CHAPTER 2

METHODOLOGY

This chapter includes a description of the methods used in the study. These methods include the following components.

(a) The selection of the data for the study;

(b) The procedure for structuring the data in a relational database;

(c) The graphical methods for displaying the data; and

(d) The methods used for the statistical analysis.

1. Selection of data

One of the objectives of the study is to assess the accuracy of the tide table at Pattani by compary the data given in the tide tables with data obtained by direct observation. The direct observation data were collected in 1996 at Pattani Bay over two periods, one week apart, each of 25 hours duration. To make an effective comparison, it is thus necessary to select the tide table data from Pattani in 1996. The direct observations were measured every five minutes at six locations in Pattani Bay. One of these locations is close to the position at which the tide table data are listed. See Figure 5.

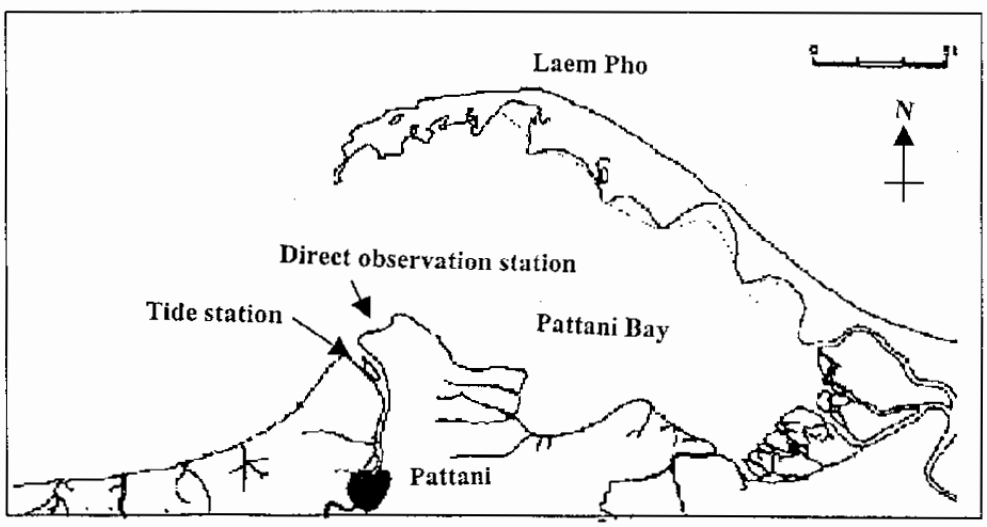

Figure 5: Map showing the approximate locations of the tide stations and direct observation station at Pattani.

Another objective of the study is to understand the pattern of variation in the tides at the selected location. To achieve this objective, it is desirable to select data for a reasonably long period of time. To allow for seasonal variation, this period should be no less than one year.

The tide tables give water levels to the nearest millimetre, relative to the lowest low water recorded over a long period, at each occurrence of a high or low tide, together with the time of occurrence to the nearest minute.

The study also aimed to gain some understanding of the spatial variation in the tides. To achieve this objective, the data obtained from the tide tables for Songkhla (*at Ko Nu)* during the same period are also investigated. The location of Ko Nu is shown in Figure 6.

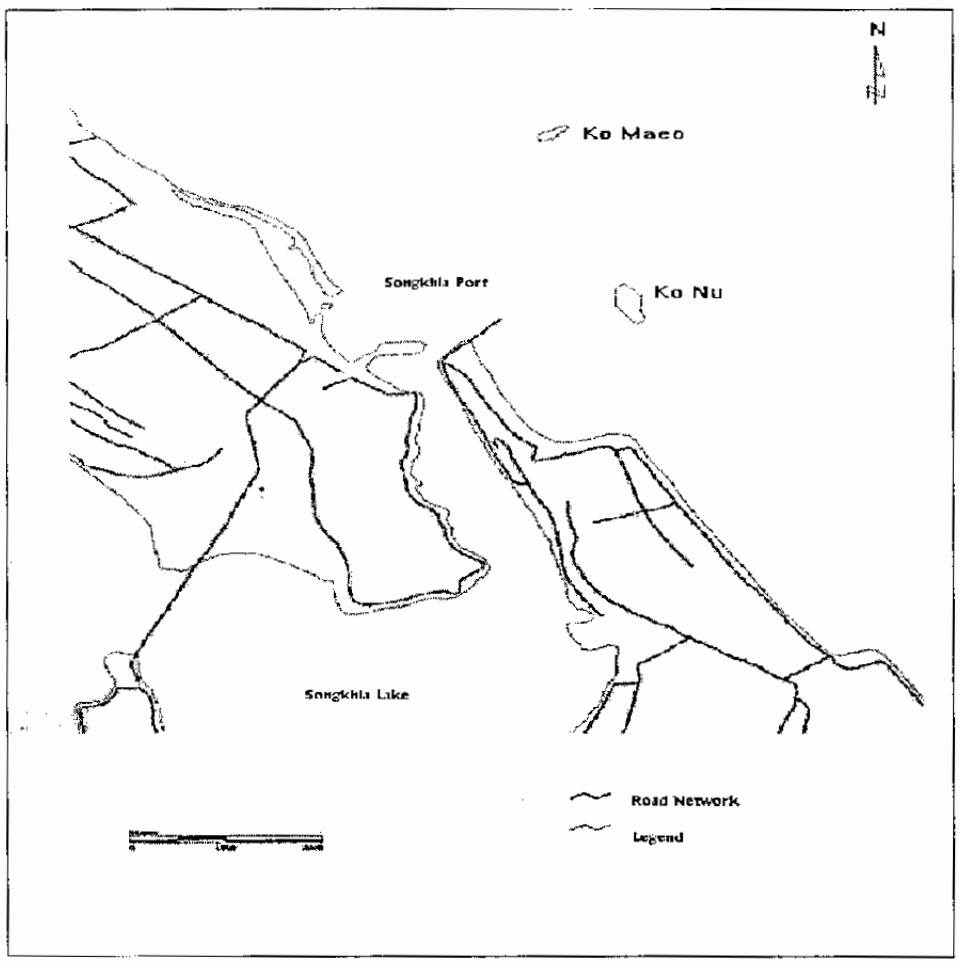

Figure 6: Map showing the location of the tide station of Songkhla.

2. Structure of data records

The direct observation data are stored simply as a time series giving the relative height of the water to the nearest centimetre at successive intervals. To reduce the effects of high frequency oscillations that are not relevant to the present study, taking averages of three measurements at 15-minute intervals smoothed these data. The data thus comprise three fields: (i) water level (metres); (ii) time in hours after midnight; (iii) study period (May 25-26 or June 1-2).

The tide data have a different structure from the direct measurement data. Instead of having data recorded at regular intervals, they are given at the times of occurrence of high and low tides in a day.

The tide tables data are kept in a Microsoft Access database file called *tide.mdb* in the directory f:\home\tides on the *garudo* file server, Department of Applied Mathematics, Faculty of Science and Technology, Prince of Songkla University, Pattani and are thus available to participating scientists authorised to access this directory. The structure for the tide data is given in Figure 7.

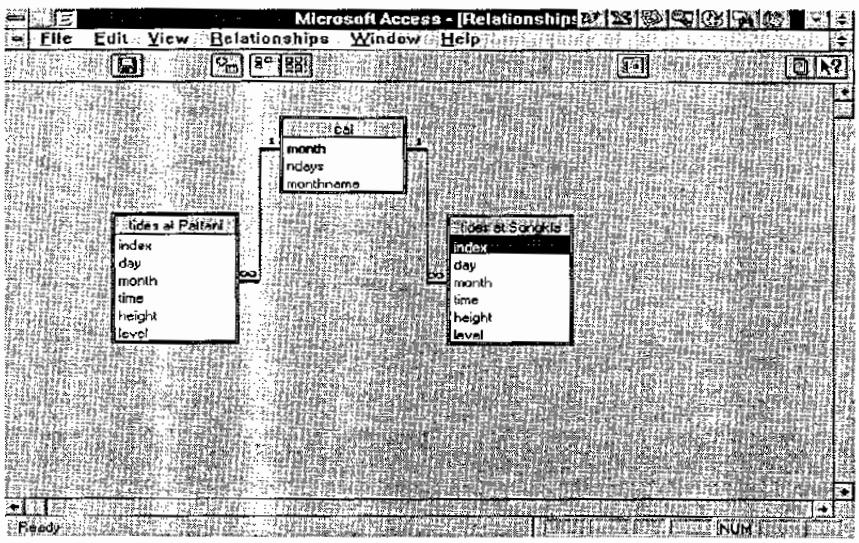

Figure 7: Relationships between tables in ***tides.mdb*** *database*

There are three tables in the database: *tide at Pattani*, *cal* and *tide at Songkhla*. In the table's *tide at Pattani* and *tide at Songkhla* records have the variables *index*, *day*, *month*, *time*, *height* and *level*. *Index* is an integer specifying the time at which a measurement is recorded in one year. *day* is an integer specifying the day at which a

measurement is recorded (1-30 or 1-31). *month* is an integer specifying the month at which a measurement is recorded (1-12). *time* is the time in 24 hour system. *height* is the height of water in centimetres at high tide and low tide. *level* is a dichotomous variable indicating water level (*high* or *low*). In the table *cal* the variables are *month*, *ndays* and *monthname*. *month* is an integer specifying the month (1-12). *ndays* is the numbers of days since January 1, *monthname* is the name of each month.

A query was used to create the data structure appropriate for statistical analysis (see Appendix A).

3. Graphical Methods

The direct observation data may be effectively graphed simply by plotting the water level on the vertical axis against the time on the horizontal axis with successive points joined by straight lines, using the Matlab software package. These data may thus be compared with the corresponding tide table data by superimposing the two data sets on the same axes.

The tide table data may be graphed using the same method. However, with over 700 occurrences of high and low tides during the year, this graph is not the most effective method for displaying the detailed variation in the data. For this reason, we also give graphs showing the data for a two months period.

Alternatively, the times of occurrence of successive high and low tides may be graphed on the vertical axis, with the index of the tide on the horizontal axis. Since the tide has a period of approximately 24 hours and 50 minutes (or more precisely, 24.812 hours) corresponding to the minimum period taken by a fixed point on the Earth's surface to arrive at the same position relative to the moon, it is more appropriate to graph the time of occurrence of the tide relative to this 'lunar day'. Using this method, the graph of the times of occurrence, for a semidiurnal tide pattern, will comprise four separate, approximately horizontal, curves corresponding to the first high tide (H1), the first low tide (L1), the second high tide (H2), and the second low tide (L2).

4. Statistical Methods

Data from the tide table fits very well for time series analysis. It is a set of numerical data measured sequentially in time with trend, cycle, seasonal variation and random disturbance or irregular movement. So the statistical methods used for the data analysis are based on time series analysis (see, for example, Chatfield 1989). These methods are described in this section.

4.1 Time Series

A time series is a set of numerical data measured sequentially in time. The measurements are assumed equidistant in the time (or nearly so). Time series data arise in economics (e.g. stock market prices), the physical sciences (e.g. barometric pressure), in biology (e.g. size of animal populations) and in many other applications. The statistical analysis of time series data is the subject of many recent texts (Diggle (1990), Tong (1990), Chatfield (1989), Abraham and Ledolter (1983), Whittle (1983), Anderson (1976), Bloomfield (1976), Box and Jenkins (1976), Fuller (1976), Anderson, (1971), Hannan (1970), Jenkins and Watts (1968), Brown (1963), Hannan (1960)), most of which involve complex mathematics. As in many areas of statistics, modern developments in computer technology have made time series methods accessible to persons with a minimal background in mathematics.

There are four important objectives in time series analysis. These are:

1. Forecasting future values of a series, eg. predicting the number of newly diagnosed cases of AIDS in a community;
2. Estimating the trend or overall characteristics of a time series, eg. seasonal components of monthly interest rates;
3. Modelling the dynamic relationship between two or more time series, eg. inflation and unemployment rates;
4. Summarising the characteristic features of a time series, eg. the cyclic behaviour of ozone levels in the upper atmosphere;

Conventional time series models are linear and multiplicative and give rise to uncorrelated statistical estimates. A crucial assumption underlying many of the methods used in time series is stationarity. For a time series to be stationary, the

statistical properties of the series should not change with time, i.e. the mean of the series should be approximately constant and the variability should be homogeneous.

If a time series is not stationary, a data transformation may be required. Inspecting a graph of the data usually assesses the need for a data transformation. A logarithmic transformation is usually appropriate for financial data, whereas a square root transformation is often better when dealing with counts data.

4.1.1 Removing a Trend

A time series may be decomposed into a deterministic component called a *signal* and a stochastic component called the *noise*. The signal may comprise a *trend* (usually just a linear function, or possibly a quadratic) and a periodic component. The trend should be removed before undertaking further statistical analysis. Once the trend has been removed, the series should be stationary.

4.1.2 Spectrum Analysis

A time series is stationary if its statistical properties do not change with time. It is unlikely that a stationary time series will repeat itself exactly, but the series is repeatable in a probabilistic sense. Another way of looking at this is to say that the character of the series persists as you move forward or backward in time, and the only aspect that changes is the sampling error, which does not contain useful information. Of course these sampling fluctuations could be relatively large compared to the persistent characteristic.

These ideas lead to the sinusoid (the simplest function that repeats itself) and to the idea of measuring the amount of periodicity or repeatability in a time series by finding its covariance or correlation with a sine wave having a given period. A sinusoid is characterized by the property that taking a linear transformation of its argument only shifts its frequency and its phase or position relative to some origin. The cosine function is just a sine function whose argument is shifted by $\pi/2$, that is

$$\cos(x) = \sin(x+ \pi/2) \qquad (1)$$

Since sinusoidal functions are periodic it is natural to use them as a basis for approximating a stationary time series. This basis comprises sine waves with different frequencies each defined on the time interval spanned by the data. The first

component appears exactly once on this time interval, the second comprises two repeated sinusoids, the third three sinusoids, and so on. These components are also called *harmonics*. The functional form for the j^{th} harmonic is a cosine wave with some phase ϕ, that is, $\cos\{2\pi j(t-1)/n+\phi\}$, $t = 1, 2, ..., n$.

Using the theory of Fourier analysis, any function defined at n equispaced points on a finite interval may be represented exactly by a constant and n−1 harmonics. The number of different frequencies in these components, m, is (n-1)/2 or n/2 (depending on whether n is odd or even) since there is a sine and a cosine harmonic at each frequency. If n is even this Fourier representation takes the form

$$y_t = a_0 + \sum[a_j\cos\{2\pi j(t-1)/n\}+b_j\sin\{2\pi j(t-1)/n\}]+a_m\cos\{\pi(t-1)\} \quad (2)$$

where the summation is from j=1 to j=m-1. (Since $\sin\{\pi(t-1)\}$ is 0 for all integers t, in this case there is no sine harmonic at the highest frequency). A similar formula applies if n is odd. Using the fact that a linear combination of a sine function and a cosine function at the same frequency may be expressed as a single sinusoid with some phase ϕ, an alternative formula for the Fourier representation is

$$y_t = a_0 + \sum A_j\cos\{2\pi j(t-1)/n\}+\phi_j\} \quad (3)$$

where the amplitude $A_j = \sqrt{(a_j^2 + b_j^2)}$ and the summation is from 1 to m.

This Fourier representation is similar to linear regression analysis, where the sinusoidal components play the role of determinants or predictor variables. Since the number of parameters is exactly equal to the number of data values, there is no residual error, the regression model provides a perfect fit to the data. Moreover it may be shown that the sum of products or sine and/or cosine harmonics over the range of frequencies is zero, which means that these harmonics are statistically uncorrelated with each other. Consequently each Fourier coefficient (a_j or b_j) is the regression coefficient of the time series y_t on the corresponding harmonic. The formulas for these coefficients (for n even) are as follows.

$$a_0 = \sum y_t /n$$

$$a_m = \sum (-1)^{t-1} y_t/n$$

$$a_j = (2/n)\sum y_t\cos\{2\pi j(t-1)/n\}$$

$$b_j = (2/n)\sum y_t\sin\{2\pi j(t-1)/n\}$$

We can see from these formulas that each Fourier coefficient may be interpreted as a covariance between the data and a sinusoid at the given frequency.

The *periodogram* of a time series (I_j, j = 1, 2, .., m) is defined in terms of the amplitudes of the harmonics in the Fourier representation as

$$I_j = (n/2)(a_j^2 + b_j^2) \qquad (4)$$

The multiplier n/2 ensures that the j^{th} periodogram value is equal to the component of the variance in the data accounted for by sinusoidal function with frequency j/n. Since the sinusoidal terms are uncorrelated with each other, it follows that

$$\Sigma(y_t - \Sigma y_t/n)^2 = \Sigma I_j \qquad (5)$$

This useful formula is known as *Parseval's theorem.*

This relation is just an analysis of variance for a time series. So the sum of the periodogram ordinates is equal to the total squared error of the data, and consequently the periodogram shows how much of the squared error of the data is accounted for by each various harmonics. For this reason it useful to graph the scaled periodogram, obtained by dividing the periodogram by its sum. The scaled periodogram thus shows what proportion of the squared error is associated with each harmonic.

Note that the frequency j/n is expressed in terms of the number of cycles per unit time. Since the values of j are 1, 2, ..., m, the lowest frequency is 1/n, corresponding to a period equal to the whole range of the data. and the highest frequency is close to 0.5 (exactly 0.5 if n is even), corresponding to cycles of length 2 with the data oscillating from one value to the next. A function *tsplot* may be used to show a periodogram of a time series.

4.1.3 Autoregressive models

We saw how the periodogram and its logarithm may be used to investigate the character of a time series. Another useful graphical tool is the *correlogram,* or sample *autocorrelation function*, which comprises the set of estimated correlation coefficients between the series and itself at various spacing. Thus the (auto)correlation coefficient at spacing (or lag) s may be estimated from the formula

$$r_s = \frac{\sum_{t=1}^{n-s}(y_t - \bar{y})(y_{t+s} - \bar{y})}{\sum_{t=1}^{n}(y_t - \bar{y})^2} \qquad (6)$$

and the correlogram is a graph of the series (r_s , s=1, 2, .., s) against the spacing s. Since the number of terms used to calculate the correlation coefficient at lag s is n-s where n is the length of the time series, the maximum spacing s should be substantially less than n.

According to statistical theory, when the sample size n is large the standard error of a correlation coefficient is approximately normally distributed with standard deviation $1/\sqrt{n}$, which tends to 0 as n gets large. This means that as the length of an observed time series increases, the sample autocorrelation function of a stationary time series stabilises, approaching a smooth curve.

For a process of independent measurements with a common normal distribution (a *white noise* process), the theoretical correlation between observations at different spacing is zero, so you would expect the graph of its sample autocorrelation function to approach the horizontal axis r = 0 as n gets large. Based on the normal distribution which has 95% of its probability within 1.96 standard deviations of its mean, a 95% confidence interval for the autocorrelation at lag s ranges from $-1.96/\sqrt{(n-s)}$ to $1.96/\sqrt{(n-s)}$. In contrast, the (unscaled) periodogram values of a white noise process, are exponentially distributed with constant standard deviation, and thus do not settle down as the length of the series increases. Instead they become more densely packed. Ljung & Box (1978) suggested using the statistic

$$Q = n(n+2)\sum_{s=1}^{m}\frac{r_s^2}{n-s} \qquad (7)$$

where m is a specified integer substantially less than the series length n, to the hypothesis that a time series is a sample from a white noise process. If it is necessary to fit a linear model involving p parameters to transform the series to a white noise process, where these parameters are estimated from the data, then Q is distributed approximately as a chi-squared distribution with m-p degrees of freedom.

The Asp function *tsplot* (See McNeil, et al., 1997) (ASP Users Manual) may be used to show a periodogram, base 10 logarithm of the periodogram and

autocorrelation function of a time series. It also has the capability of removing a linear or quadratic trend, fitting specified harmonic terms, and estimating autoregressive coefficients at specified lags.

4.2 Harmonic Analysis of the Tides

At lectures to the Lowell Institute in Boston in 1897, George Darwin, Plumiam Professor of Physics at the University of Cambridge, proposed a model for the movement of the tides based on harmonic analysis (Darwin, 1898, reprinted in 1962: 193-210). This model is based on spectrum analysis of a long series of measurements at (hourly) intervals at a specified location, with up to 20 or 25 harmonic components. As a result, it is possible to forecast the water levels accurately at any given location.