Chapter 2

Methodology

This chapter describes the methods used in the study. The following topics are covered.

(a) the data management;

(b) the methods used for the quantitative analysis;

(c) the methods used for creating the geographical maps.

2.1 The data management

The following computer programs were used for data analysis and thesis preparation.

Microsoft Office

Excel was mainly used to manage the data for this research. It has many functions that are useful for manipulating data, including functions for calculating Poisson and Gamma distributions. *Word* was used to create some of the figures and to format the thesis.

WebStat, a suite of web-database software engineering tools written in HTML and ASP (Microsoft's Active Server Pages) is used to create graphs for statistical analysis, including histograms and scatterplot matrices, when the data are stored in an SQL Server database. It runs on a local area intranet, and is available in the postgraduate computer laboratory in the Department of Mathematics and Computer Science at Prince of Songkla University in Pattani.

EcStat, a statistical add-in to Excel, is used to create scores plots, used for assessing the statistical assumptions.

The open source statistical package R is used to fit the negative binomial distributions to the disease incidences.

2.2 Methods used for the Quantitative Analysis

Graphical methods

The following graphical displays were created using WebStat.

1. Histograms with statistical summaries of raw and transformed data for all variables representing the distribution and summaries including the sample size, mean, standard deviation, minimum and maximum of a set of data.

2. Scatterplot matrix showing relationships between disease incidence rates before and after transformation using logarithms.

Statistical Methods

The statistical methods used for the data analysis may be described as follows.

Correlation Coefficient

The correlation coefficient is a measure of the strength of the linear or straight-line relationship between the two variables. The correlation coefficient is defined as (McNeil et al, 1998: 181)

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \tag{2.1}$$

It may be shown that r ranges from a minimum of -1 to a maximum value of 1. The correlation equal to 0 indicates no linear association between the two variables.

Poisson distribution

The Poisson distribution is for a random variable X is defined as

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \ldots, \tag{2.2}$$

where λ is density or mean occurrence per unit of time or space. In a meteorite example, this could be expressed in impacts per year. For disease occurrence, it could represent

the number of cases to occur in a subdistrict of a province over a specified period of time such as a year, in which case λ is the size of the population at risk in the subdistrict.

The Poisson distribution arises from observations that are statistically independent. It can be shown that X has mean λ and variance λ.

Negative binomial distribution

The negative binomial distribution for a random variable X is defined as

$$P(X = k) = \binom{r-1+j}{r-1} p^r (1 \quad p)^{r-1+j}, \quad k = 0, 1, 2, \ldots, \tag{2.3}$$

where $0 < p < 1$ and $r > 0$. For this distribution it can be shown that X has mean $r(1-p)/p$ and variance $r(1-p)/p^2$.

Defining $\lambda = r(1-p)/p$ and $\alpha = (1-p/r)/(1-p)$, so that $p = 1 - \lambda/(1+\lambda\alpha)$ and $r = 1-\lambda+\lambda\alpha$, it follows that X has mean λ and variance $\lambda(1+\lambda\alpha)$, and it can be shown that the Poisson distribution is the special case of the negative binomial distribution when $\alpha = 0$.

It can also be shown that as λ tends to infinity the distribution of the standardised random variable $Z = (X - \lambda)/\sqrt{(\lambda+\lambda^2\alpha)}$ has a left-shifted Gamma distribution with mean 0, scale parameter $1/\sqrt{\alpha}$, and shape parameter $1/\alpha$, that is, a distribution with probability density function

$$f(z) = \frac{\alpha^{0.5/\alpha}}{\Gamma(1/\alpha)} \left(z + 1/\sqrt{\alpha}\right)^{1/\alpha - 1} e^{-z\sqrt{\alpha}}, \quad z > -1/\sqrt{\alpha}. \tag{2.4}$$

P-value

The p-value is of the probability of observing an event at least as extreme as the specific event observed, based on a specified statistical assumption about the distribution of events. In our study, a p-value may be associated with the incidence rate for each

disease in each subdistrict. If a p-value is sufficiently small, it indicates that an unlikely event has occurred. P-values are useful for the control of diseases, because a sufficiently small p-value should trigger an alarm, resulting in intervention such as emergency relief.

If the statistical assumption is made that all persons at risk of a specified disease in a specified period of time are equally likely to become infected, independently of each other, then it follows that the number infected in a subdistrict i has a Poisson distribution with parameter λ_i. The p-value is then given by the formula

$$p_i = \sum_{j=n_i}^{\infty} \frac{\lambda_i^j}{j!} \exp(-\lambda_i), \tag{2.5}$$

where n_i is the number of infected cases observed in subdistrict i.

If the independence assumption is not satisfied, the negative binomial distribution is preferable to the Poisson distribution for describing the number of observed events in a region over a period of time, and the p-value corresponding to n_i observed events is given by the formula

$$p_i = \sum_{j=n_i}^{\infty} \binom{r_i - 1 + j}{r_i - 1} p_i^{r_i} (1 - p_i)^{r_i - 1 + j}. \tag{2.6}$$

If n_i is sufficiently large, the asymptotic distribution (2.4) may be used as an approximation.

2.3 The methods used for creating the geographical maps

Software used: MapInfo 7.0

Range maps

A thematic map is a type of map that uses a variety of graphic styles (usually colours of fill patterns) to graphically display information about a map's underlying data. Thus a thematic map using data in regions might show one region in dark red (to indicate that

the region has high values), while showing another region in very pale red (to indicate that the region has low values).

A range map is a type of thematic map that displays data according to ranges set by the users. The ranges are shaded using colour or patterns.

Contour map

A contour map is a way of graphing a variable that takes values in a two-dimensional region. It's not the only way of graphing such data: you could also use a 3D wire-frame map or a prism map.