Chapter 3

Modeling overdispersed response data

This chapter reports on data collection, data management, and preliminary data analysis for our three studies, and includes the published article and manuscripts that were written as part of the thesis.

Graphical and statistical analyses were carried out using the program available from the R package library (R Development Team 2008). The methods for canonical correspondence analysis used CANOCO Version 4.5 (Ter Braak and Šmilauer 2002).

3.1 Data source and data management

Epidemiological data used for the first two studies were obtained from the National Notifiable Disease Surveillance (Report 506) collected routinely in each of Thailand's 76 provinces by the Ministry of Public Health. Reported cases come from health stations located in each province (Leelarasamee et al 2004). Data for each year are available in computer files with records for disease cases and fields comprising characteristics of the subject and the disease, including dates of sickness and disease diagnosis, the subject's age, gender, and address, and the severity of the illness including date of death for mortality cases. Extensive cleaning was required to correct or impute data entry errors.

For the first study the records for Surat Thani province for the nine years from 1999 to 2007 were stored in an SQL database. SQL programs were used to create pneumonia disease counts by age group (less than 1 or 1-4 years) from children by month and district in Surat Thani province for the nine years from 1999 to 2007 and stored in an

SQL database. Incidence rates were computed as the number of cases per 1000 residents in the appropriate demographic group in each district according to the 2000 Thai Population and Housing Census. For simplicity the denominator populations in the two age groups were estimated approximately by apportioning the Census populations aged 0-4 for each gender in the ratio 1:4. Microsoft Excel was used to manage the data analysis. Data cleaning was undertaken for correct coding and dealing with the missing values by using MySQL. WebStat (a web-database engineering tool developed by Don McNeil) was used for preliminary data analysis, and we extracted GIS district boundary data from MapInfo for producing maps using R programs. The flow diagram for data management is summarized in Figure 3.1.

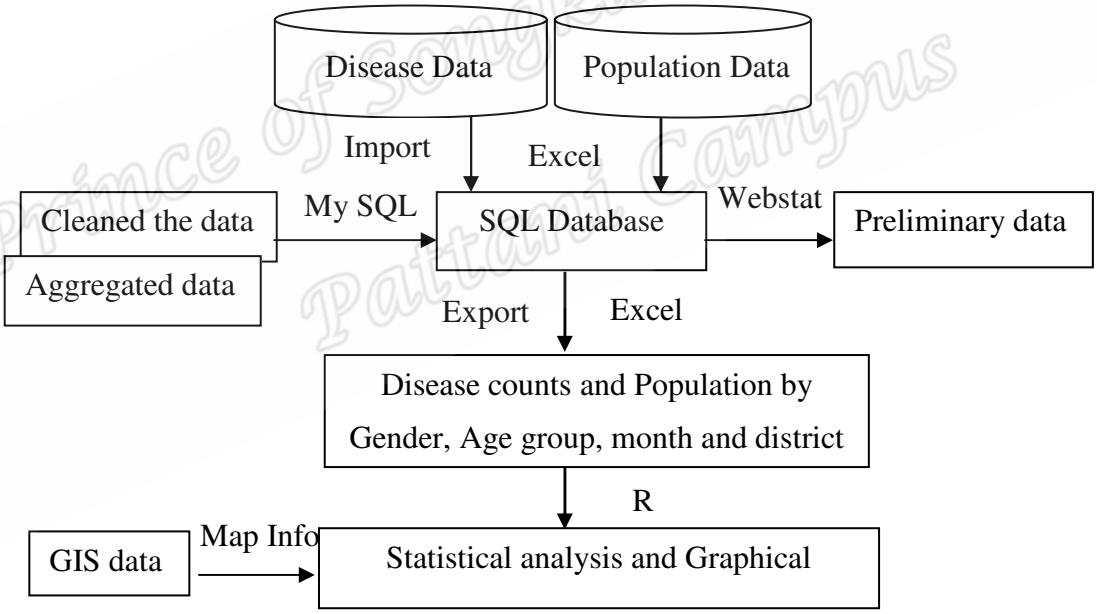

Figure 3.1: Flow diagram for data management

We used a similar procedure for the second study of tuberculosis (TB) in the fourteen provinces in southern of Thailand during 1999–2004.

With respect to the third study, Angsupanich et al (2005a) collected macrobenthic fauna from the nine sampling stations. The assemblages were conducted with 11 samples for each station at bimonthly intervals from April 1998 to February 1999.

The densities of macrobenthic fauna were recorded as individuals per cubic meter volume for each species. In many cases the species could not be identified exactly, so we used the family instead of the species in our model. A total of 161 species of macrobenthic fauna were found. They were classified into 81 families. With nine locations and six bimonthly data collection periods, we defined the coverage for a specified family as the proportion of these 54 occasions on which at least one organism was found. We then selected the 24 families (93.2% total assemblages) with greater than 35% coverage for developing the statistical model.

Environmental variables comprised water depth (wDep), water temperature (wTemp), salinity (Sal), water pH (wpH), dissolved oxygen (DO), total suspended solids (TSS), sediment pH (spH), total nitrogen content (TN), organic carbon content (OC), and soil structure (percentages of sand, silt and clay). These were measured with three samples on the same occasions as the biotic data.

3.2 Studies completed

The first study aimed to compare the negative binomial GLM with a log transformed linear model using an extensive set of data for quarterly pneumonia incidence rates among children reported from districts in Surat Thani province in Thailand from 1999 to 2007. The manuscript was accepted for publication in the *Chiang Mai Journal Science*, and appear in Volume 37 No.1 pages 29-38 in January 2010.

The second study aimed to model the joint effects of gender-age, quarterly season and location associated with the incidence of tuberculosis in the southern region of Thailand during 1999–2004. We again compared the negative binomial GLM with a log transformed linear model. This paper has been submitted to the *Southeast Asian Journal of Tropical Medicine and Public Health, 41(3).*

The third paper examined the relations between the sediment densities of 24 selected macrobenthic families and a reduced set of environmental predictors in the middle of Songkhla Lake using multivariate multiple regression (MMR) and canonical correspondence analysis (CCA) and compared the results for the two methods. This paper has been submitted to the *Chiang Mai Journal Science*.

3.3 Preliminary analyses

***Study 1*:** Pneumonia incidence rate among children in Surat Thani, 1999-2007

Figure 3.2 shows overall incidence rates by gender, age-group and period. The bar chart in the left panel shows that incidence rate were highest in males age less than one. The time series plot of childhood pneumonia incidence rates from January 1999 to December 2007 is shown in the right panel. Figure 3.3 shows a thematic map of classifying district incidence rates as high (red), average (orange), and low (yellow).

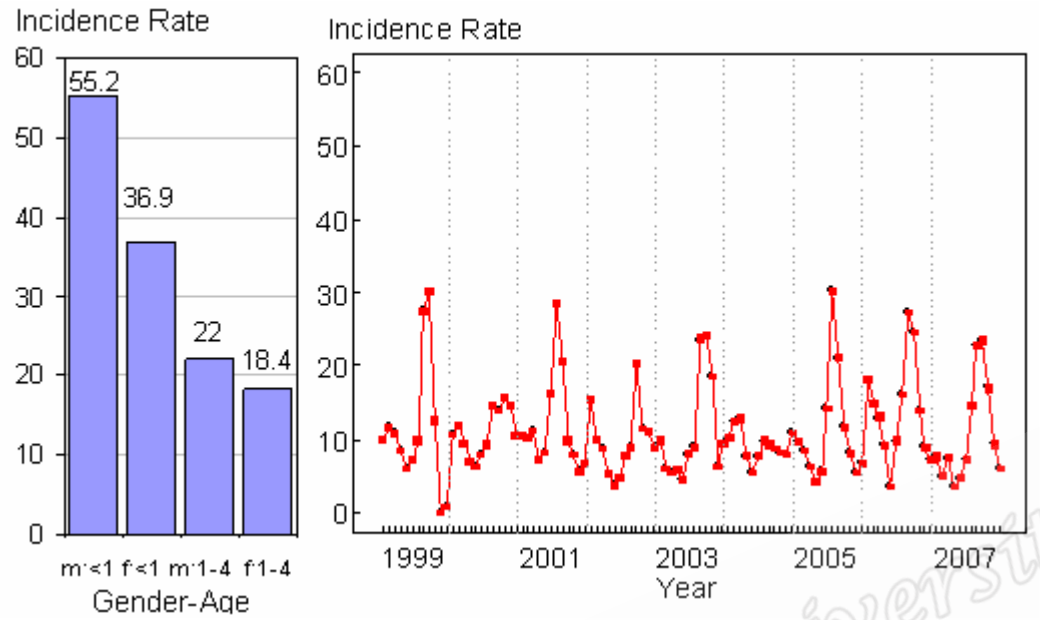

Figure 3.2: The total annual incident rates in gender age group and time

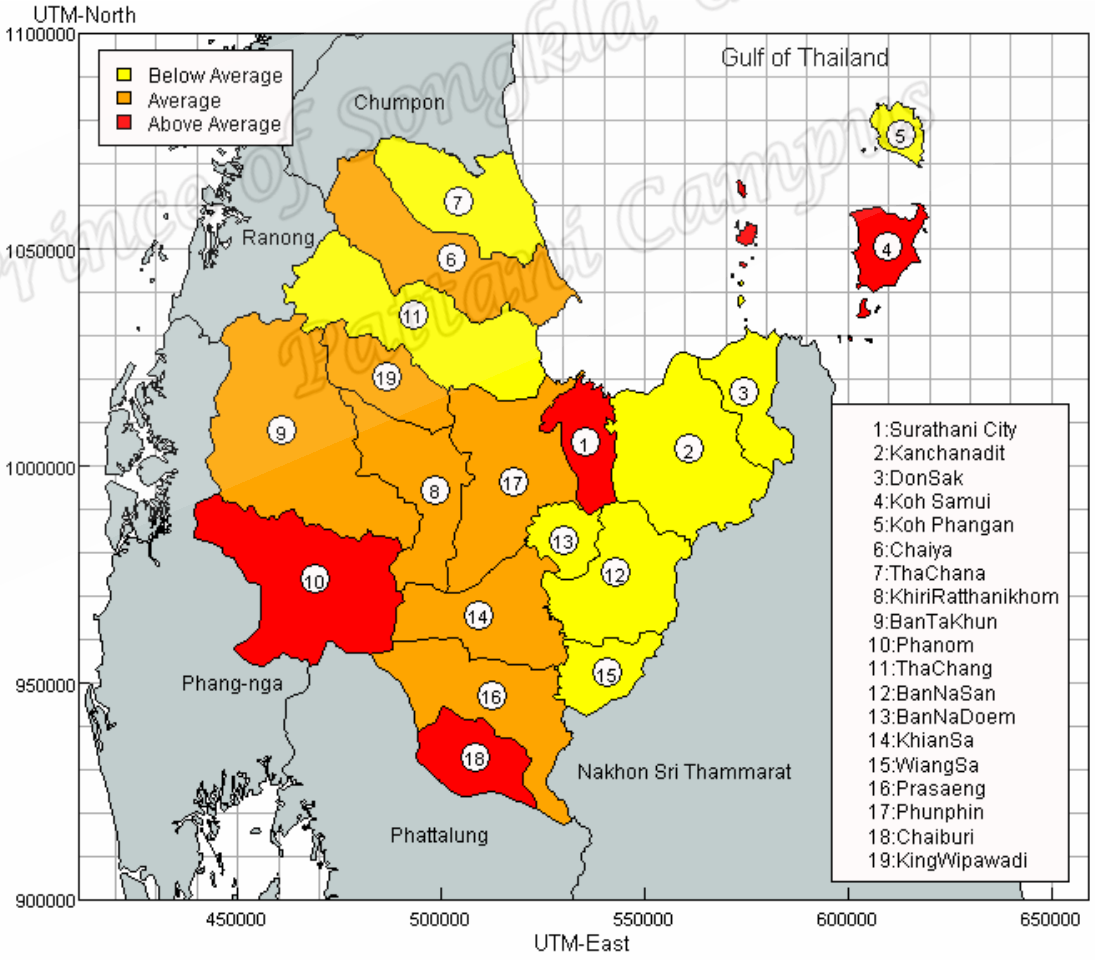

Figure 3.3: Map of Surat Thani showing incidence rate in Surat Thani province

Study 2: A total of 21,831 tuberculosis cases were reported from the 14 southern provinces of Thailand from January 1999 to December 2004. The incidence rate of tuberculosis in this region was calculated to be 143. Table 3.1 shows the TB incidence rate for each demographic group by gender and age-group. The highest incidence was found among people aged 60 or more, and lowest incidences were revealed in people aged below 25 years. Among males, the incidence rate was 5.4. The highest incidences rates were found in Songkla (22.3), Pattani (21.6), and Surat Thani (21.9) provinces, respectively. The lowest incidence rate was found in Satun province (2.48). The results are shown in Table 3.2.

Table 3.1: TB incidence per 100,000 by age group and gender

Determinant	Number of TB cases	Population	Incidence rate
Age in years			
below 25 yrs	2,755	267,827,400	1.03
25-39 yrs	7,515	145,285,272	5.17
40-59 yrs	6,284	113,836,320	5.52
60+ yrs	5,277	553,489,20	9.53
Gender			
Male	15,589	288,946,296	5.4
Female	6,242	293,351,616	2.13

Table 3.2 TB incidence per 100,000 in each province in the southern region of Thailand

Province	Number of districts	Number of TB cases (N=21,831)	Population (N=8.087,471)	Incidence rate
Nakhon Si Thammarat	23	3,304	109,426,392	5.78
Krabi	8	456	24,207,120	4.47
Phang-nga	8	957	16,861,536	17.64
Phuket	3	959	17,960,112	5.67
Surat Thani	19	2,706	62,597,520	21.93
Ranong	5	519	11,607,120	3.59
Chumpon	8	950	32,126,832	5.34
Songkhla	16	3,345	90,407,664	22.29
Satun	7	442	17,847,000	2.48
Trang	10	830	42,847,920	5.32
Phatthalung	11	963	35,889,912	5.36
Pattani	12	3,226	42,910,920	21.61
Yala	8	1,473	29,918,664	9.94
Narathiwat	13	1,701	47,689,200	10.75

The preliminary analysis also showed age and gender differences, with substantially higher rates for males, and regional differences, with higher rates in Surat Thani, Phang-nga and Songkla provinces. The further analysis involved developing a statistical regression model containing effects for age-gender, location and time after log transforming incidence rates.

***Study* 3:** Macrobenthic families in the middle of Songkhla Lake.

Tables 3.3 and Table 3.4 show descriptive statistics and correlation coefficients of 13 environmental parameters across nine sampling stations located in the Middle Songkhla Lake during the study period, with correlations greater than 0.4 in bold. Percent of clay had a strong negative correlation with percent of sand (-0.88). Salinity had a strongly positive correlation with water pH (0.68) and negative correlation with water depth (-0.52). Factor analysis was used for removing correlation between environmental parameters that mask their effects on the macrobenthic family densities.

Table 3.3: Descriptive statistics for 13 environmental variables

Determinant	Min.	Max.	Med.	Mean	SD.
OC	0.38	3.69	0.87	1.02	0.64
TN	0.01	0.45	0.07	0.10	0.11
pHSed	5.06	7.48	6.60	6.48	0.54
wDepth	0.33	2.73	1.34	1.36	0.54
wpH	5.56	8.59	7.26	7.21	0.82
DO	5.62	8.73	7.13	7.17	0.74
TSS	2.40	144.50	51.95	56.43	29.97
Sal	0.00	30.20	18.30	14.51	10.16
wTemp	26.76	33.83	29.20	29.47	1.61
Clay	1.49	68.95	41.92	38.76	16.12
Silt	1.70	65.27	41.85	40.07	15.38
Sand	0.24	96.81	9.85	21.17	27.66

Table 3.4: Pearson correlation coefficients of environmental variables

	OC	TN	spH	wDepth	wpH	DO	TSS	Sal	wTemp	clay	sand
OC	1.00										
TN	**0.42**	1.00									
pHSed	-0.16	-0.05	1.00								
wDepth	-0.21	-0.37	-0.15	1.00							
wpH	0.09	**0.41**	-0.02	-0.25	1.00						
DO	0.02	0.09	-0.03	0.09	0.11	1.00					
TSS	0.14	0.23	0.00	-0.29	0.13	0.09	1.00				
Sal	0.17	**0.50**	0.19	**-0.52**	**0.68**	-0.03	0.32	1.00			
wTemp	0.25	**0.61**	-0.25	-0.32	0.38	0.00	0.37	**0.59**	1.00		
Clay	0.29	0.10	-0.14	-0.07	0.01	0.05	0.19	0.30	0.27	1.00	
Sand	-0.13	0.10	-0.05	0.09	0.05	-0.14	-0.24	-0.24	-0.08	**-0.88**	1.00

Table 3.5 shows descriptive statistics of densities (individuals per cubic meter) of the 24 macrobenthic families with their dominant greater than 35% for the raw data (left panel) and after using the transformation (right panel). These distributions are all highly positively skewed. To reduce skewness and stabilize the error variance in the model, the transformation ln (1+c×density) was used, with c chosen as 100 to best approximate normality. This transformation enables zeros to be natural log-transformed.

Table 3.5: Descriptive statistics for 24 macrobenthic family densities (ind m^{-2})

Family	Raw data					After transformation			
	Min.	Max.	Med.	Mean	SD.	Max.	Med.	Mean	SD.
01:*Capitellidae*	0.0	165.5	10.9	22.7	30.9	9.7	7.0	6.3	2.7
02:*Goniadidae*	0.0	101.8	0.0	8.2	20.4	9.2	0.0	2.6	3.5
03:*Hesionidae*	0.0	114.5	1.8	12.9	24.2	9.4	5.2	3.9	3.6
04:*Nephtyidae*	0.0	310.9	12.7	41.1	69.3	10.3	7.2	6.6	2.9
05:*Nereididae*	0.0	876.4	50.0	157.5	214.9	11.4	8.5	8.4	2.0
06:*Pilargiidae*	0.0	189.1	3.6	30.1	49.5	9.9	5.9	5.2	3.6
07:*Pholoidae*	0.0	114.5	1.8	12.2	25.2	9.4	5.2	4.0	3.5
08:*Spionidae*	0.0	978.2	21.8	93.6	198.7	11.5	7.7	7.4	2.6
09:*Terebellidae*	0.0	854.5	0.0	21.0	116.2	11.4	0.0	2.5	3.5
10:*Aoridae*	0.0	805.5	3.6	44.8	131.8	11.3	5.9	4.5	4.0
11:*Isaeidae*	0.0	1581.8	22.7	127.8	274.1	12.0	7.7	7.2	3.2
12:*Melitidae*	0.0	776.4	38.2	82.2	133.8	11.3	8.3	7.7	2.4
13:*Oedicerotidae*	0.0	63.6	3.6	12.4	17.6	8.8	5.9	4.6	3.4
14:*Apseudidae*	0.0	5043.6	164.5	742.3	1164.0	13.1	9.7	8.9	3.5
15:*Pseudotanaidae*	0.0	1467.3	0.0	79.0	295.6	11.9	0.0	2.8	3.9
16:*Anthuridae*	0.0	820.0	5.5	70.7	181.7	11.3	6.3	5.6	3.5
17:*Cirolanidae*	0.0	203.6	0.0	7.9	31.2	9.9	0.0	2.4	3.2
18:*Alpheidae*	0.0	18.2	0.0	1.8	3.3	7.5	0.0	2.4	2.9
19:*Marginellidae*	0.0	632.7	26.4	73.4	128.0	11.1	7.9	7.0	3.2
20:*Retusidae*	0.0	1958.2	1.8	102.5	341.5	12.2	5.2	4.3	4.2
21:*Skeneopsidae*	0.0	700.0	0.0	17.7	95.8	11.2	0.0	2.6	3.4
22:*Stenothyridae*	0.0	383.6	0.0	10.8	52.6	10.6	0.0	2.3	3.3
23:*Pelecypoda*	0.0	110.9	0.0	6.3	16.2	9.3	0.0	3.0	3.4
24:*Tellinidae*	0.0	3494.5	34.5	317.3	807.2	12.8	8.2	7.0	3.8

3.4 Published article and manuscripts written as part of the thesis

Chiang Mai J. Sci. 2010; 37(1) : 001-010
www.science.cmu.ac.th/journal-science/josci.html
Contributed Paper

Methods for Modeling Incidence Rates with Application to Pneumonia among Children in Surat Thani, Thailand

Noodchanath Kongchouy* [a], Chamnein Choonpradub [b], and Metta Kuning [b]**
[a] Department of Mathematics, Prince of Songkla University, Hat Yai, Songkhla, 90112, Thailand.
[b] Department of Mathematics and Computer Science, Prince of Songkla University, Pattani, 94001, Thailand.
Author for correspondence; e-mail: *nootchanath.k@psu.ac.th, **cchamnein@bunga.pn.psu.ac.th

Received: 10 July 2009
Accepted: 29 October 2009

ABSTRACT

In statistical studies, generalized linear models (GLMs) are usually preferred for modeling incidence rates, often with extensions to zero-inflated GLMs when the proportion of zero counts is large. However Warton has shown that for many ecological studies, simple linear models fitted to log-transformed counts do surprisingly well. In this study, we used data comprising a sizable set of pneumonia incidence rates. We compared the negative binomial GLM with a log transformed linear model, and found further support for this simpler alternative method.

Keywords: log-linear regression model, zero counts, negative binomial GLM, pneumonia incidence rates.

1. INTRODUCTION

While linear regression models based on normal errors are used extensively in statistical analysis, they are usually not preferred for analyzing counts and incidence rates. Crawley [1] recommended against using linear regression models in this situation because (a) such responses are necessarily non-negative and simple linear models can give negative fitted values; (b) the variance of such responses usually increases with the mean, whereas the linear model assumes a constant variance; (c) the errors are usually not normally distributed; and (d) transformation of the response is difficult when zero counts occur. Thus it is now common practice among statisticians to prefer more complex models. The Poisson GLM is usually preferred as a starting point, with an extension to a negative binomial or quasi-Poisson model (see, for example, Venables and Ripley [2], Jansakul and Hinde [3], Bulsara et al. [4], Faraway [5]) to handle overdispersion. When excessive zero counts occur in such models, they are extended further to zero-inflated or "hurdle" models by combination with logistic models that can have separate sets of coefficients (see, for example, Cameron and Trivedi [6,7], Dalrymple et al. [8]).

Despite this research effort, there are many situations to which these complex advanced methods have yet to be applied. For example, generalized linear models have not

been fully extended to outcome data with correlated errors, although similar models with correlated Gaussian errors such as those based on multivariate multiple regressions, generalized estimating equations [9], time series and spatially correlated processes [10, 11], and GARCH models [12] are available. Many of the applications for which these methods are needed also involve count data.

Despite Crawley's claim with respect to the difficulty of achieving normality by transforming the responses, particularly those taking zero values, transformations can be both effective - as Tukey [13] demonstrated with many examples - and theoretically uncomplicated - as Miller [14] showed. Moreover, it is common in biological and environmental science applications to transform counts by adding 1 and then taking logarithms, thus retaining the zeros in the outcomes as Clarke and Warwick [15] recommended. It is also important to realize that the statistical assumption of normality applies to the errors from the fitted model rather than the responses, and with currently available software it is straightforward to assess the plausibility of this assumption by examining plots of residuals against normal quantiles.

Based on an extensive study of ecological data, Warton [16] found that after appropriate data transformation, the more traditional models based on Gaussian error assumptions provided surprisingly good fits – in many cases superior to those based on zero-inflated models even when the responses had high proportions of zeros.

While biostatisticians and epidemiologists are among the strongest advocates of the complex new models [17, 18], it is possible that Warton's findings may also apply to medical data. In this paper we investigate the use of standard linear regression after data transformation for analyzing incidence rates, with application to an extensive set of data for quarterly pneumonia incidence rates among children reported from districts in Surat Thani province in Thailand from 1999 to 2007.

2. METHODS

Suppose that n_{ijqt} is the number of observed cases in cells defined by demographic (age-gender) group i, region j, season q and year t and P_{ij} is the corresponding population at risk. Denoting the corresponding mean incidence rate by λ_{ijqt}, we consider additive models of the form

$$\ln(\lambda_{ijqt}/P_{ij}) = \mu+\alpha_i+\beta_j+\eta_q+\gamma_t. \qquad (1)$$

The terms α_i, β_j, η_q and γ_t represent demographic group, region, season and year effects, respectively, and are centred at 0, so that μ is a constant encapsulating the overall incidence rate. Since the observations n_{ijqt} take non-negative integer values, the simplest generalized linear model is based on the Poisson distribution with mean λ_{ijqt} where P_{ij} is an offset [2, 19]. Observed incidence rates are often over-dispersed due to spatial clustering, so the negative binomial GLM - an extension of the Poisson GLM that allows for over-dispersion - may be preferred. The variance of this distribution is $\lambda_{ijqt}(1+\lambda_{ijqt}/\theta)$ with the Poisson model arising in the limit as $\theta\to\infty$. The model fit is assessed by comparing deviance residuals with normal quantiles, and it is also informative to plot observed counts and appropriately-scaled incidence rates against corresponding fitted values based on the model. The model also gives adjusted incidence rates for each factor of interest, obtained by suppressing the subscripts in Equation (1) corresponding to the other factors, and replacing these terms with a constant. The constant satisfies the condition that the sum of the counts based on the adjusted incidence rates matches the total.

Sum contrasts [2, 19] may be used to obtain confidence intervals for comparing the adjusted incidence rates within each factor with the overall incidence rate. An advantage of these confidence intervals is that they provide a simple criterion for classifying levels of a factor into three groups according to whether each corresponding confidence interval exceeds, crosses, or is below the overall mean.

An alternative additive log-linear model for the incidence rates with normally distributed errors is

$$\ln(n^*_{ijqt} / P_{ij}) = y_{ijqt} = \mu + \alpha_i + \beta_j + \eta_q + \gamma_t + \varepsilon_{ijqt}. \quad (2)$$

In this model n^*_{ijqt} is a simple modification of the disease count n_{ijqt} to ensure that the incidence rates are positive, and thus can be log-transformed. In this case, the model fit can be assessed by plotting studentized residuals against normal quantiles, by again plotting observed counts and appropriately-scaled incidence rates against corresponding fitted values based on the model, and also by using the r-squared to see how much of the variation in the data is accounted for by the model.

Various methods may be considered for this data modification. As a starting point, cases where the count is zero could simply be omitted, and the fitted model then used to impute counts for these cases before refitting the model. This method may be desirable in situations where under-reporting is known or suspected. Another method involves adding a constant c (say 1) to all counts so that $n^*_{ijqt} = n_{ijqt} + c$. A third method involves replacing the zeros by a suitably chosen constant d without changing any values of n_{ijqt} greater than 0.

3. PNEUMONIA IN SURAT THANI PROVINCE

In Thailand, all hospital-diagnosed infectious disease cases are routinely recorded by the Ministry of Public Health in each of its 13 regional health areas, and these records include pneumonia. Among the diseases reported, pneumonia was by far the most lethal in seven provinces of the upper southern region, accounting for 47.7% of all deaths from hospital-diagnosed cases of infectious diseases in the area during 1999-2007. Of the seven provinces, Surat Thani province recorded the highest average incidence rate of pneumonia cases (8.3%) during this period.

Previous publications on pneumonia mortality and morbidity in Thailand are not extensive. They include a study by Kanlayanaphotporn et al. [20] of pneumonia cases reported in 1999-2001 by the Ministry of Public Health surveillance system in Sakaeo province near the Cambodian border. Suwanjutha et al. [21] studied risk factors associated with mortality and morbidity of community acquired pneumonia in Thai children younger than 5 years of age. Reechaipichitkul and Tantiwong [22] studied clinical features of community - acquired pneumonia among patients treated at Srinagarind Hospital in Khon Kaen province in the north-eastern region.

Surat Thani province is divided into 19 districts. Incidence rates were computed as the number of cases per 1,000 residents in cells defined by demographic group, district, and quarterly period according to the 2000 Thai Population and Housing Census. Since most cases occurred among children under 5 years of age, we defined four demographic groups, separating boys and girls, and those aged less than 1 and those aged 1-4.

The R program [23] was used for all statistical analysis, graphs and maps.

4. RESULTS

During the study period from January 1999 to December 2007, 15,919 cases of

pneumonia were reported to have been diagnosed at district hospitals in Surat Thani province among children below 5 years of age. Table 1 and Figure 1 show the pneumonia incidence rates for each demographic group by district and month, respectively. The number of cases reported per month for any district varied from zero to a maximum of 51. The highest average annual incidence rates occurred in the south-western district of Phanom (45.9 per 1,000) and in the urban central coastal district of Surat Thani City (43.1 per 1,000).

Table 1. Reported annual pneumonia incidence in children in Surat Thani classified by demographic and geographic factors during 1999-2007.

District	Number of Cases				Annual Incidence Rate /1,000					Population
	m:<1	f:<1	m:1-4	f:1-4	m:<1	f:<1	m:1-4	f:1-4	total	
1. Surat Thani City	823	535	1,463	1,236	85.8	57.9	38.1	33.5	43.1	10,457
2. Kanchanadit	278	192	415	313	39.1	29.3	14.6	12.0	17.5	7,588
3. Don Sak	86	48	196	128	32.0	19.7	18.2	13.1	17.8	2,851
4. Koh Samui	152	109	281	232	66.2	46.6	30.6	24.8	33.4	2,572
5. Koh Pha-ngan	7	9	15	11	10.4	13.3	5.6	4.1	6.2	747
6. Chaiya	173	98	214	160	54.6	32.9	16.9	13.4	21.0	3,415
7. Tha Chana	115	71	190	120	32.4	20.9	13.4	8.8	14.3	3,856
8. Khiri Ratnikhom	188	98	347	222	67.0	37.4	30.9	21.2	31.5	3,013
9. Ban Ta Khun	45	38	100	92	44.6	46.4	24.8	28.1	30.1	1,015
10. Phanom	237	140	397	343	93.1	60.3	38.9	36.9	45.9	2,706
11. Tha Chang	86	61	105	91	43.0	34.1	13.1	12.7	18.1	2,103
12. Ban Na San	147	94	140	145	32.2	21.1	7.6	8.1	11.6	5,018
13. Ban Na Doem	48	22	58	52	35.8	16.3	10.8	9.7	13.4	1,495
14. Khian Sa	199	115	337	270	62.5	36.6	26.5	21.5	29.1	3,513
15. Wiang Sa	195	122	243	175	43.8	30.7	13.7	11.0	17.4	4,685
16. Phrasaeng	272	184	329	217	60.4	42.6	18.3	12.6	22.7	4,897
17. Phunphin	310	173	582	447	49.2	29.7	23.1	19.2	24.9	6,738
18. Chaiburi	137	86	194	152	77.7	54.3	27.5	24.1	34.0	1,859
19. King Vibhavadi	60	39	70	45	65.4	53.5	19.2	15.5	26.1	911
Total	**3,558**	**2,234**	**5,676**	**4,451**	**56.4**	**55.2**	**36.9**	**22.0**	**25.5**	**69,439**

Note: m:<1= male and aged less than 1, f:<1= female and aged less than 1
m:1-4= male and aged 1-4 , f:1-4= female and aged 1-4

Since no cases were recorded for any district in November and December 1999 (clearly due to under-reporting), the disease counts used for fitting the models to data from these two months for each district and demographic group were imputed using simple interpolation based on the corresponding frequencies for October 1999 and January 2000.

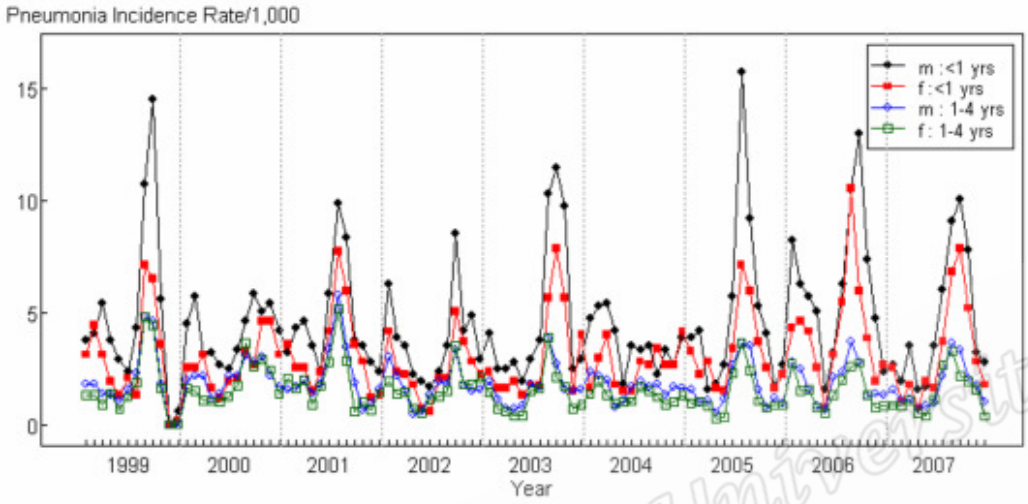

Figure 1. Monthly pneumonia incidence rates for children in Surat Thani Province.

The time series plots show a pronounced seasonal pattern with annual peaks in the July-September months proceeded by lower rates around May of each year, but little evidence of any trend. The rates have a similar pattern for each demographic group, with higher rates in the younger age groups, particularly for males.

To reduce the correlation between residuals in successive time periods, while still enabling seasonal patterns to be seen, we used quarterly aggregates of disease counts as a basis for fitting the models and estimating annual incidence rates. The pneumonia cases (including the additional imputed cases) were thus grouped by gender-age group, district, and quarter, assigning 16,253 individual cases to 2,736 records (4 gender-age groups × 19 districts × 36 quarters).

Figure 2 shows plots of studentized residuals against normal quantiles based on model (2) with various ways of handling the zero counts. The left panel was obtained simply by omitting all zero counts so that the sample size was reduced to 2,186, whereas the other panels compare the two methods for recoding these counts after choices of constants $c = 1$ and $d = 1$, respectively. These constants were selected to give residuals plots closest to a linear pattern corresponding to the model in each case.

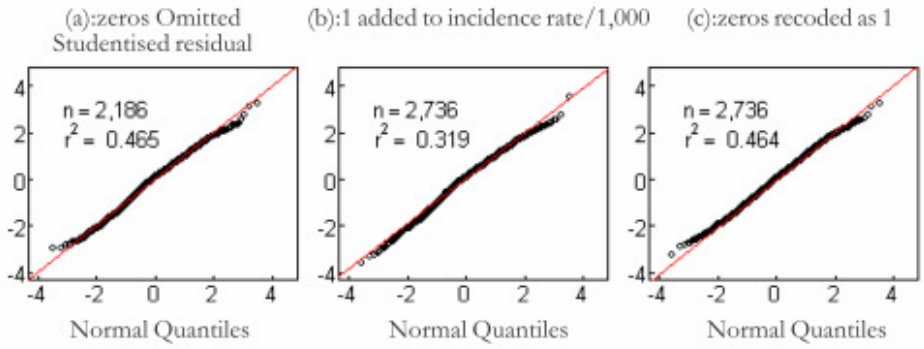

Figure 2. Studentised residuals after omitting zeros (left panel); adding 1 to incidence rates (middle panel); replacing zeros by 1 (right panel).

Since omitting zero counts could give biased results as well as eliminating most the data in districts with small populations, and there was no firm evidence of under-reporting apart from the two months where data were imputed, method (a) was not pursued further in the study. For the method based on normally distributed errors, in further analysis the zero counts were replaced by a positive constant because method (c) gave a substantially higher r-squared (0.464) than that given by the alternative method (b) (0.319).

The left and middle panels of Figure 3 show plots of observed counts and annual incidence rates versus corresponding fitted values based on the negative binomial model (upper panels); and the log-linear alternative with zero counts replaced by 1 (lower panels). A cube root scale was used to improve the clarity of the plots of counts and incidence rates against fitted values.

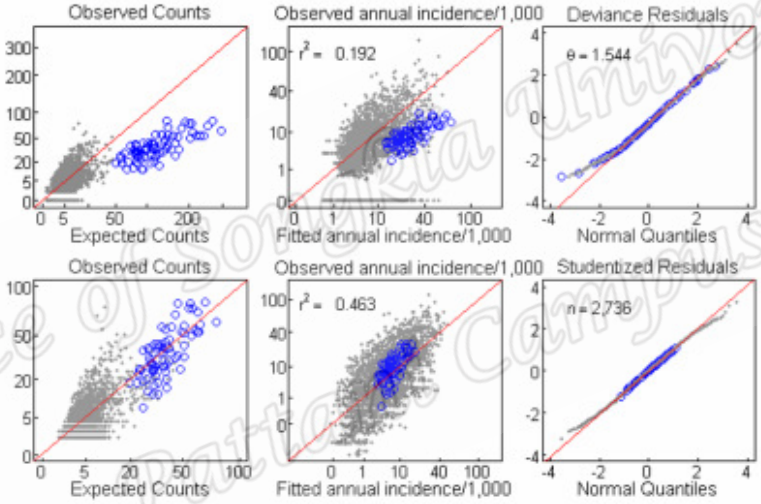

Figure 3. Results from negative binomial model (above) and log-linear model (below).

The circles in the plots correspond to data from Surat Thani City district, where the population is much higher than in other districts (Table 1). It can be seen that the expected counts and incidence rates given by the negative binomial model for this district are all substantially higher than their corresponding observed values (gray dots). On the other hand, no such bias occurs with the log-linear model. Furthermore, for the negative binomial model, the plot of deviance residuals against normal quantiles shows some departure from linearity, whereas the corresponding plot for the log-linear model is much straighter. We computed the correlation coefficients of residuals between districts after fitting the log-linear model. For all 171 pairs of the 19 districts, these correlations ranged from -0.17 (between districts 11 and 19) to +0.29 (between districts 2 and 7), with mean 0.036 and standard error 0.0078 (95% CI: 0.021-0.051). For the 33 pairs of districts with a common land border (or less than 3 km distant by sea if an island), the correlations ranged from -0.17 to +0.25 (between districts 14 and 16), with mean 0.066 and standard error 0.017 (95% CI: 0.031-0.101). These spatial correlations are statistically significant but not sufficiently large to justify any adjustment to the model.

Figure 4 shows confidence intervals for adjusted incidence rates obtained using the log-

linear model. The dotted horizontal line on each graph gives the overall mean annual incidence rate (26.9 per 1,000).

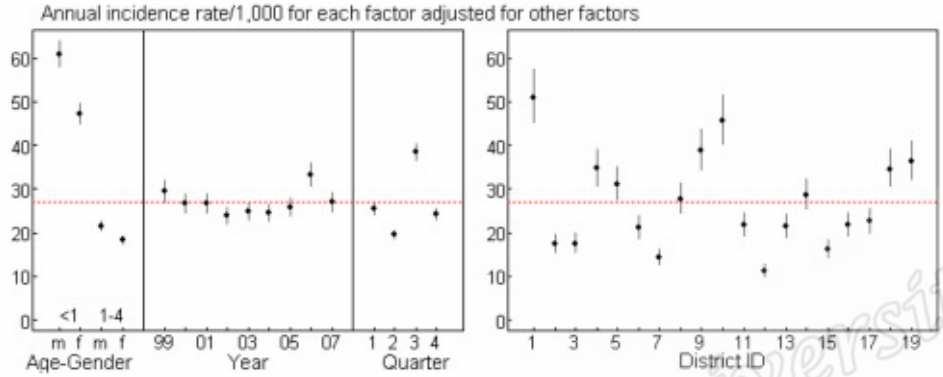

Figure 4. Annual pneumonia incidence per 1,000 among children aged 0-4 in Surat Thani for levels of each demographic factor adjusted for other factors.

These plots show that with respect to age and gender the pneumonia rates were highest among boys aged less than one year and lowest among girls aged 1-4 years. There is evidence (not so apparent in Figure 1) that the annual trend peaked in 2006 after decreasing slightly from 1999 to 2002, but then decreased again in 2007. The seasonal effect was more pronounced, high over the September quarter and low in the June quarter. There are also notable differences in incidence rates between districts, with Surat Thani City (district 1) having the highest rate.

Figure 5 shows a thematic map of the

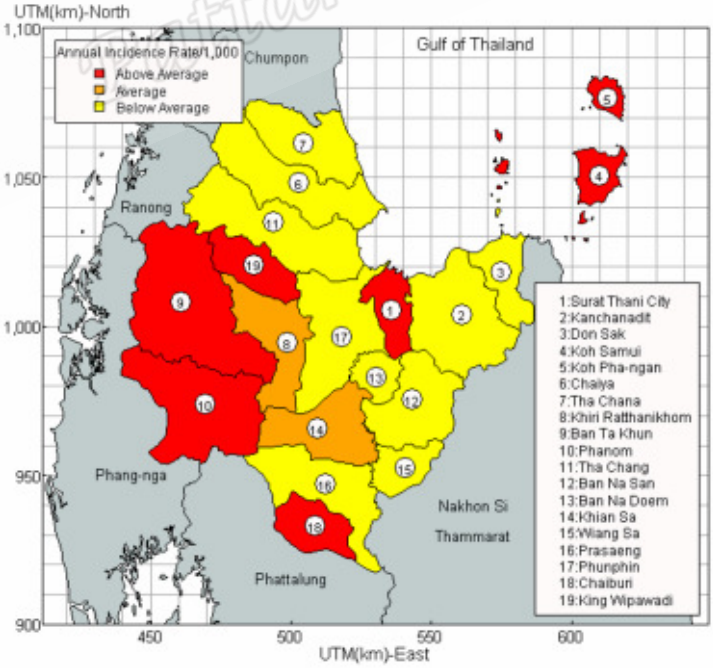

Figure 5. Map of Surat Thani showing annual pneumonia incidence patterns of children aged 0-4 years during 1999-2007 adjusted for age, gender, season and year.

adjusted annual incidence by districts, using the confidence intervals plotted in Figure 4 to classify districts as above the mean (darkest shade), below the mean (lightest shade) or not evidently different from the mean (intermediate shade), for each province studied.

Higher pneumonia incidence occurred in seven districts. These comprise (a) the urban district Surat Thani City, (b) the resort islands Samui and Pha-ngan, and (c) the mountainous districts Ban Ta Khun, Phanom, King Wipawadi, and Chaiburi. Two districts - also mountainous - had average disease rates, and the ten remaining districts had lower than average rates.

5. DISCUSSION

The log-linear regression modeling revealed that pneumonia incidence rates of hospital-reported cases among young children in Surat Thani province of Thailand during 1990-2007 based on aggregated data by quarter, gender, age group and district were well fitted by this simple. Add "model" = by this simple model of the 2,736 cells, 550 (20%) had zero counts. Each zero was replaced by a constant (we used 1) before log-transforming the incidence rates, using a similar method to that Sriwattanapongse and Kuning [24] used to identify hospital diagnosed malaria incidences in North-western districts of Thailand. To allow for zero counts in their model, they replaced zeros by 0.25 before log-transforming. This method gave a substantially higher r-squared than an alternative based on adding a constant to each count before making the log-transformation. In a similar vein, Warton [16] conducted studies with ecology data. Abundance data have many zeros (often 50–80 per cent of all values), and zero-inflated count distributions are often used to model the high frequency of zeros in such data. Examples include the "Hurdle" model developed by Mullahy [25], and the zero-inflation model introduced by Lambert [26]. and so on. However, for these data the negative binomial model gives a poor fit, whereas simply taking logarithms of the incidence rates after replacing the zeros by an appropriate constant and fitting a simple linear regression model with normal errors gives a quite satisfactory fit, even though the proportion of zero counts is substantial.

The reason for this finding may be that the standard negative binomial models (including their zero-inflated extensions) fail to cover the range of overdispersion situations that commonly occur in practice, particularly for biological data. For the data used in this study, the estimated relation between the variance and the mean of the incidence rates based on the 76 samples of 36 quarterly rates for each combination of gender, age group and district was $variance = 1.88 \times mean^{1.43}$. For the conventional negative binomial generalized linear model the corresponding relation is $variance = mean \times (1+mean/\theta)$ as in Equation (1), and for the log-normal model the relation is $variance = \exp(\sigma^2-1) \times mean^2$ where σ is the standard deviation of the underlying normal error distribution. Using statistical theory it can be shown that if the fourth root transformation as suggested by Clarke and Warwick [27] for biological counts is used, the relation takes the asymptotic form $variance = \text{const} \times mean^{1.5}$ for large values of the mean, which is close to what was found for the pneumonia incidence rates. While using a root transformation in preference to logarithms avoids the need to omit or adjust zero counts before taking the transformation, such an adjustment may still be needed to provide a satisfactory fit.

The log-normal model could possibly be improved further by incorporating weights into the linear regression model, with the weights increasing with the number of disease

counts in a cell (see, for example, Faraway [5]). The advantage of such a model is that the standard errors of the district effects would then take the sample sizes into account. However, this investigation is beyond the scope of the present paper.

A further advantage of the log-linear model is that software for handling spatial correlation as well as time series autocorrelation is more easily available.

The model should be useful for health planning in countries such as Thailand where routine epidemiological reports of pneumonia and other diseases cases are provided at the district level, The model can also be used to identify unusually high incidence rates within the season of their occurrence, and thus enable health authorities to reduce the severity of ensuing epidemics by putting preventative measures in place for the demographic group at risk.

ACKNOWLEDGEMENTS

This study was funded by the Graduate School, Prince of Songkla University.

The authors would like to thank the Ministry of Public Health for providing the data. We are grateful to Prof. Don McNeil and David Broadfoot for his assistance and helpful guidance, Assoc. Prof. Surin Khanabsakdi and referees for helpful comments and suggestions.

REFERENCES

[1] Crawley M., *Statistics an introduction using R*, 1st Edn., John Wiley and Sons: New York, 2005.

[2] Venables W.N. and Ripley B.D., *Modern Applied Statistics with S.*, 4th Edn. Springer, 2002.

[3] Jansakul N. and Hinde J.P., Linear mean-variance negative binomial models for analysis of orange tissue-culture data, *Songklanakarin J. Sci. Technol.,* 2004; **26**: 683-696.

[4] Bulsara M., Holman C., Davis E. and Jones T., Evaluating risk factors associated with severe hypoglycemia in epidemiology studies-what method should we use?, *Diabetic Medicine*, 2004; **21**: 914-919.

[5] Faraway J.J., *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models,* 1st Edn., Chapman and Hall, London, 2006.

[6] Cameron A.C. and Trivedi P.K., *Regression Analysis of Count Data*, 1st Edn., Cambridge University Press: New York, 1998.

[7] Cameron A.C. and Trivedi P.K., *Microeconometrics: Methods and Applications*, 1st Edn., Cambridge: Cambridge University Press, 2005.

[8] Dalrymple M., Hudson I. and Ford R., Finite mixture, zero-inflated Poisson and hurdle models with application to SIDS, *Stat. Data An.*, 2003; **41**: 491-504.

[9] Diggle P.J., Heagerty P., Liang K.Y. and Zeger S.L., *Analysis of Longitudinal Data*, 1st Edn., Oxford University Press: United Kingdom, 2002.

[10] Kissling W.D. and Cart G., Spatial autocorrelation and the selection of simultaneous autoregressive models, *Global Ecol. Biogeogr.,* 2008; **17**: 59-71.

[11] Dormann C., Effects of incorporating spatial autocorrelation into the analysis of species distribution data, *Global Ecol. Biogeogr.*, 2007; **16**: 129-138.

[12] Engle R.F., *ARCH: selected readings*, 1st Edn., Oxford University Press, 1995.

[13] Tukey J.W., *Exploratory Data Analysis*, 1st Edn., Addison-Wesley Reading, Massachusetts, 1977.

[14] Miller R.J., *Beyond ANOVA, Basics of Applied Statistics*, 1st Edn., John Wiley and Sons: New York, 1986.

[15] Clarke K.R., and Warwick R.M., *Change in marine communities: an approach to statistical analysis and interpretation*, 1st Edn., Plymouth Marine Laboratory, United Kingdom, 1994.

[16] Warton D.I., Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. *Environmetrics*, 2005; **16**: 275-289.

[17] Wakefield J., Disease mapping and spatial regression with count data, *Biostatistics*, 2007; **8**: 158-183.

[18] Welsh A.H., Cunningham R.B. and Chambers R.L., Methodology for estimating the abundance of rare animals: seabird nesting on North East Herald Cay, *Biometrics*, 2000; **56**: 22-33.

[19] Tongkumchum P. and McNeil D., Confidence intervals using contrasts for regression model, *Songklanakarin J. Sci. Technol.*, 2009; **31**: 151-156.

[20] Kanlayanaphotporn J., Brady M.A., Chantate P., Chantra S., Siasiriwattana S., Dowell S. and Olsen S.J., Pnuemonia surveillance in Thailand : current practice and future needs. *Southeast Asian J. Trop. Med. Public. Health,* 2004; **35**: 711-716.

[21] Suwanjutha S., Ruangkanchanasetr S., Chantarojanasiri T. and Hotrakitya S., Risk factors associated with morbidity and mortality of pneumonia in Thai children under 5 years, *Southeast Asian J. Trop. Med. Public Health,* 1994; **25**: 60-66.

[22] Reechaipichitkul W. and Tantiwong P., Clinical features of community-acquired pneumonia treated at Srinagarind Hospital, Khon Kaen, Thailand, *Southeast Asian J. Trop. Med. Public Health,* 2002; **33**: 355-361.

[23] Venables W. and Smith D., The R Development Core Team. An introduction to R: notes on R: a programming environment for data analysis and graphics version 2.8 (2008-2-08). Available from: http://cran.r-project.org/doc/manuals/R-intro.pdf.

[24] Sriwattanapongse W. and Kuning M., Modeling Malaria Incidence in North-Western Thailand, *Chiang Mai J. Sci.* 2009; **36**: 403-410.

[25] Mullahy J., Specification and testing of some modified count data models. *J. Econometrics,* 1986; **33**: 341-365.

[26] Lambert, D., Zero-inflated Poisson regression, with an application to defects in manufacturing, *Technometrics*, 1992; **34**: 1-14.

[27] Clarke K. and Warwick R., *Change in Marine Communities: an approach to statistical analysis and interpretation*, 2nd Edn., PRIMER-E: Plymouth, United Kingdom, 2001.

MODELIMG THE INCIDENCE OF TUBERCULOSIS IN THE SOUTHERN PROVINCES OF THAILAND

Noodchanath Kongchouy[1], Sampurna Kakchapati[2] and Chamnein Choonpradub[2]

[1]Department of Mathematics, Faculty of Science, Prince of Songkla University, Hat Yai, Songkhla; [2]Department of Mathematics and Computer Science, Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, Pattani, Thailand

Abstract

The aim of this study was to examine the trend, seasonal and geographic effects on tuberculosis (TB) incidence in the fourteen southern provinces of Thailand from 1999 to 2004. Data were obtained from National Notifiable Disease Surveillance (Report 506), Ministry of Public Health. The joint effects of gender-age, quarterly season and location on the TB incidence rates were modeled using both negative binomial distributions for the numbers of cases and log-linear distributions for the incidence rate, and these models were compared. The linear regression models provided a good fit, as indicated by residual plots and the R^2 (0.64). The model showed that males and females aged less than 25 years had similar risks for TB in the study area. Both sexes had their risk increased with age but to a much greater extent for men than women, with the highest rate noted in males aged 65 years and over. There was no evidence of a trend in the annual incidence of TB during 1999-2004, but the incidence has a significant season variation with peaks in the first quarter over the six year period. There were also differences in the incidence rate of TB both within and between provinces. The high risk areas were in upper western and lower southern parts of the region. The log-linear regression model can be used as a simpler method for modeling TB incidence rates. These findings highlight the importance of selectively monitoring geographic location when studying TB incidence patterns.

Key words: *log-linear models, negative binomial model, tuberculosis incidence*

INTRODUCTION

The scale of the global *tuberculosis* (TB) epidemic is very substantial. About a third of the world's population is infected with *Mycobacterium tuberculosis*, with an estimated nine million new cases and two million deaths occur yearly due to the disease. It is a leading cause of death among people who are HIV-positive (World Health Organization, 2009b). It affects the mostly economically productive age group comprising adults aged 15 to 54 years, with males being disproportionately affected. The male/female ratio among newly detected cases is 2:1. The impact of TB is particularly evident in Asia (South-East Asia and Western Pacific Regions) and Africa Approximately 86% of all TB cases reported worldwide occurs in these regions, where 60% of the world's populations live. The South East Asia region carries one third of global burden of TB. However, TB epidemic in Asia could gradually worsen by variations in the quality and coverage of various TB control interventions, population demographics, urbanization, changes in socio-economic standards, HIV and, more recently, emerging drug resistance (World Health Organization, 2009a).

Correspondence: E-mail: cchamnein@bunga.pn.psu.ac.th, noodchanath.k@psu.ac.th

TB is a serious public health problem in Thailand. The country is ranked 18th on the list of 22 high TB-burden countries. The prevalence of TB was estimated at 192 per 100,000 persons for all forms in 2007, with an incidence rate of 62 new smear-positive cases per 100,000 (World Health Organization, 2009a). Several studies suggest marked regional differences in incidence of TB in Thailand (Wibulpolprasert, 2007). Recent research based on National Tuberculosis Program (NTP) surveillance data from 2001 to 2005 shows higher numbers of cases and deaths from TB occurred in the southern and northern provinces of Thailand due to low treatment success rates (Thongraung, 2008).

The Thailand Health Profile Report indicates that the number of TB cases declined between 1985 and 1989 but increased slightly from 1990 to 2005. The tuberculosis prevalence has risen by 2% each year during the past five years and there was no tendency to decline during the period 1995-2002 (Wibulpolprasert, 2007). Thus, TB remains a major infectious disease in Thailand (Phomborphub *et al*, 2008; Jittimanee *et al*, 2009; Wibulpolprasert, 2007; Thongraung, 2008).

Public health officials are often required to evaluate disease incidence in the country. They compare the standardized disease incidence rate within the area and time frame for planning various interventions. Statistical modeling may provide the necessary quantitative framework for investigating main issues related to disease. Since the 1960s, statistical models have been used to understand tuberculosis transmission dynamics and to predict the effects of different interventions (Waaler *et al*, 1962). The models have been used to understand the impact of tuberculosis control strategies of tuberculosis, (Legrand *et.al,* 2008) HIV and TB joint epidemic (Willam *et.al*, 2005; Baclear, 2008) and spatial and temporal variation of TB incidence (Uthman, 2008; Nunes, 2007) and drug-resistance tuberculosis (Castillo-Chavez, *et.al* 1997; Dye, *et.al* 1998).

The purpose of this study to examine the trend, seasonal and geographic effects on TB incidence in the fourteen southern provinces of Thailand from 1999 to 2004. In our study we used two alternative statistical models, a negative binomial generalized linear model and a simple linear model after logarithmic transformation of incidence rates, and we compared the results obtained from applying these methods. After fitting the models, we analyzed the findings based on the best fitted model.

MATERIALS AND METHODS

Study area and data source

The study design was a retrospective analysis of reported TB cases obtained from the National Notifiable Disease Surveillance (Report 506), Bureau of Epidemiology, Ministry of Public Health, Thailand. The study population included all such cases in the six-year period 1999-2004, for the fourteen southern provinces of Thailand. These data are available in computer files with individual records for disease cases and fields comprising characteristics of the subject and the disease, including dates of sickness and diagnosis, the subject's age, gender, address, and the severity of the illness, including date of death for mortality cases. The resident population denominators used to compute incidence rates were obtained from the Population and Housing Census of 2000

undertaken by the National Statistics Office of Thailand. To simplify the effect of location of residence when calculating incidence rates, smaller contiguous districts in each province were grouped together to form 32 super-districts containing populations ranging from 161,210 to 360,000, as shown in Table 1, where they are listed in order of geographical location from north to south (keeping super-districts within the same province together) with their 2000 census populations.

Table 1: Definitions and populations of super-districts

Code	Super-district	Population	Code	Super-district	Population
1	Chumpon North	246,279	17	Trang South	264,412
2	Chumpon South	199,927	18	Trang City	330,698
3	Ranong	161,210	19	Pattalung City	251,029
4	SuratThani North West	206,713	20	Pattalung West	247,442
5	SuratThani City	241,373	21	Songkla North Coast	329,133
6	SuratThani East	168,801	22	Songkla City	340,096
7	SuratThani South	252,523	23	Hat Yai	350,776
8	Phang-nga	234,188	24	Songkla South West Coast	235,657
9	NakornST North	240,392	25	Satun	247,875
10	NakornST North West	220,430	26	Pattani City-West	332,757
11	NakornST City	267,560	27	Pattani East	263,228
12	NakornST West	273,583	28	Yala City West	211,180
13	NakornST South Coast	194,771	29	Yala South	204,357
14	NakornST South West	323,075	30	Narathiwat Central	224,646
15	Krabi	336,210	31	Narathiwat West	214,412
16	Phuket	249,446	32	Narathiwat South	223,292

Age, gender, residential area (by super-district), quarter of year and year were selected as the explanatory variables in studying the incidence rates of TB. Age was divided into four groups (0-24, 25-39, 40-59 and 60+ years) and age and gender were combined together to form eight gender-age groups. The year was divided into four 'quarter' periods, defined as January-March, April-June, July-September and October-December, giving twenty-four quarter periods over the six years.

Statistical methods

We first calculated the disease incidence in cells defined by gender-age group *i*, region *j*, quarterly season *q* and year *t* as the ratio of the number of reported cases n_{ijqt} to P_{ij}, the corresponding population at risk in 1000s.

The negative binomial generalized linear model (GLM) (Venables and Ripley, 2002) is an extension of the Poisson regression model that allows for over-dispersion. If λ_{ijqt} denotes the mean incidence rate in gender age group *i*, quarter *q*, super *j* and year *t*, an additive model with this distribution is expressed as

$$\ln(\lambda_{ijqt}) = \ln(P_{ij}) + \mu + \alpha_i + \beta_q + \gamma_j + \eta_t \qquad (1)$$

The terms α_i, β_q, γ_j and η_t represent gender-age group, season, super-district and year effects that sum to zero so that μ is a constant encapsulating the overall incidence. The variance of this distribution is $\lambda_{ijqt}(1+\lambda_{ijqt}/\theta)$ with the Poisson model arising in the limit as $\theta \to \infty$. The model fit is assessed by comparing deviance residuals with normal quantiles, and it is also informative to plot observed counts and appropriately scaled incidence rates against corresponding fitted values based on the model.

The model also gives adjusted incidence rates for each factor of interest, obtained by suppressing the subscripts in Equation (1) corresponding to the other factors and replacing these terms with a constant satisfying the condition that the sum of the disease counts based on the adjusted incidence rates matches the total. Sum contrasts (Venable and Ripley, 2002; Tongkumchum and McNeil, 2009) were used to obtain confidence intervals for comparing the adjusted incidence rates within each factor with the overall incidence rate.

The alternative additive log-linear model for incidence rates with normally distributed errors is

$$\ln\left(\frac{n_{ijqt}}{P_{ij}}\right) = y_{ijqt} = \mu + \alpha_i + \beta_q + \gamma_j + \eta_t . \qquad (3)$$

Since some cells had no reported cases, thus disallowing log-transformation, we replaced the zeroes by a suitably chosen small constant *d*, without changing any values of n_{ijqt} greater than 0.

The model fit was assessed by plotting standardized residuals against normal quantiles, and by plotting observed counts and appropriately scaled incidence rates against corresponding fitted values based on the model, and also by using the r-squared to see how much of the variation in the data was accounted for by the model.

To obtain estimates of incidence rates in cells we used the formula

$$\hat{r}_{iqjt} = \exp(\hat{y}_{iqjt} + c) , \qquad (4)$$

where $\hat{y}_{ijqt}$ is the fitted value of y_{ijqt} and *c* is a constant chosen to match the total number of observed cases with the total given by the model.

After fitting the model, incidence rates for levels of each factor adjusted for other factors were calculated similarly. Standard errors for these adjusted incidence rates were obtained by using sum contrasts to compare the incidence rates for each level of a factor with the overall mean incidence rate. By using this method, the pattern of TB was identified for each factor. Since the confidence intervals for factor-specific incidence rates obtained in this way divide naturally into three groups according to their location entirely above the mean, around the mean, or entirely below the mean, we used this trichotomy to create thematic maps of districts according to their estimated tuberculosis annual incidence rates.

All models were fitted by maximum likelihood and the analysis was undertaken using purpose-written code in R version 2.7.1 (R Development Core Team, 2008).

RESULTS

The results of the model fitting are shown in Figure 1. The left and right upper panels show plots of the deviance residuals against the normal quantiles based on the negative binomial model (1) (upper panels) and the log-linear alternative (2) with zero counts replaced by 1 (lower panels). The left and right lower panels show plots of observed counts and observed annual incidence rates per 1000 versus corresponding fitted values using the linear model. Clearly, the residuals plot from the negative binomial model did not fit the data as well as the linear model on the log-transformed incidence rates.

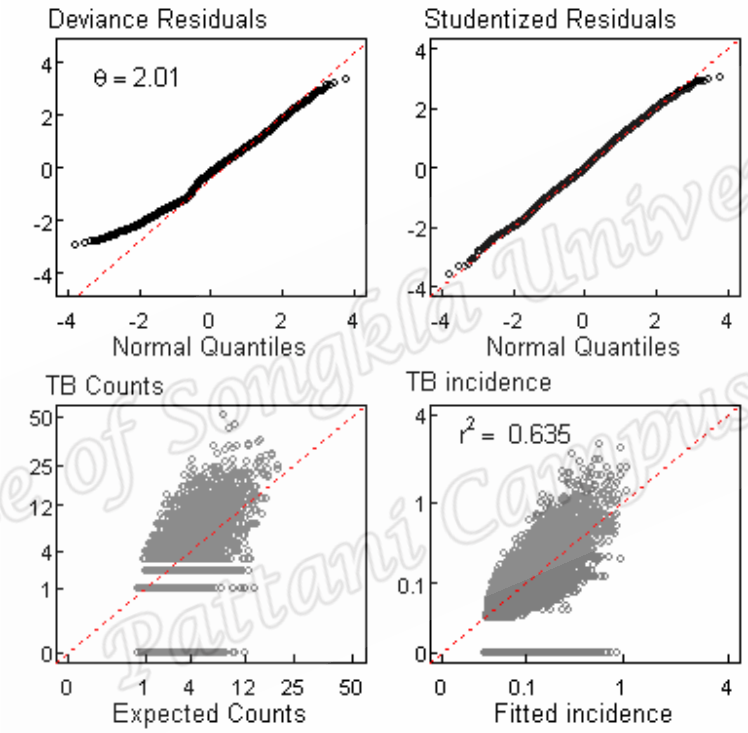

Figure 1: Diagnostic residual plots for negative binomial (top left) and log-normal (top right) models, and plots of counts and incidence rates for the log-normal model (lower panels) for tuberculosis incidence rates in Southern Thailand super-districts. Note that Theta (θ) is a dispersion parameter in the Negative binomial distribution.

Figure 2 shows 95% confidence intervals of annual tuberculosis incidence rate by gender-age group (upper left panel), year and season (upper right panel) and super-district (lower panel), each adjusted for the effects of the other two factors in the model based on the log linear model. The dotted horizontal lines on each graph represent the overall mean annual incidence rate (0.45 per 1,000). The dotted vertical lines in the lower panel separate the fourteen provinces of Southern Thailand. The male to female incidence rate ratio was calculated to be 2.16.

For those aged below 25 years, no gender difference was found. But there was a notable difference in incidence between males and females aged above 25 years with the highest

difference observed in the older than 60 age group (0.87). There was no evidence of a trend in the incidence of TB from 1999 to 2004. Higher incidences occurred in each first quarter, with lower rates in the second half of each year.

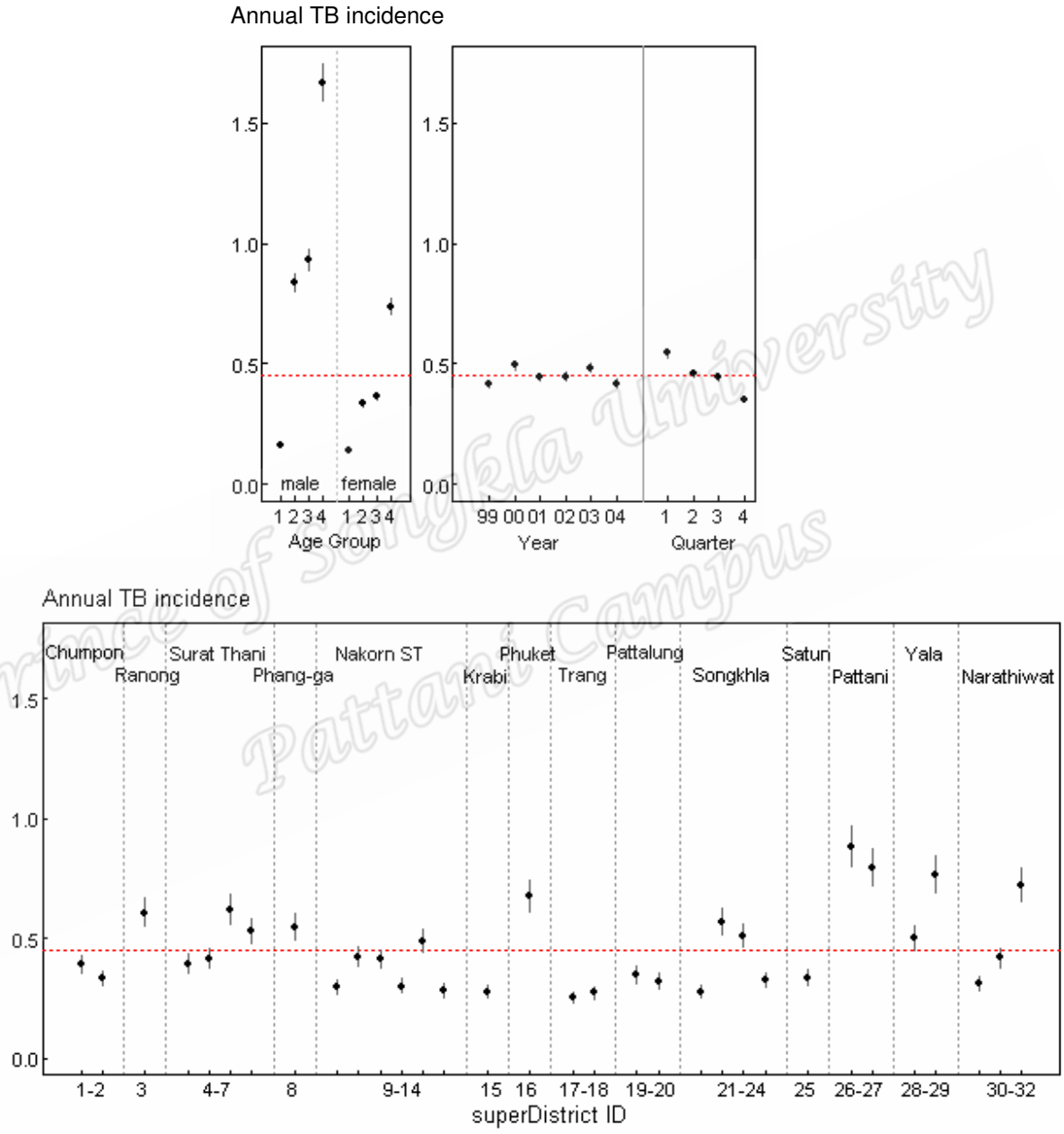

Figure 2: Annual TB incidence/1000 in southern Thailand by each demographic factor adjusted for other factors, with 95% confidence intervals for differences from the mean.

Figure 3 shows a thematic map of the adjusted annual incidence by super-district, using the confidence intervals plotted in Figure 2 to classify districts as above the mean (darkest shade), below the mean (lightest shade) or not evidently different from the mean (intermediate shade). The map shows that the higher TB incidence occurred in all the super districts of Pattani, Ranong, Phang-nga, Phuket and Yala provinces. Similarly Hat Yai and Songkla City had higher TB incidence. Three adjoining super-districts of Phang-nga, Ranong and Phuket provinces also had higher than average incidence rates.

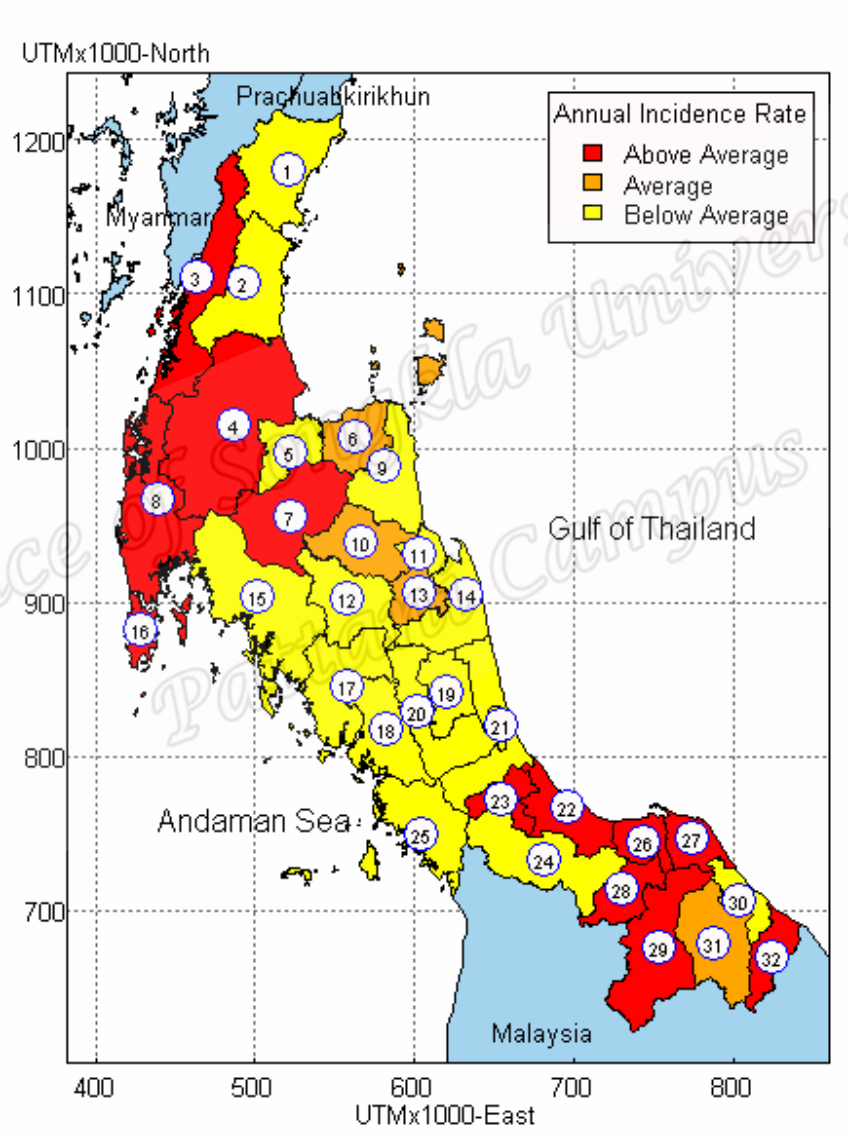

Figure 3: TB incidence patterns in the super-districts of southern of Thailand

DISCUSSION

This study used statistical modeling of TB incidence to gain insights into its dependence on several demographic and geographic factors (age, gender, quarterly season, year, and super-district) in the southern region of Thailand from 1999 to 2004. When the dependent variable is the disease count, Poisson and negative binomial generalized linear models are usually considered to be most statistically appropriate. The Poisson distribution assumes events to be independent and does not account for clustering, over-dispersion or serial correlation. A negative binomial GLM is an extension of the Poisson regression model that allows for over-dispersion. Negative binomial models containing gender, age, quarterly season and super-district as factors were fitted to the disease incidences. However, for these data the negative binomial model gave a poor fit as indicated by the residual plot. Linear regression models containing gender, age, quarterly season and super-district as factors were fitted to the log-transformed disease incidences, with zero cell counts replaced by a constant before log-transformation. This model provided an acceptable fit with an adjusted R-squared of 0.64. The overall annual incidence of TB in the 14 southern provinces of Thailand was 0.45 per 1000 population.

No gender difference in the incidence rate of TB was found in people aged below 25 years. However, substantial differences between males and females were observed among those aged a 25 years or over. This is consistent with age and gender patterns found in other recent studies. Similar studies by the World Health Organization (WHO) in South East Asia have shown that the male: female ratio of the incidence rates of TB registered in 2006 was lower in lower age groups, gradually increasing in older age groups, with cases among men becoming 3.5 times higher than among women in higher age groups (World Health Organization, 2009b). Epidemiological findings indicate that in most settings, TB incidence rates are higher for males at all ages except in childhood, when they are higher for females. Studies have reported that gender differences in TB incidence begin to appear between 10 and 16 years of age and incidence rates remain higher for males than females thereafter (World Health Organization, 2009b).

In the six-year period of the current study, 1999 to 2004, there was no evidence of any trends in the incidence of TB. The highest incidence was observed in year 2000 followed by 2003. This finding was consistent with Thailand Health Profile that also indicate high TB incidence in 2000. However, the incidence rates were higher in each first quarter (January to March), with lower rates in the second half of each year. Studies from the UK (Douglas *et al*, 1996) and Spain (Rios *et al*, 2000) have also shown seasonal variations in TB rates and higher levels of notifications over summer months. The mechanism for the increase in notification rates in the summer months is unclear, although it has been hypothesized that seasonal fluctuations in vitamin D serum levels may contribute to impair host defense mechanisms against TB bacilli (Davies, 1997). However, studies in Russia have shown a contrary result, with the hospital admissions of TB increasing in the winter months (from October to March) and declining in the warmer months (from April to September) (Atun *et al*, 2005).

A striking finding in this study is that the highest TB incidences occurred in all of the super-districts of Pattani and Yala and in one super-district of Narathiwat province (the three southernmost provinces). However, the adjoining super-districts of Ranong, Phang-nga and Phuket provinces (west coast) to the north also showed higher incidences of TB. These findings are consistent with a recent study which found higher incidences of TB in the west coast and southernmost provinces (Jittimanee *et al*, 2009). Possible reasons for higher than average incidence rates in these provinces were not investigated in our study, and could be due to HIV, which contributes to high incidence of TB. In Thailand, high TB case-fatality rates have been reported in areas with high HIV rates in the general population (Wibulpolprasert, 2007). Besides this, urban migration and cross-border population movements are also contributing factors to TB epidemics in these areas (World Health Organization, 2005). Our study findings also revealed that high TB incidence occurred in border provinces in Thailand including Ranong, Songkhla, Yala and Narathiwat. Lower socio-economic status, drug addiction and increasing poverty seem to be linked with the reemergence of TB (Carolyn, 1996) and further research in this area is needed.

Perhaps the major limitation of this study is that the precise burden of TB in Southern provinces Thailand is unknown as that the surveillance data from the Ministry of Public Health is known to be under-reported for major infectious diseases (Lumbiganon *et al*, 1990; Saengwonloey *et al*, 2003; Intusoma *et al*, 2008). It is not based on registration in TB clinics in most public hospitals. Despite this limitation, the findings should reflect the relative patterns of TB incidence with respect to demographic, spatial and temporal variation for the southern provinces of Thailand, even though the absolute extent of these incidence rates is inaccurate.

ACKNOWLEDGEMENTS

We thank the Ministry of Public Health, for providing the data. This study was funded by the Graduate School, Prince of Songkla University. We are grateful to Prof. Don McNeil who supervised our research and to Dr Petchawan Pungrassami and the referees for helpful comments and suggestions.

REFERENCES

Atun RA, Samyshkin YA, Drobniewski F, et al. Seasonal variation and hospital utilization for tuberculosis in Russia: hospitals as social care institutions. *Eur J Pub Health* 2005; 15: 350–4.

Bacaër N, Ouifki R, Pretorius C, et al. Modeling the joint epidemics of TB and HIV in a South African township. *J Math Biol* 2008; 57: 557–93.

Carolyn S. Healthy Cities or Unhealthy Island? The Health and Social Implications of Urban Inequality. Environ Urban 1996; 8: 9–30.

Castillo-Chavez C & Feng Z. To treat or not to treat: the case of Tuberculosis. *J. Math. Biol* 1997; 35: 629-56

Davies PD. Seasonality of tuberculosis. *Thorax* 1997; 52: 398.

Douglas AS, Strachan DP, Maxwell JD. Seasonality of tuberculosis: the reverse of other diseases in the UK. *Thorax* 1996; 51: 944–6.

Dye C, Garnett GP, Sleeman K & Williams BG. Prospects for worldwide Tuberculosis control under the WHO DOTS strategy: Directly Observed Short Course Therapy. *Lancet* 1998; 352:1886-91

Intusoma U, Sornsrivichai V, Jiraphongsa C, et al. Epidemiology, clinical presentations and burden of rotavirus diarrhea in children under five seen at Ramathibodi Hospital, Thailand. *J Med Assoc Thai* 2008; 91: 1350–5.

Jittimanee S, Vorasingha J, Mad-asin W, et al. Tuberculosis in Thailand: epidemiology and program performance, 2001–2005. *Int J Infection Dis* 2009; 13: 436–42.

Legrand J, Sanchez A, Le Pont E, et al. Modeling the impact of tuberculosis control strategies in highly endemic overcrowded prisons. *PLoS One* 2008; 3:e2100.

Lumbiganon P, Panamonta M, Laopaiboon M, et al. Why are Thai official perinatal and infant mortality rates so low? *Int J Epidemiol* 1990; 19: 997–000.

Nunes C Tuberculosis incidence in Portugal: spatiotemporal clustering. *Int J Health Geogr* 2007; 6–30.

Phomborphub B, Pungrassami P, Boonkitjaroen T. Village health volunteer participation in tuberculosis control in Southern Thailand. *Southeast Asian J Trop Med Public Health* 2008; 39: 542–8.

R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2008. [Cited 2008 Feb 8]. Available from:URL: http://www.R-project.org

Rios M, Garcia JM, Sanchez JA, et al. A statistical analysis of the seasonality in pulmonary tuberculosis. *Eur J Epidemiol* 2000; 16: 483–8.

Saengwonloey O, Jiraphongsa C, Foy H. Thailand report: HIV/AIDS surveillance 1998. *J Acquir Immune Defic Syndr* 2003; 32(Suppl 1): S63–7.

Thongraung W, Chongsuvivatwang V, Pungrassamee P. Multilevel factors affecting tuberculosis diagnosis and initial treatment. *J Eval Clin Pract* 2008; 14: 378–84.

Tongkumchum P. and McNeil D., Confidence intervals using contrasts for regression model, *Songklanakarin J Sci Technol* 2009; 31: 151-156.

Uthman OA. Spatial and Temporal Variations in Incidence of Tuberculosis in Africa, 1991 to 2005. *World Health Popul* 2008; 10(2):5-15.

Venables WN, Ripley BD. Modern Applied Statistics with S. 4th ed. Springer; 2002: 139–208.

Waaler HA, Geser A, Andersen S. The use of mathematical models in the study of the epidemiology of tuberculosis. *Am J Public Health Nations Health* 1962; 52: 1002–13.

Wibulpolprasert S. Thailand health profile 2005–2007. Bangkok: Printing Press, The War Veterans Organization of Thailand; 2007.

Williams BG, Granich R, Chauhan LS, Dharmshaktu NS, Dye C. The impact of HIV AIDS on the control of tuberculosis in India. *PNAS* 2005; 102: 9619-24.

World Health Organization. Overview of Thai/Myanmar boarder health situation 2005. Bangkok; WHO country office for Thailand, 2005.

World Health Organization. Global tuberculosis control 2009: epidemiology, strategy, financing: WHO report 2009. Geneva: World Health Organization; 2009b.

World Health Organization. Tuberculosis control in the South East Asia Region: WHO annual report 2009. Geneva: World Health Organization; 2009a.

World Health Organization. Overview of Thai/Myanmar Boarder Health Situation. WHO Thailand Boarder Health Program 2005.

Regression-Based Modeling of macrobenthic fauna density in the Middle Songkhla Lake, Thailand

Uraiwan Sampantarak [1,*], Noodchanath Kongchouy [2], Saowapa Angsupanich [3]

[1] *Inland Fisheries Research and Development Bureau, Department of Fisheries, Bangkok 10900, Thailand*
[2] *Department of Mathematics, Faculty of Science, Prince of Songkla University, Songkhla 90112, Thailand*
[3] *Department of Aquatic Science, Faculty of Natural Resources, Prince of Songkla University, Songkhla 90112, Thailand*

Abstract

This study examined distributional patterns of macrobenthic fauna assemblages in relation to environmental characteristics in the middle of Songkhla Lake, Thailand. Macrobenthic fauna and water quality parameters including sediment characteristics were obtained from nine sampling sites at bimonthly intervals from April 1998 to February 1999. Factor analysis was used to define five predictors including three composite variables based on salinity, physical sediment characteristics, and physico-chemical properties of water and sediment, together with total suspended solids and dissolved oxygen as single variables. A multivariate multiple regression model (MMR) was used to examine relationships between these predictors and the densities of twenty four selected macrobenthic families with greater than 35% occurrence. To remove skewness, the densities were log-transformed before fitting the model. Results were compared with those obtained using canonical correspondence analysis. It was found that MMR could be used as additional or alternative method to analyse relationship between environmental variables and abundance of benthic organisms in coastal ecosystem.

***Keywords**:* Macrobenthic fauna, Multivariate Multiple Regression model, Factor analysis, Canonical Correspondence Analysis, Tropical lagoon

1. Introduction

Macrobenthic fauna are recognized as sensitive indicators of environmental disturbance (Pearson and Rosenberg, 1978; Rygg, 1985; Engle et al., 1994; Weisberg et al., 1997; Borja et al., 2000; Ranasinghe et al., 2004). They have limited mobility. Many of them are unable to avoid adverse conditions brought about by natural stresses or human impacts. Moreover, their relative longevity, with many species having life spans in excess of two years, allows them to integrate responses to environmental processes over extended time periods (Gray et al., 1988). In addition, observed distribution of macrobenthic fauna are useful in diagnostic studies and environmental monitoring (Warwick, 1986).

Clarke and Warwick (1994) outlined the basic methods now commonly used by biological scientists for analysis of their data.

[*]Corresponding author, E-mail: uraiwan111@hotmail.com

For descriptive studies these methods include data transformation using square roots, fourth roots or logarithms (after adding 1 to cell counts or densities to handle zeros) to remove skewness, principal components analysis of covariance matrices, and ordination procedures to cluster taxa in space and time, as well as more complex multivariate analytical techniques such as dendrograms based on similarity matrices and multidimensional scaling. Measures of association in assemblage data such as the Bray-Curtis similarity index are preferred to Pearson correlation coefficients "for sound biological reasons" (Clarke et al., 2006), but such measures do not satisfy the positive-definiteness assumptions that underpin conventional multivariate statistical analysis.

For comparative studies to assess associations between species abundance outcomes and environmental predictor variables, canonical correspondence analysis (Ter Braak, 1986) is now used extensively in the biological literature (von Wehrden et al., 2009). Some important studies using this method include those reported by Rakocinski et al. (1997); Hawkins et al. (2000); Joy and Death (2000); Guerra-García et al. (2003); Hajisamae and Chou (2003); Morrisey et al. (2003); Ysebaert et al. (2003); Ellis et al. (2006); Frédou et al. (2006); Quintino et al. (2006); Anderson (2008); and Glockzin and Zettler (2008).

Although exceptions exist such as studies by Liang et al. (2002) using structural equation modeling and by Warton and Hudson (2004) using multivariate analysis of variance, multivariate multiple regression analysis is not commonly used in the biological literature for analyzing species abundance patterns. However, this method would appear to be an ideal statistical method for assessing relationships between species abundance outcomes and their environmental predictors, for the simple reason that it is the natural extension of ordinary regression analysis involving a single outcome to any number of mutually correlated outcomes such as species abundances. It is thus of interest to compare this method with its biologically preferred counterpart using common sets of biological data relating taxonomic abundances to environmental determinants, and this is the object of our study.

For this comparison we use data from a study involving macrobenthic fauna abundances and various water and sediment characteristics collected at specified locations in an estuarine lake over a period of one year reported by Angsupanich et al (2005a). The methods compared are canonical correspondence analysis (CCA) using CANOCO Version 4.5 (Ter Braak and Šmilauer, 2002) and multivariate multiple regression (MMR) using R Version 2.10.0 (R Development Team 2009).

2. Materials

Songkhla Lake is a shallow coastal lagoon, located in a tropical coastal ecosystem in Southern Thailand. It covers an area of 1040 km^2 with 20 km width and 75 km length, approximately. It is divided into three parts as the Upper Lake in Phatthalung Province, the Middle Lake between borders of Songkhla Province and Phatthalung Province, and the Lower Lake in Songkhla Province connected to the Gulf of Thailand. Some canals pour fresh water into the lake. The salinity slowly increases where the freshwater and seawater meet. Thus, the water in the Middle Lake is brackish, and becomes saltier in the area around the lake mouth (Lower Lake). The zone of interest for this study covers an area of 390 km^2 located between UTM

635000E and 660000E in the west-east direction and between UTM 840000N and 805000N in the north-south direction (Figure 1).

Angsupanich et al (2005a) collected macrobenthic fauna using a Tamura's grab (0.05 m^2) from the nine sampling stations. The assemblages were conducted with 11 replications for each station at bimonthly intervals from April 1998 to February 1999. The samples were sieved consecutively through the screens (5, 1, and 0.5 mm of sieve mesh size, respectively) and fixed in 10% Rose Bengal-formalin for later identification.

The densities of macrobenthic fauna were recorded as the number of individuals per square meter (ind m^{-2}) for each species. A total of 161 taxa of macrobenthic fauna were found and classified into 81 families. In many cases the species could not be identified exactly, so in our model the outcomes were classified by family instead of species. With nine locations and six bimonthly data study periods, we defined the occurrence for a specified family as the proportion of these 54 occasions on which at least one organism was found. We then selected the 24 families with greater than 35% occurrence (93.2% total assemblages) for data analysis.

Environmental variables comprised water depth (wDep), water temperature (wTemp), salinity (Sal), water pH (wpH), dissolved oxygen (DO), total suspended solids (TSS), with sediment pH (spH), total nitrogen content (TN), organic carbon content (OC), and soil structure (percentages of sand, silt, and clay). These were measured with three replications on the same occasions as the biotic data.

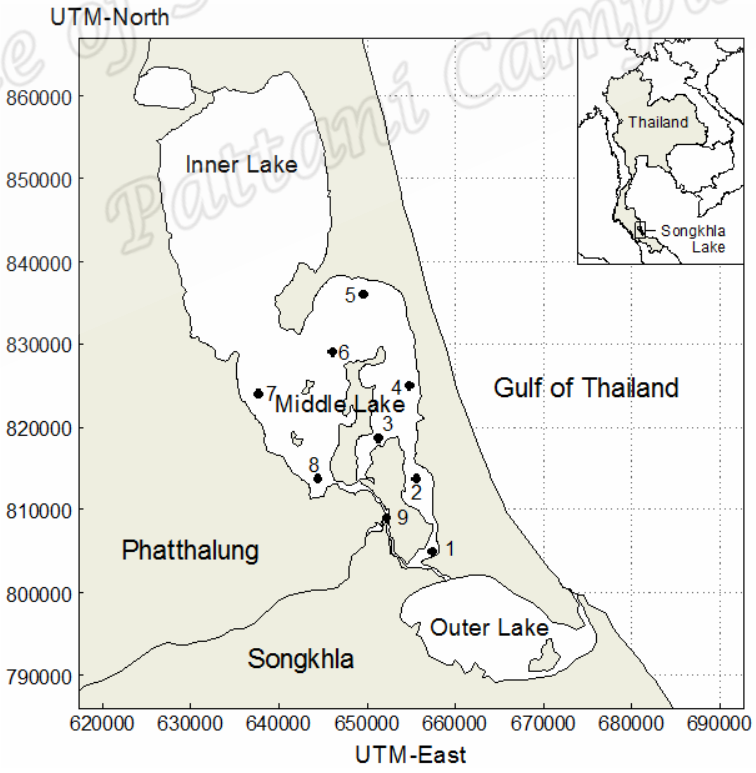

Figure 1: Songkhla Lake and sampling sites (labeled 1-9).

3. Methods

Figure 2 shows a path diagram. The nine sampling stations and six bimonthly periods were combined as 54 station-month measurement occasions. The response variable was taken as log (1 + c × density) with the multiplier c chosen to approximate normality of error distributions. The predictors comprised environmental components derived from a factor analysis together with unique variables not accommodated by the factor analysis.

Factor Analysis

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).

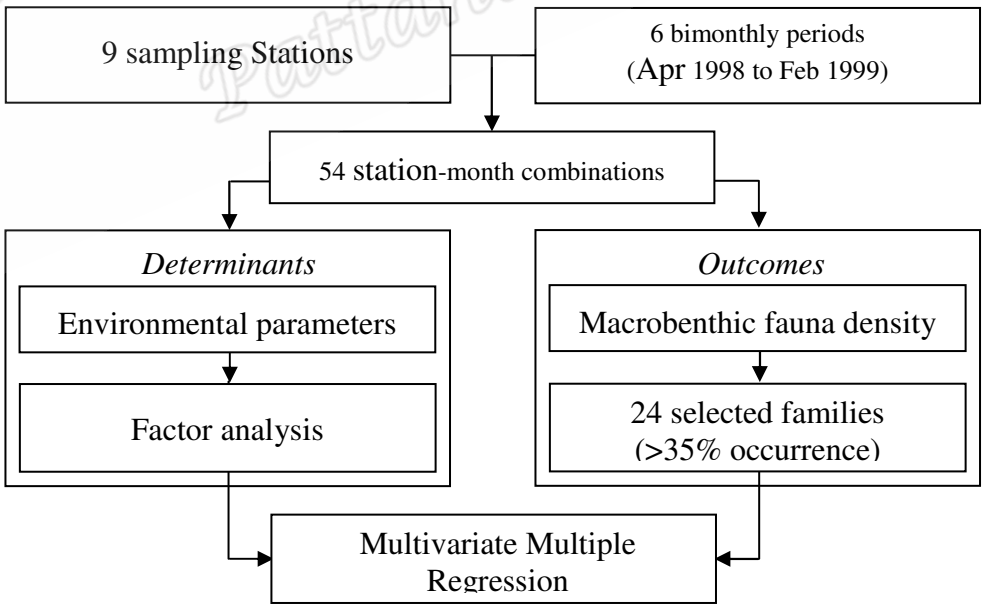

Figure 2: Steps used for the data analysis

Multivariate Multiple Regression

Multivariate multiple regression (MMR) is used to evaluate the effects of multiple predictor variables on multiple response variables. The model (Mardia et al., 1979) may be defined in matrix form , that is,

$$\mathbf{Y}_{(n \times p)} = \mathbf{X}_{(n \times q)} \mathbf{B}_{(q \times p)} + \mathbf{E}_{(n \times p)} . \quad (1)$$

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

$$y_{ij} = \mu_j + \sum_{k=1}^{q-1} \beta_j^{(k)} f_i^{(k)}. \quad (2)$$

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).

Canonical Correspondence Analysis

Assuming that the data structure comprises the **Y** and **X** matrices with rows corresponding to measurements of outcomes and predictors taken on the same occasions, canonical correspondence analysis (Ter Braak, 1986) produces a two-dimensional *biplot* comprising arrows of variable lengths and directions (*gradients*) emanating from a common origin representing the predictor variables, together with superimposed points denoting the outcome variables. The relative lengths of the arrows and the angles between them are based on the correlation matrix of the predictor variables, and the coordinates of the points are planar projections of the density outcomes, computed in such a way that their positions relative to the arrows portray their associations with the environmental predictors. The method also produces coordinate scores and p-values for the overall associations based on Monte Carlo permutation tests.

4. Results

Environmental parameters

Figure 3 plots the water characteristics in Middle Songkhla Lake from April 1998 to February 1999. The water depth varied to a lesser extent, but was also higher during the rainy season, varying with location from an average of less than 1 m at stations

four and nine to more than 2 m at station eight. The water temperature showed decreased values in the rainy season, with range 27-34°C. The salinity increased from close to zero during the rainy season (December to February) to an average close to 20 in other months. The pH of water was also lowest in December.

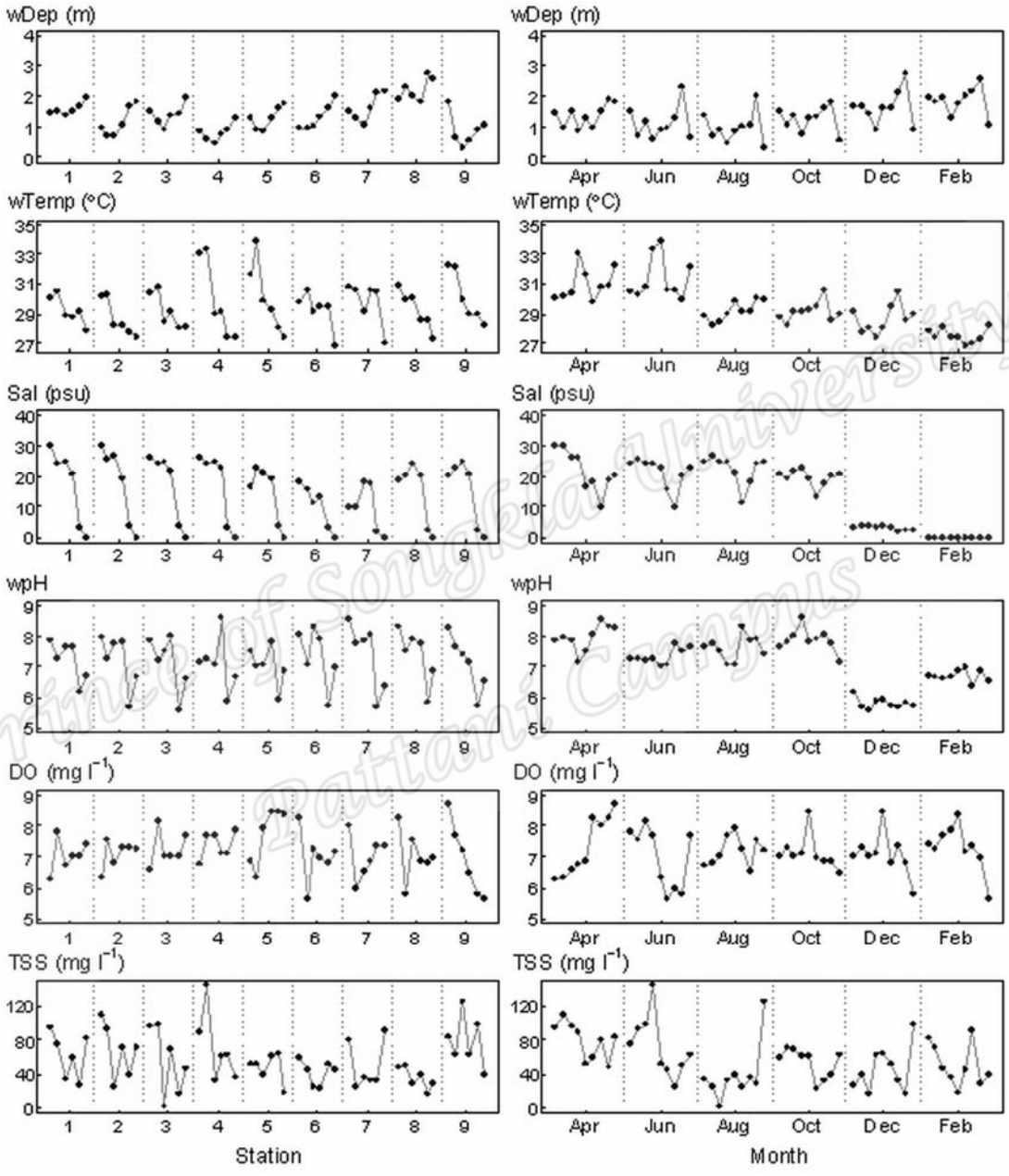

Figure 3: Water characteristics in Middle Songkhla Lake from April 1998 to February 1999 by station (left panel) and month (right panel)

Figure 4 plots sediment characteristics measured on the same occasions as the water characteristics. The total nitrogen content at each station was very low (0.02%) from October to February. The organic carbon content was relatively constant with respect to month, but showed the highest value at station nine in every month except August.

The lake bed at station six was mostly characterized by sand (mean = 84.6%) and station 9 was mostly characterized by clay (mean = 53.2%), also with high values of organic carbon. Note that the sand, silt, and clay percentages sum to 100%.

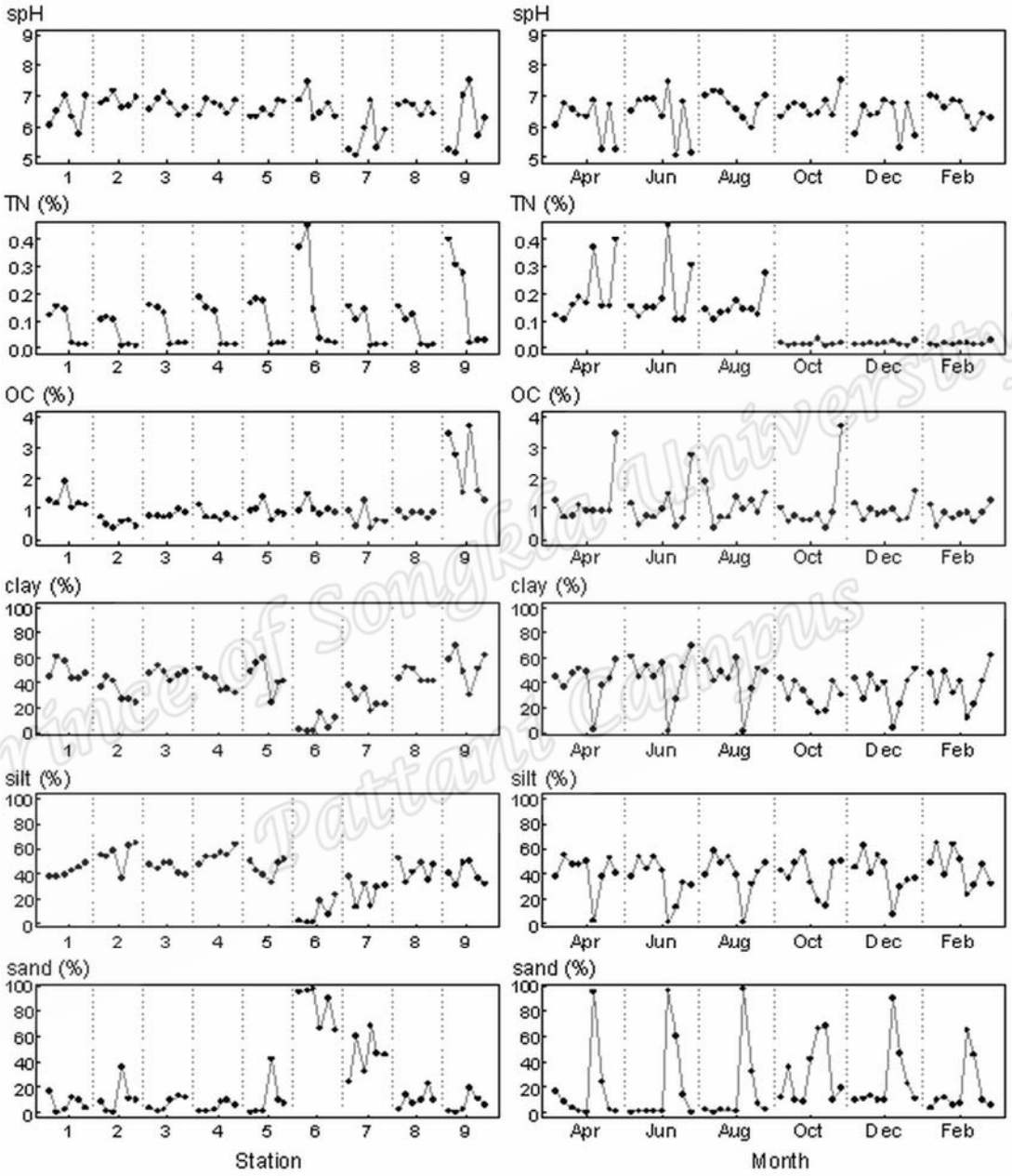

Figure 4: Sediment characteristics in Middle Songkhla Lake from April 1998 to February 1999 by station (left panel) and month (right panel)

Occurrence and abundance of macrobenthic fauna

Table 1 shows the taxa percentages of occurrence and density in individuals per square meter of the 24 families of macrobenthic fauna measured with the water characteristics. A total of 24 families were classified in three phyla of Annelida (Polychaeta), Arthropoda (Crustacea) and Mollusca (Gastropoda and Bivalvia), which comprised the most diverse group (35.2-98.2% of occurrence). The Polychaeta was

represented by nine families (Capitellidae, Goniadidae, Hesionidae, Nephtyidae, Nereididae, Pilargiidae, Pholoidae, Spionidae and Terebellidae). The Crustacea was also represented by nine families (Aoridae, Isaeidae, Melitidae, Oedicerotidae, Apseudidae, Pseudotanaidae, Anthuridae, Cirolanidae and Alpheidae). Marginellidae, Retusidae, Skeneopsidae and Stenothyridae were in the Gastropoda whilst the two remaining families (Tellinidae and unidentified species were in the Bivalvia). Nereididae was the most commonly observed family with 98.2% occurrence whereas Terebellidae and Stenothyridae had lowest occurrence (35.2%). Apseudidae was the most abundant family with average density of 40 083.6 ind m^{-2}; on the other hand, the Alpheidae was least abundant with average density 98.2 ind m^{-2}.

Table 1: Taxa occurrence (%) and density (individuals per square meter) of 24 families of macrobenthic fauna in Middle Songkhla Lake from April 1998 to February 1999

Phylum	Class	Order	Family	% occ	Density (ind m^{-2})
Annelida	Polychaeta	Capitellida	1: Capitellidae	87.0	1 227.3
	Polychaeta	Phyllodocida	2: Goniadidae	37.0	443.6
	Polychaeta	Phyllodocida	3: Hesionidae	55.6	698.2
	Polychaeta	Phyllodocida	4: Nephtyidae	87.0	2 218.2
	Polychaeta	Phyllodocida	5: Nereididae	98.2	8 507.3
	Polychaeta	Phyllodocida	6: Pilargiidae	70.4	1 625.5
	Polychaeta	Phyllodocida	7: Pholoidae	59.3	658.2
	Polychaeta	Spionida	8: Spionidae	92.6	5 056.4
	Polychaeta	Terebellida	9: Terebellidae	35.2	1 136.4
Arthropoda	Crustacea	Amphipoda	10: Aoridae	59.3	2 421.8
	Crustacea	Amphipoda	11: Isaeidae	87.0	6 900.0
	Crustacea	Amphipoda	12: Melitidae	94.4	4 438.2
	Crustacea	Amphipoda	13: Oedicerotidae	66.7	667.3
	Crustacea	Tanaidacea	14: Apseudidae	90.7	40 083.6
	Crustacea	Tanaidacea	15: Pseudotanaidae	37.0	4 265.5
	Crustacea	Isopoda	16: Anthuridae	75.9	3 816.4
	Crustacea	Isopoda	17: Cirolanidae	37.0	427.3
	Crustacea	Decapoda	18: Alpheidae	40.7	98.2
Mollusca	Gastropoda	Neogastropoda	19: Marginellidae	85.2	3 963.6
	Gastropoda	Cephalaspidea	20: Retusidae	55.6	5 536.4
	Gastropoda	Mesogastropoda	21: Skeneopsidae	38.9	956.4
	Gastropoda	Mesogastropoda	22: Stenothyridae	35.2	581.8
	Bivalvia	Unidentified	23: Unidentified	44.4	338.2
	Bivalvia	Veneroida	24: Tellinidae	81.5	17 134.5

Factor analysis

DO and TSS were omitted from the factor analysis due to high uniquenesses (0.975 and 0.848, respectively). The model provided an adequate fit using three factors (chi-squared = 20.23, 12 df, *p*-value = 0.063).
Table 2 shows the loadings, with values less than 0.20 in magnitude suppressed. If only loadings greater in magnitude than 0.45 are considered, the three factors do not contain any overlapping variables.

Factor 1 encompasses salinity, containing positive loadings for Sal and wpH, and a negative loading for wDep as expected, with deeper water during the rainy season. Factor 2 represents the effect of sediment characteristics in the lake bed (sand-clay habitat), consisting of a positive loading for sand and a similar negative loading for clay. Factor 3 characterizes physical and chemical compositions in the lake,

comprising positive loadings for TN, OC, and wTemp, and a negative loading for spH. Thus Factor 1 was defined as –0.53 × wDep + 0.70 × wpH + 0.99 × Sal, Factor 2 as 0.94 × Sand – 0.95 × Clay, and Factor 3 as 0.47 × OC + 0.58 × TN – 0.57 × spH + 0.54 × wTemp. The three factors respectively accounted for 24.6%, 20.7%, and 13.7% of the variance in the environmental data, a total of 59.0%. The three factors were included in the regression model as predictors together with the two singleton variables omitted from the factor analysis, with each of these five predictor variables scaled to have mean 0 and standard deviation 1. The correlation coefficients of the predictors ranged from –0.17 (between Factor 1 and Factor 2) to 0.40 (between Factor 1 and Factor 3).

Table 2: Factor analysis (with loadings below 0.2 omitted)

Variable	Factor	Factor	Factor
Organic carbon (OC)	-	-	**0.47**
Total nitrogen (TN)	0.34	-	**0.58**
Sediment pH (spH)	0.39	-	**-0.57**
Water depth (wDep)	**-0.53**	-	-
Water pH (wpH)	**0.70**	-	-
Salinity (Sal)	**0.99**	-	-
Water temperature (wTemp)	0.42	-	**0.54**
Sand	-	**0.94**	-
Clay	-	**-0.95**	-
% Total variance	24.6	20.7	13.7
% Cumulative variance	24.6	45.3	59.0

Regression analysis

The choice $c = 100$ in the transformation $\log(1 + c \times \text{density})$ gave residuals satisfying the normality assumption. The correlation coefficients of the residuals ranged from –0.34 (between Cirolanidae and Goniadidae) to 0.62 (between Anthuridae and Pseudotanaidae).

The left panel of Table 3 shows the corresponding individual regression coefficients and standard errors and r-squared values for each family after fitting the MMR model with all five environmental predictors included. The right panel shows the corresponding results for a reduced model containing only the two predictors that were statistically significant in the MANOVA, as shown in Table 4.

The coefficients listed are the ones statistically significant at 5% and 1% (in bold). Since there are 120 regression coefficients in all and 5% of these would be expected to have *p*-values less than 0.05 even if all their corresponding population parameters was zero, the six largest *p*-values less than 0.05 are italicized to indicate failure to achieve "honest" significance. Table 3 also shows additional coefficients (labeled *ns*) that were not statistically significant in their fitted model, but achieved significance in

the other model. Note that the coefficients for TSS are reversed in sign because most were negative, so this predictor is labeled –TSS.

Table 3: Coefficients and standard errors (in parenthesis) from fitting multivariate multiple regression models with all five environmental predictors (left panel) and with only the two statistically significant predictors in the MANOVA (right panel). Coefficients with *p*-values greater than 0.05 in both models are omitted; those adjudged not honestly statistically significant are shown in italics and those with *p*-values less than 0.01 are shown in bold.

Family	5 predictors						2 predictors		
	Factor 1	Factor 2	Factor 3	–TSS	DO	r^2	Factor 1	Factor 2	r^2
Hes	**0.76** (0.28)	-	-	-		0.23	0.67 (0.25)	-	0.18
Uni	0.71 (0.27)	-	-	-	-	0.16	0.53 (0.25)	-	0.08
Spi	0.44 (0.21)	-	-	-	-	0.13	0.43 (0.19)	-	0.09
Gon	**1.01** (0.26)	-	–0.76	1.08 (0.45)	-	0.31	0.61 (0.25)	-	0.11
Pho	–0.61 (0.25)	0.43^{ns}(0.22)	-	-	-	0.33	**–0.79** (0.23)	0.47 (0.22)	0.28
Pse	–0.64 (0.28)	**0.81** (0.24)	-	-	-	0.38	–0.65	**0.85** (0.24)	0.32
Ant	-	**0.69** (0.24)	-	-	-	0.19	-	**0.73** (0.24)	0.17
Pil	-	–0.68	-	-	-	0.17	-	**–0.66** (0.24)	0.16
Aor	-	**0.87** (0.26)	-	*–1.11*	-	0.26	-	**0.75** (0.27)	0.14
Ter	-	**1.04** (0.20)	*0.60* (*0.28*)	*0.92* (*0.41*)	-	0.45	-	**1.08** (0.21)	0.35
Cir	0.42^{ns} (0.24)	0.51 (0.21)	0.67 (0.29)	-	-	0.30	**0.69** (0.23)	0.42^{ns} (0.21)	0.19
Ner	0.27^{ns} (0.15)	*0.30* (*0.14*)	*0.42* (*0.19*)	-	-	0.25	**0.41** (0.14)	0.27^{ns} (0.14)	0.17
Ste	0.35^{ns} (0.25)	-	*0.65* (*0.30*)	-	-	0.25	0.52 (0.23)	-	0.17
Mel	0.20^{ns} (0.18)	-	0.58 (0.22)	-	-	0.24	0.42 (0.18)	-	0.11
Ret	-	-	-	**1.56** (0.58)	-	0.22	-	-	0.05
Cap	-	-	-	-	–0.86	0.15	-	-	0.07
Nep	-	-0.16^{ns} (0.24)	-	-	-	0.12	-	–0.42 (0.20)	0.08
Isa	-	0.46^{ns}(0.23)	-	-	-	0.16	-	0.47 (0.23)	0.09
Mar	-	0.42^{ns}(0.22)	-	-	-	0.16	-	*0.45* (*0.21*)	0.15
Alp	-	-	-	-	-	0.09	-	-	0.04
Oed	-	-	-	-	-	0.08	-	-	0.04
Aps	-	-	-	-	-	0.05	-	-	0.04
Ske	-	-	-	-	-	0.02	-	-	0.01
Tel	-	-	-	-	-	0.10	-	-	0.08

Canonical Correspondence Analysis

Figure 5 shows biplots based on the CCA matching the two MMR analyses. The families represented by dot whereas the environmental predictors represented by arrows. Each arrow determines an axis in the plots, obtained by extending the arrows in both directions. In addition, each dot (family) must drop a perpendicular to this axis (see for example in the right panel).

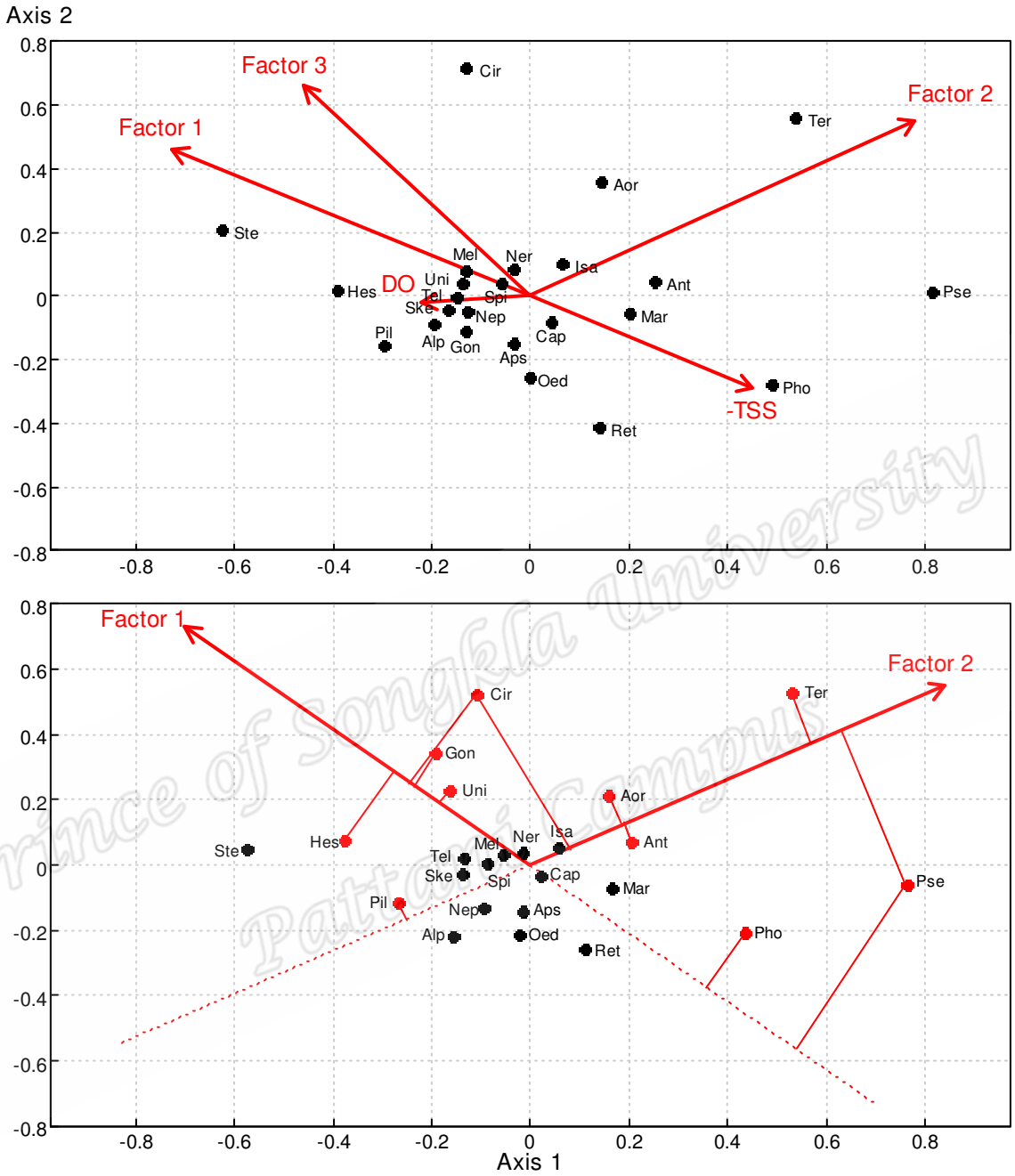

Figure 5: Upper panel: biplot of first two axes of CCA ordination diagram with 24 Macrobenthic fauna families against five environmental predictors. Lower panel: similar biplot omitting the three environmental variables not statistically significant in the MANOVA, with families showing highly significant coefficients in the MMR model (p-values < 0.01) connected to the corresponding arrows representing the predictors.

5. Discussion

Comparison of methods

This study aim was to identify potential environmental "key factors" causing distribution of macrobenthic fauna communities, to improve understanding concerning benthic/abiotic interactions and ecosystem functioning. By replacing groups of correlated predictors by single variables, factor analysis was used to remove correlations between environmental parameters that mask their effects on the macrobenthos densities.

Multivariate multiple regression and canonical correspondence analysis were used to examine the relations between the macrobenthic fauna family densities and the reduced set of environmental predictors. Although each of these five predictors was found to be associated with at least one family, the corresponding MANOVA decomposition found only two of them to be statistically significant overall. The biplot produced by the canonical correspondence analysis is seen to be more informative when only these two predictors were included.

The MMR model containing all five predictors gives seven associations between a family density and an environmental determinant that are highly statistically significant (p-value < 0.01), and a further nine with p-value between 0.01 and 0.05. Ten families showed no evidence of an association with any of the five determinants. Most of these associations can also be seen in the CCA biplot.

The most noticeable difference between the results of the two methods is that Spionidae is found to be associated with the salinity factor in the MMR model but this association is not seen in the biplots. Since Spionidae is a marine benthos (Day and Blake, 1979) with typical dominant species *Pseudopolydora kempi* and *Prionospio cirrifera* (Angsupanich et al., 2005a), there is evidence supporting the MMR result in this case. In the biplot containing all five environmental predictors the arrows for Factor 1 and TSS have identical directions, but the correlation between these variables (0.30) is not high.

Scientific findings

The results, both by MMR and CCA, clearly indicate that the salinity factor was positively associated with the densities of the Goniadidae, Hesionidae, and Spionidae and the unidentified families in the Bivalvia, and negatively associated with the densities of Pholoidae and Pseudotanaidae. This is in contrast with those analyzed based on the same data using BIOENV by Angsupanich (2005a) which indicating that no impact of salinity on benthos density. In general, salinity is an important factor affecting the distribution and structure composition of macrobenthic fauna in brackish water of costal habitats (Mannino and Montagna, 1997; Ogunwenmo and Osuala, 2004; Nanami et al., 2005). Although Middle Songkhla Lake is not connected to the sea directly, this zone receives the effect of salinity from the saltwater inflow through Lower Lake which is open to the Gulf of Thailand. Often salinity is regarded as a primary descriptor in estuarine ecosystems (Gaston, 1988; Lamptey and Armah, 2008).

A sedimentary habitat contains information mirroring the functional biodiversity and activity patterns of macrobenthic fauna (Rosenberg et al., 2009). The results show the

sedimentary factor was positively associated with the densities of Terebellidae, Aoridae, Pseudotanaidae, and Anthuridae, while a negative association was found with Pilargiidae. The main characteristics at the bottom of Middle Songkhla Lake are clay and silt (Angsupanich et al., 2005a) except for station six, which is mainly sand (84.6%). A typical genus *Sigambra* within Pilargiidae (Angsupanich et al., 2005a) was found to be negatively related with sand-clay excess, a finding supported by a study in the southeastern Gulf of California reporting that the genus *Sigambra* was dominant in the areas of sand percentage of 1% or mud of 60-70% (Méndez, 2007).

In addition, the genus *Marginella* within Marginellidae was also listed as being present in Middle Songkhla Lake (Angsupanich et al 2005a) thus showing a positive association with sand. This finding also agrees with a study of invertebrate species identified in Fresh Creek, Bahamas where *Maginella* was listed as most commonly having the habitat type of sandflat (Layman and Silliman, 2002).

Ten families (Nereididae, Stenothyridae, Nephtyidae, Isaeidae, Marginellidae, Alpheidae, Oedicerotidae, Apseudidae, Skeneopsidae, and Tellinidae) showed no evidence of association with any of the environmental variables. Although, Alpheidae was found to have the lowest density among the families included in our study, it is commonly found in the stomach contents of the dominant bottom feeding fish (*Osteogeneiosus militaris* and *Arius maculates*) in Middle Songkhla Lake. Angsupanich et al. (2005b) implied that these catfish species feed opportunistically on a variety of prey in their environment coupled with preferential feeding. So the low occurrence of Alpheidae may have been due to its swift movement and consequent catching difficulty.

Nereididae is one of the most important polychaete due to its diversity and abundance, found not only in marine environments (Gonzalez-Escalante and Salazar-Vallejo, 2003) but also in brackish water such as occurs in Middle Songkhla Lake. Fourteen species of Nereididae were reported in a former study (Angsupanich et al., 2005a) and it seems that Nereididae is widespread in Middle Songkhla Lake where it was the highest species richness. No evidence of Nereididae variation with salinity was found, possibly due to species diversity within this family. Some species, such as *Ceratonereis hircinicola*, were widely spread in the high salinity areas (Angsupanich and Kuwabara, 1995), whereas *Namalycastis indica* has been found to inhabit fresh to slightly brackish water in cisterns, pools and lagoons (Glasby, 1999).

Songkhla Lake nowadays suffers from the use of coastal land and water resources for uncontrolled shrimp farming, the destruction of both mangrove areas and peat swamp forest, construction of intake and outfall structures, and the construction of a deep sea port (Chufamanee et al., 2003). The analytic methods we have used are designed to gain a better understanding of the environmental factors associated with macrobenthic fauna and it is able to be used as additional or alternative method for an analysis of relationship between environmental variables and abundance of benthic organisms. This knowledge is useful for the natural resource management that needs to be conducted for effective management of similar estuarine environments.

Acknowledgements

We thank Emeritus Prof. Dr. Don McNeil and Asst. Prof Dr. Sukree Hajisamae for their advice on the statistical analysis. The paper has been substantially improved by

constructive comments from anonymous referees. The data were collected using funds provided by the Biodiversity Research and Training Program (BRT), under the National Science and Technology Development Agency (NSTDA), together with the Thailand Research Fund (TRF) with the grant number 142016.

References

Abdi, H., 2003. Factor rotations in factor analyses; in M. Lewis-Beck, A. Bryman, T. Futing (Editors), Encyclopedia of Social Sciences Research Methods. Thousand Oaks, CA, Sage, pp. 1-8.

Anderson, M.J., 2008. Animal-sediment relationships re-visited: Characterising species distributions along an environmental gradient using canonical analysis and quantile regression splines. Journal of Experimental Marine Biology and Ecology 366, 16-27.

Angsupanich, S., Kuwabara, R., 1995. Macrobenthic fauna in Thale Sap Songkla, a brackish lake in southern Thailand. Lakes and Reservoirs: Research and Management 1, 115-125.

Angsupanich, S., Siripech, A., Charoenpornthip, M., 2005a. Macrobenthic fauna community in the Middle Songkhla Lake, Southern Thailand. Songklanakarin Journal of Science and Technology 27, 365-390 (in Thai, with English Abstr.).

Angsupanich, S., Somsak, S., Phrommoon, J., 2005b. Stomach contents of the catfishes *Osteogeneiosus militaris* (Linnaeus, 1758) and *Arius maculatus* (Thunberg, 1792) in the Songkhla Lake. Songklanakarin Journal of Science and Technology 27, 391-402 (in Thai, with English Abstr.).

Borja, A., Franco, J., Pérez, V., 2000. A marine biotic index to establish the ecological quality of soft-bottom benthos within European estuarine and coastal environments. Marine Pollution Bulletin 40, 1100-1114.

Browne, M.W., 2001. An overview analytic rotation in exploratory factor analysis. Multivariate Behavioral Research 36, 111-150.

Chufamanee, P., Boromthanarat, S., Lønholdt, J., 2003. A case experience on integrated environment and water management towards people's livelihoods, Case story 1: Partnership policy in Songkhla Lake Basin. Workshop report. Songkhla workshop on linking management of catchments and coastal ecosystems. Royal Thai Ministry of Natural Resources and Environment. pp. 15-32.

Clarke, K.R., Somerfield, P.J., Chapman, M.G., 2006. On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted Bray-Curtis coefficient for denuded assemblages. Journal of Experimental Marine Biology and Ecology 330, 55-80.

Clarke, K.R., Warwick, R.M., 1994. Change in Marine Communities: an Approach to Statistical Analysis and Interpretation. Natural Environment Research Council, Plymouth, UK, 144 pp.

Day, R.L., Blake, J.A., 1979. Reproduction and larval development of Polydora Giardi Mesnil (Polycheata: Spionidae). The Biological Bulletin 156, 20-30.

Ellis, J., Ysebaert, T., Hume, T., Norkko, A., Bult, T., Herman, P., Thrush, S., Oldman, J., 2006. Predicting macrofaunal species distributions in estuarine gradients using logistic regression and classification systems. Marine Ecology Progress Series 316, 69-83.

Engle, V.D., Summers, J.K., Gaston, G.R., 1994. A benthic index of environmental condition of Gulf of Mexico estuaries. Estuaries 17, 372-384.

Frédou, T., Ferreira, B.P., Letourneur, Y., 2006. A univariate and multivariate study of reef fisheries off northeastern Brazil. Journal of Marine Science 63, 883-896.

Gaston, G.R., Lee, D.A.L., Nasci, J.C., 1988. Estuarine macrobenthos in Calcasieu Lake, Louisiana: Community and trophic structure. Estuaries and Coasts 11, 192-200.

Glasby, C.H., 1999. The Namanereidinae (Polychaeta: Nereididae). Part 1, Taxonomy and phylogeny. Records of the Australian Museum, 129 pp.

Glockzin, M., Zettler, M.L., 2008. Spatial macrozoobenthic distribution patterns in relation to major environmental factors- A case study from the Pomeranian Bay (southern Baltic Sea). Journal of Sea Research 59, 144-161.

Gonzalez-Escalante, L.E., Salazar-Vallejo, S.I., 2003. A new estuarine species, *Nereis garwoodi* (Polychaeta: Nereididae), from Bahia Chetumal, Macican Caribbean coast. Revista de Biología Tropical 51, 155-164.

Gray, J.S., Aschan, M., Carr, M.R., Clarke, K.R., Rosenberg, R., Warwick, R.M., 1988. Analysis of community attributes of the benthic macrofauna of Frierfjord/ Langesundfjord and in a mesocosm experiment. Marine Ecology Progress series 46, 151-165.

Guerra-García, J.M., González-Villa, F.J., García-Gómez, J.C., 2003. Aliphatic hydrocarbon pollution and macrobenthic assemblages in Ceuta harbour: a multivariate approach. Marine Ecology Progress Series 263, 127-138.

Hajisamae, S., Chou, L.M., 2003. Do shallow water habitats of an impacted coastal strait serve as nursery grounds for fish? Estuarine, Coastal and Shelf Science 56, 281-290.

Hawkins, C.P., Norris, R.H., Hogue, J.N., Feminella, J.W., 2000. Development and evaluation of predictive models for measuring the biological integrity of streams. Ecological Applications 10, 1456-1477.

Johnson, R.A., Wichern, D.W., 1998. Applied Multivariate Statistical Analysis, forth ed. Prentice-Hall, USA, 816 pp.

Joy, M.K., Death, R.G., 2000. Development and application of a predictive model of riverine fish community assemblages in the Taranaki region of the North Island, New Zealand. New Zealand Journal of Marine and Freshwater Research 34, 241-252.

Lamptey, E., Armah, A.L., 2008. Factors affecting macrobenthic fauna in a tropical hypersaline coastal lagoon in Ghana, West Africa. Estuaries and Coasts 31, 1006-1019.

Layman, C.A., Silliman, B.R., 2002. Preliminary survey and diet analysis of juvenile fishes of an estuarine creek on Andros Island, Bahamas. Bulletin of Marine Science 70, 199-210.

Liang, S., Shieh, B., Fu ,Y., 2002. A structural equation model for physiochemical variables of water, benthic invertebrates, and feeding activity of waterbirds in the Sitsao wetlands of Southern Taiwan. Zoological studies 41, 441-451.

Mannino, A., P.A., Montagna., P.A., 1997. Small-scale spatial variation of macrobenthic community structure. Estuaries 20, 159-173.

Méndez, N., 2007. Relationships between deep-water polychaete fauna and environmental factors in the southeastern Gulf of California, Mexico. Scientia Marina 71, 605-622.

Morrisey, D.J., Turner, S. J., Mills, G.N., Williamson, R.B., Wise, B.E., 2003. Factors affecting the distribution of benthic macrofauna in estuaries contaminated by urban runoff. Marine Environmental Research 55, 113-136.

Nanami, A., Saito, H., Akita, T., Motomatsu, K., Kuwahara, H., 2005. Spatial distribution and assemblage structure of macrobenthic invertebrates in a brackish lake in relation to environmental variables. Estuarine, Coastal and Shelf Science 63, 167-176.

Ogunwenmo, C.A., Osuala, I.A., 2004. Physico-chemical parameters and macrobenthos of an estuarine creek and an artificial pond in Lagos, South-Western Nigeria. Acta Satech 1, 128-132.

Olson, C.L. 1976. On choosing a test statistic in multivariate analysis of variance. Psychological Bulletin 83, 579-586.

Pearson, T.H., Rosenberg, R., 1978. Macrobenthic succession in relation to organic enrichment and pollution of the marine environment. Oceanography and Marine Biology Annual Review 16, 229-311.

Quintino, V., Elliott, M., Rodrigues, A.M., 2006. The derivation, performance and role of univariate and multivariate indicators of benthic change: Case studies at differing spatial scales. Journal of Experimental Marine Biology and Ecology 330, 368-382.

R Development Core Team, 2009. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rakocinski, C.F., Brown, S.S., Gaston, G.R., Heard, R.W., Walker, W.W., Summers, J.K., 1997. Macrobenthic responses to natural and contaminant-related gradients in Northern Gulf of Mexico estuaries. Ecological Applications 7, 1278-1298.

Ranasinghe, J.A., Thompson, B., Smith, R.W., Lowe, S., Schiff, K.C., 2004. Evaluation of benthic assessment methodology in Southern California Bays and San Francisco Bay. Technical Report No. 432, Southern California Coastal Water Research Project, pp. 1-70.

Rosenberg, R., Magnusson, M., Nilsson, H.C., 2009. Temporal and spatial changes in marine benthic habitats in relation to the EU Water Framework Directive: The use of sediment profile imagery. Marine Pollution Bulletin 58, 565-572.

Rygg, B., 1985. Distribution of species along pollution-induced diversity gradients in benthic communities in Norwegian fjords. Marine Pollution Bulletin 16, 469-474.

Ter Braak, C.J.F., 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67, 1167-1179.

Ter Braak, C.J.F., Šmilauer, P., 2002. CANOCO Reference manual and CanoDraw for Windows User's guide: Software for Canonical Community Ordination (version 4.5). Microcomputer Power (Ithaca, NY, USA), 500 pp.

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S, fourth ed. Springer Science + Business Media, New York, USA, 495 pp.

von Wehrden, H., Hanspach, J., Bruelheide, H., Wesche, K., 2009. Pluralism and diversity-trends in the use and application of ordination methods 1990-2007. Journal of Vegetation Science 20, 695-705.

Warton, D.I., Hudson, H.M., 2004. A MANOVA statistic is just as powerful as distance-based statistics, for multivariate abundances. Ecology 85, 858-874.

Warwick, R.M., 1986. A new method for detecting pollution effects on marine macrobenthic communities. Marine Biology 92, 557-562.

Weisberg, S.B., Dauer, D.M., Schaffner, L.C., Frithsen, J.B., 1997. An estuarine benthic index of biotic integrity (B-BI) for Chesapeake Bay. Estuaries 20, 149-158.

Ysebaert, T., Herman, P.M.J., Meire, P., Craeymeersch, J., Verbeek, H., Heip, C.H.R., 2003.A large-scale spatial pattern in estuaries: estuarine macrobenthic communities in the Schelde estuary, NW Europe Estuarine, Coastal and Shelf Science 57, 335-355.