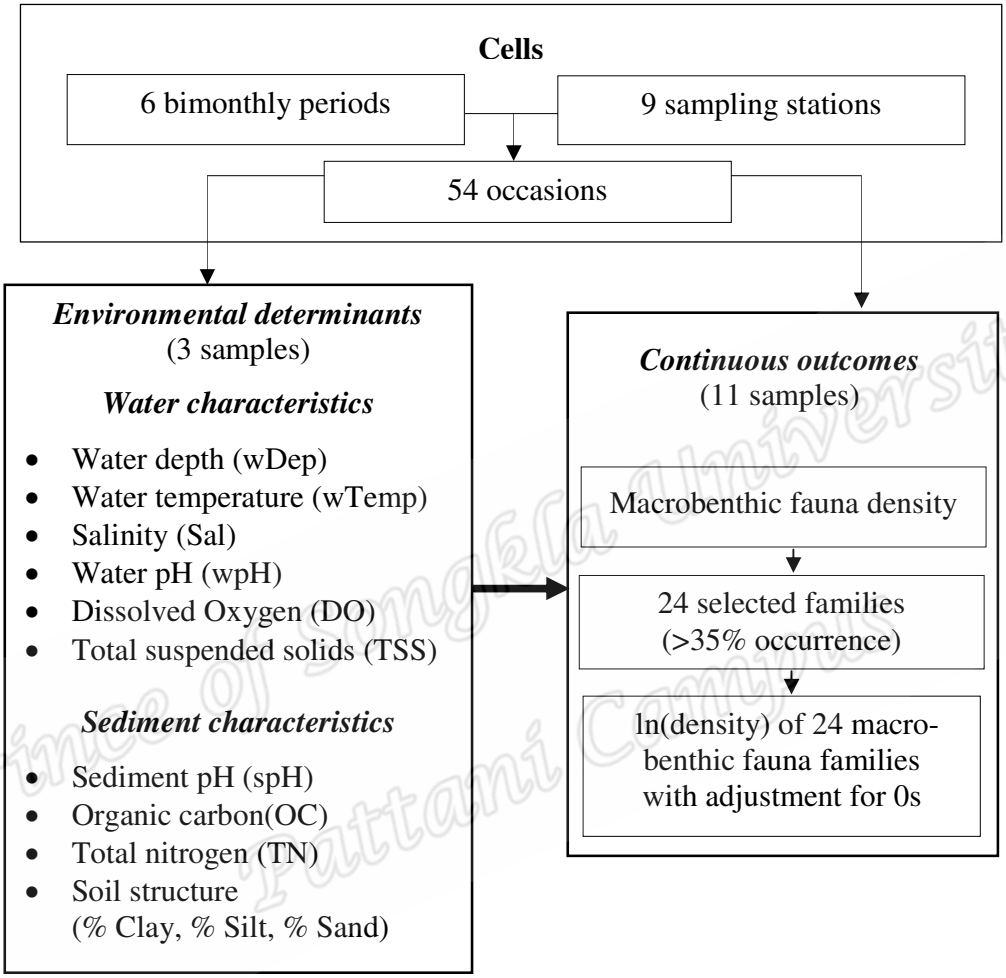


Figure 1.2: Path diagram for variables used in ecological study

1.4 Objectives and plan of thesis

Appropriate statistical models were used to model incidence rates and macrobenthic fauna densities. These models attempted to identify the associations between demographic factors (location, season, age, and gender) and data outcomes (incidence rates), so linear regression, Poison regression, and negative binomial regression models

and log-transformed linear model were applied to fit these first two studied on epidemiology data. Moreover for ecologist, they tried to identify the associations between continuous (environmental data) determinants and continuous outcomes (macrobenthic fauna densities).

The objectives of studies were thus as follows:

(a) To develop and applied statistical methods for modeling overdispersion of incidence rates.

(b) To applied alternative method to examine the distributional patterns of macrobenthic fauna assemblages in relation to environmental predictor variables.

This thesis contains four chapters. The introductory chapter discusses the rationale, the scope and the aim of the study, and also includes a review of some relevant literature.

Chapter 2 provides a description of the methodology including an overview of the statistical methods for data analysis aligned to the statistical models.

Chapter 3 reports on the data collection and preliminary data analysis and includes the published article and manuscripts that were written as part of the thesis.

The last chapter states the summaries and conclusions in each study. Suggestions for further research are also provided in this chapter.