CHAPTER 2

METHODOLOGY

This chapter describes (a) the methods for data management, (b) graphical presentation, (c) statistical modelling, and (d) reconstruction of the tide heights using the fitted model.

1. Data management

1.1 Database creation

Tide tables in the Gulf of Thailand and the Andaman sea are produced by the Hydrographic Department of the Royal Thai Navy (Figure 2.1 shows a typical page from a tide table). Hourly heights are recorded each day of the year at 19 locations. These data are stored in a database using Microsoft Access. Windows technology with multitasking facilitates viewing the data whereas analysing them was performed by another program.

BANG NARA

January 1994

	HOURS																							
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
DATE	HEIGHT OF WATER IN DECIMETERS																							
1	9	7	6	5	4	4	5	5	6	6	6	5	4	3	2	2	3	3	5	6	8	9	10	10
2	9	8	7	6	5	4	4	5	5	6	6	6	5	4	3	3	3	3	4	5	7	8	9	10
3	9	8	7	6	5	5	4	5	5	6	6	7	6	5	4	4	3	4	4	5	6	7	8	10
4	9	8	7	7	6	5	5	5	5	5	6	6	7	6	6	5	5	4	4	5	5	6	7	7
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
29	7	6	5	4	4	4	5	6	6	6	6	5	4	3	3	3	4	5	6	8	9	10	10	10
30	8	6	5	5	4	4	5	6	6	7	7	6	5	4	3	3	3	4	5	7	8	9	10	10
31	8	7	6	5	4	4	4	5	6	7	7	7	6	5	4	4	4	4	5	6	7	8	9	9

Figure 2.1 Tide table for Bang Nara in January 1994

The data from the tide tables is incorporated into a relational database as follows. The database table has six columns. Column 1 is an identification number, columns 2 and 3 are month and day of the year, respectively, column 4 is tide type (0=low tide, 1=high tide), column 5 is the height of water predicted in decimeters above the lowest low water, and column 6 is hour of the day. Figure 2.2 shows the database table corresponding to the tide table shown in Figure 2.1.

ID	Month	Day	L/H	HT	Time
1	1	1	0	4	4.5
2	1	1	1	6	9.5
3	1	1	0	2	14.5
4	1	1	1	10	22.5
5	1	2	0	4	5.5
6	1	2	1	6	10
7	1	2	0	2.5	15.5
8	1	2	1	9.5	23
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
1241	12	30	0	4	11.5
1242	12	30	1	10.5	18.5
1243	12	31	0	5	3.5
1244	12	31	1	6	6.5
1245	12	31	0	3.5	12.5
1246	12	31	1	11	19.5

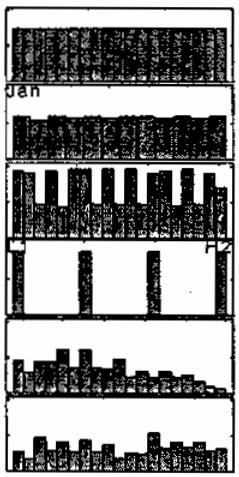
Figure 2.2 Database structure for Bang Nara

1.2 Data exploration and cleaning

Cleaning a data set is claimed to be the first step by the data collector and the keyboard operator. The data analyst should not assume that data are ready for statistical analysis. There are always human errors. The best thing to do is to minimize errors by preventing them. Diagnosis of the errors and proper treatment is always essential. This can be done by using the functions *describe* and *relate* of *Asp* (McNeil et al., 1997) run under Matlab (MathWorks, 1994), which considers variables one or two at a time.

1.3 Data structure

Statistical data for analysis must be stored as a rectangular array of numeric data with variables in columns. From Microsoft Access it is straight forward to export data to an ASCII data file that is used to produce the analysis. Next, the characteristics of the tides should be decomposed into the first high tide (H1), the first low tide (L1), the second high tide (H2) and the second low tide (L2). In a uniform diurnal or semidiurnal tidal system, the greatest height to which the tide rises on any day is known as high water, and the lowest point to which it drops is called low water. In a mixed-tide system, it is necessary to refer to higher high water and lower low water, as well as higher low water and lower low water. Tidal measurements taken from the tidal data are used to describe the data into higher high water (HH), lower high water (LH), higher low water (HL) and lower low water (LL) for semidiurnal or mixed tides. So it is necessary to change column 4 in the database to separate the type of tide as follows: 0 is the first high tide (H1), 1 is the first low tide (L1), 2 is the second high tide (H2) and 3 is the second low tide (L2). With this restructuring, the data summaries for Bang Nara, for all tides in 1994, are shown in Figure 2.3.



col	variable	size	mean	st dev	min	max
1	ID	1410	705.5	407.2	1	1410
2	month	1410	6.524	3.449	1	12
3	day	1410	15.76	8.796	1	31
4	tide type	1410	1.499	1.118	0	3
5	height	1244	5.744	2.596	1.5	12
6	hour	1244	12.04	6.803	0	23.5

Bang Nara

Figure 2.3 Summaries of restructured data

2. Graphical presentation

There are two reasons for graphing the tides separately as high-1, low-1, high-2 and low-2. The first is based on the fact that the earth rotates in 24 hours, not in 12 hours. The second is for empirical reasons. There is a marked difference in the characteristic patterns of the four tides when plotted. The data as structured in Figure 2.3 may be regarded as a bivariate set of outcomes for each tide type, indexed by a positive integer identifying each successive tide. This index is synonymous with lunar day, a period of 24 hours and 50.48 minutes. The two outcomes comprise the height h_i and the time of occurrence t_i for the tide on lunar day. These outcomes may be graphed as separate time series. For example, Figure 2.4 shows a graph of H_i for the first 58 tides of type L1 at Bang Nara in 1994.

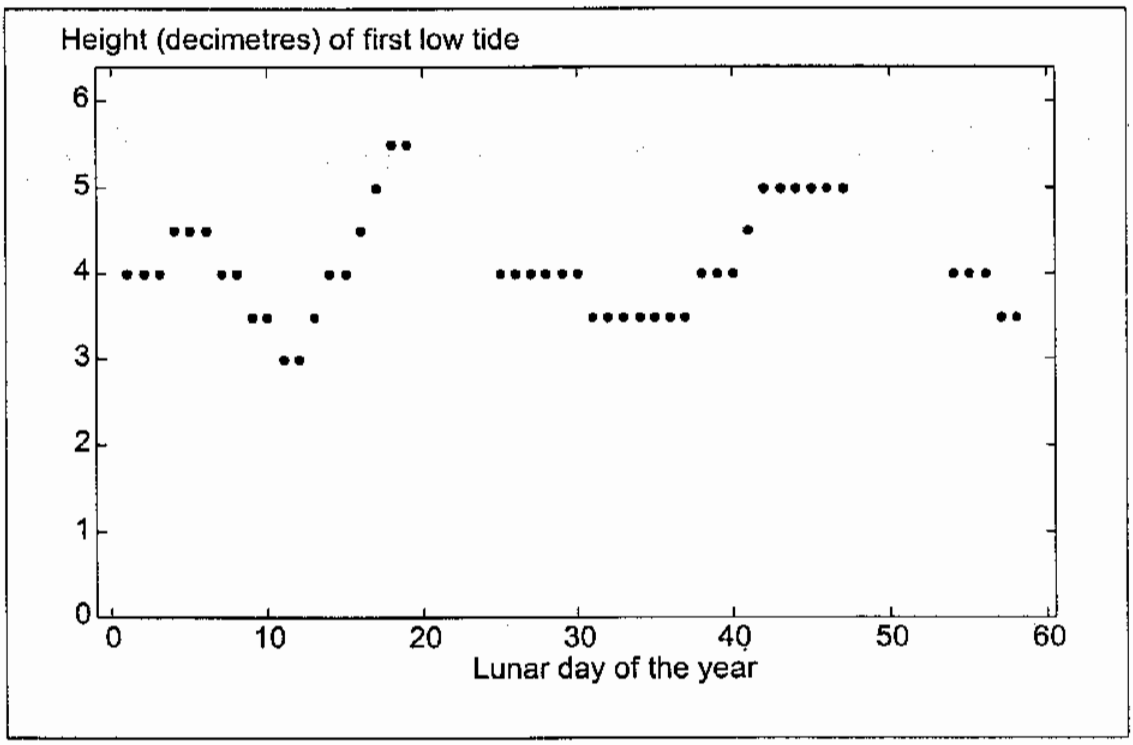


Figure 2.4 Height in decimeters of tide type L1 at Bang Nara in 1994

Note that there are no points plotted for indexes 20-24 and 48-53. The reason for these gaps is that the tides at Bang Nara are a mixture of diurnal and semidiurnal, and at certain periods during the lunar month there are only two tides each lunar day rather than four. Thus on lunar days 20, 21, 22, 23 and 24 in the first lunar month the

first low tide does not occur. Similarly, there is no first low tide on the lunar days 48, 49, 50, 51, 52 and 53.

The time of occurrence of the first low tide in each lunar day may be graphed in the same way. Figure 2.5 shows this graph for the first 58 tides of type L1 in 1994.

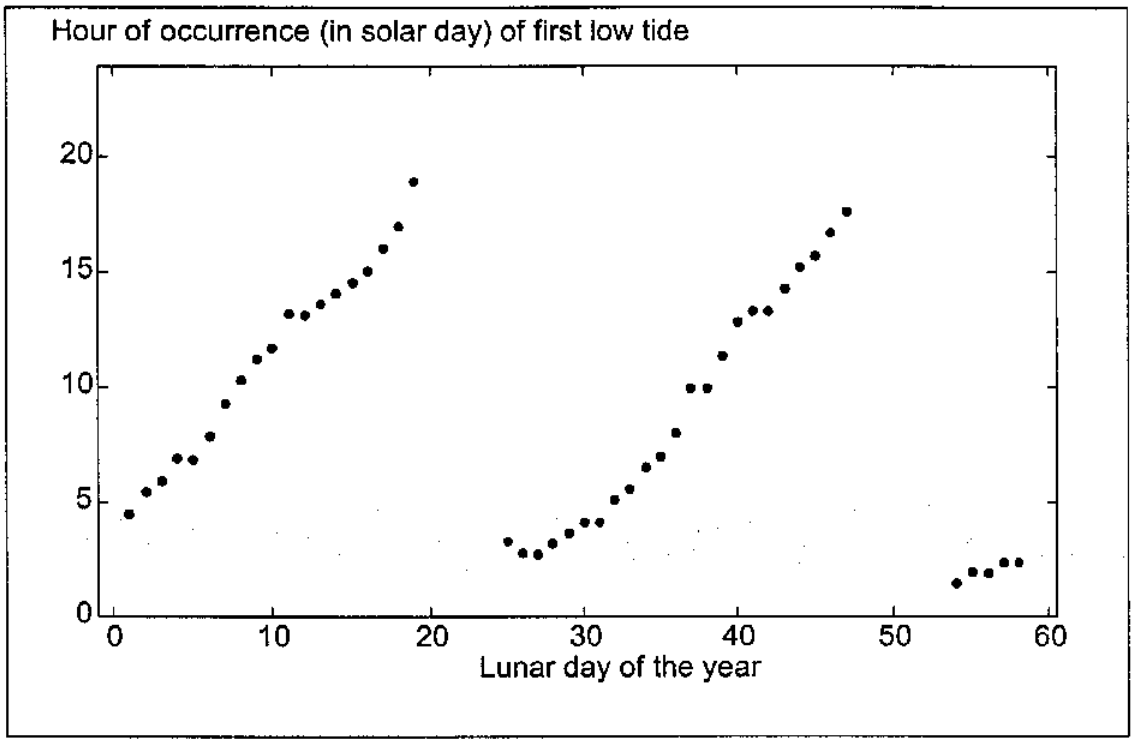


Figure 2.5 Times of occurrence of tide type L1 at Bang Nara in 1994

Note that there is a positive trend in Figure 2.5. Due to the fact that the interval between two occurrences of a tide of given type is on average, 24 hours and 50.48 minutes rather than 24 hours. Figure 2.6 shows the graph of the times of occurrence in the *lunar* day for the first 58 tides of type L1 in 1994. Recall that the lunar day is defined as a period of 24 hours and 50.48 minutes. For definiteness, we have taken the midnight on December 31, 1993 as the origin of the first lunar day in the year.

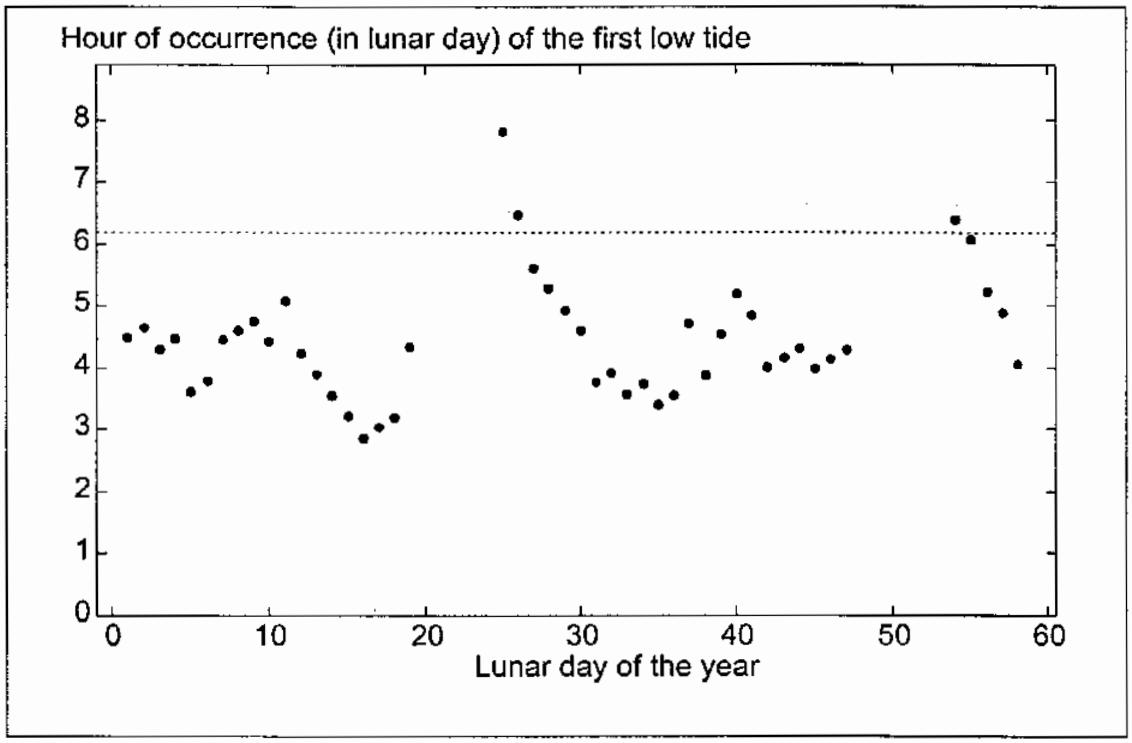


Figure 2.6 Times of occurrence of tide type L1 in the lunar day at Bang Nara in 1994

The information in Figures 2.4 and 2.6 may be combined, by replacing the curve joining the points in Figure 2.6 by two curves in which the vertical distance separating them corresponds to the height of the tide, and the average of the two curves corresponds to the time of occurrence (as in Figure 2.6). This graph may be called *a ribbon graph,* and is a useful method for simultaneously showing both the height of a tide and its time of occurrence. Its construction is shown in Figures 2.7 and 2.8. The thickness of the ribbon is chosen on aesthetic grounds. This method also enables the four tides to be graphed together, and thus easily compared.

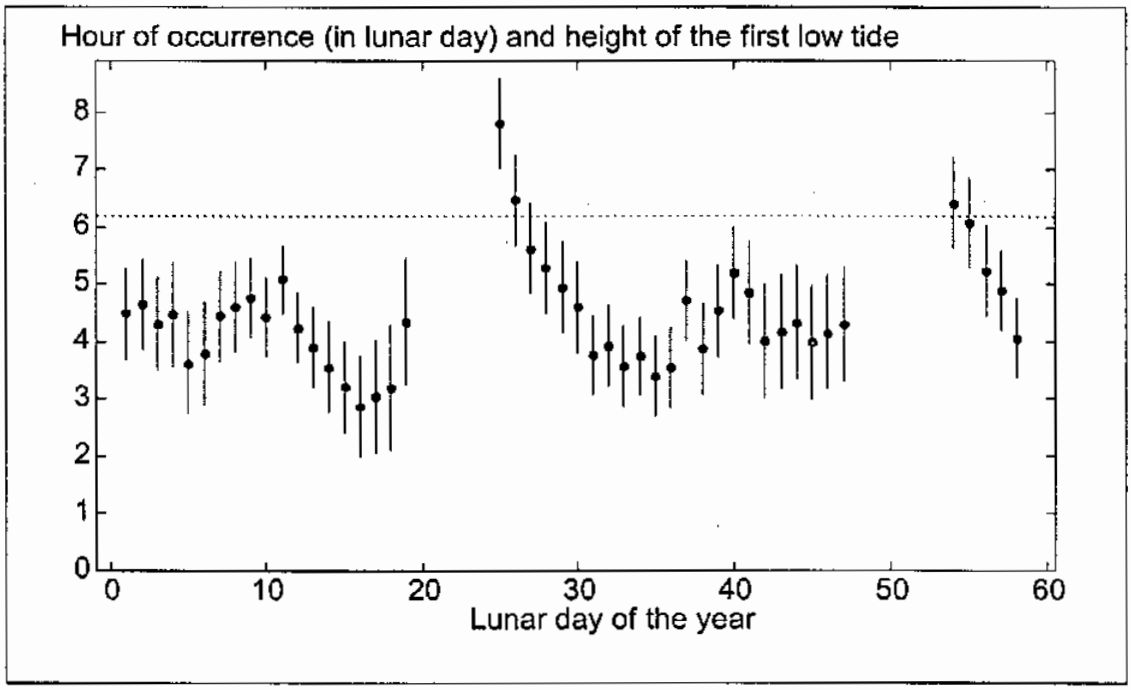


Figure 2.7 Times of occurrence of tide type L1 at Bang Nara during the lunar day with the corresponding heights shown as vertical bars

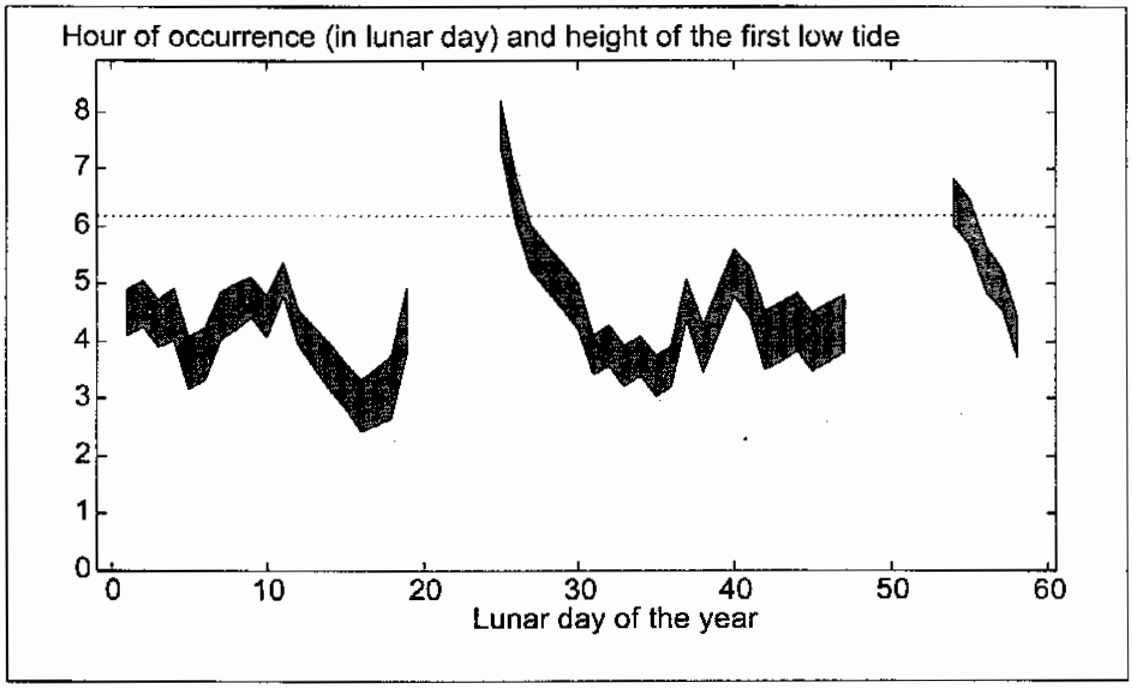


Figure 2.8 Ribbon graph obtained by joining and filling the bars in Figure 2.7

The function *plottide.m* that runs under Matlab is used to produce a ribbon graph. It requires that the data structure be changed so that columns 2-3 become day of year by using a function *days.m*. Files containing the text labels, (eg. bang.dn, bang.fn, bang.lab) are also needed, as shown in Figure 2.9. The function *bang.m.* that runs under Matlab is then used to produce the ribbon graph for Bang Nara.

Note that missing tides (observed from Figure 2.3) are included in the Bang.num file, with heights and time of occurrence denoted as *nan* (Matlab's code for missing data).

bang.num

1	1	1	1	4	4.5
2	1	1	3	6	9.5
3	1	1	0	2	14.5
4	1	1	2	10	22.5
5	1	2	1	4	5.5
6	1	2	3	nan	nan
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
1243	12	31	1	3.5	12.5
1244	12	31	3	11	19.5

bang.fn

ID
month
day
tide type
height
hour

bang.dn

Bang Nara

bang.lab

2,1 Jan,2 Feb,3 Mar,4 Apr,5 May,6 Jun,7 Jul,8 Aug,9 Sep,10 Oct,11 Nov,12 Dec
4,0 L1,1 H1,2 L2,3 H2

Figure 2.9 Data structure used for producing a ribbon graph

3. Statistical modelling

Time series analysis based on fitting harmonic components, is used to model both the heights and time of occurrence of each four tides. This method requires that each outcome comprises a sequence of the form (y1, y2, …, y_n) with no gaps. Thus the method may be applied directly to pure semidiurnal tides. However, mixed tides have gaps (as seen in Figures 2.4 and 2.6), so the method is not directly applicable to them.

A function *tsplot.m* (Asp User's manual, McNeil et al., 1997) is used to fit a time series model to data with no gaps. For data with gaps, an alternative function (*tsplot1.m*) is used. This function uses the E-M algorithm to handle the gaps. The method is described as follow.

3.1 E-M Algorithm

If a time series has missing data the E-M algorithm (Dempster et al., 1977) may be used to fit a harmonic model. This method may be described as follow :

(1) Replace all missing data by the mean of the non-missing data.

(2) Fit the model.

(3) Replace all missing data by corresponding values given by the fitted model.

(4) Repeat steps (2) and (3) until the estimates cover.

The term E-M algorithm arises from the fact that missing data are **E**stimated using the **M**eans given by the model. Figure 2.10 shows an example of a time series (in this case heights of L1 tide at Pak Phun) where the missing values are replaced by the overall mean. Figure 2.11 shows the same data after using the E-M algorithm to estimate the missing values.

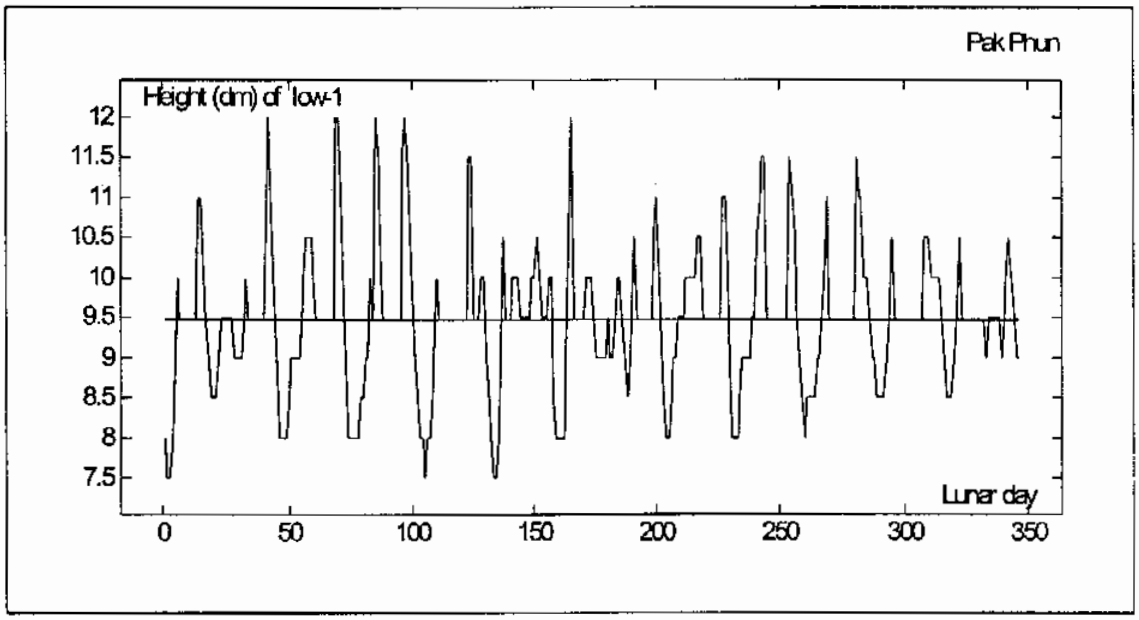


Figure 2.10 Time series of heights of low-1 tide at Pak Phun where missing data were replaced by the mean

Figure 2.11 shows the result of fitting a model containing harmonic components using the E-M algorithm.

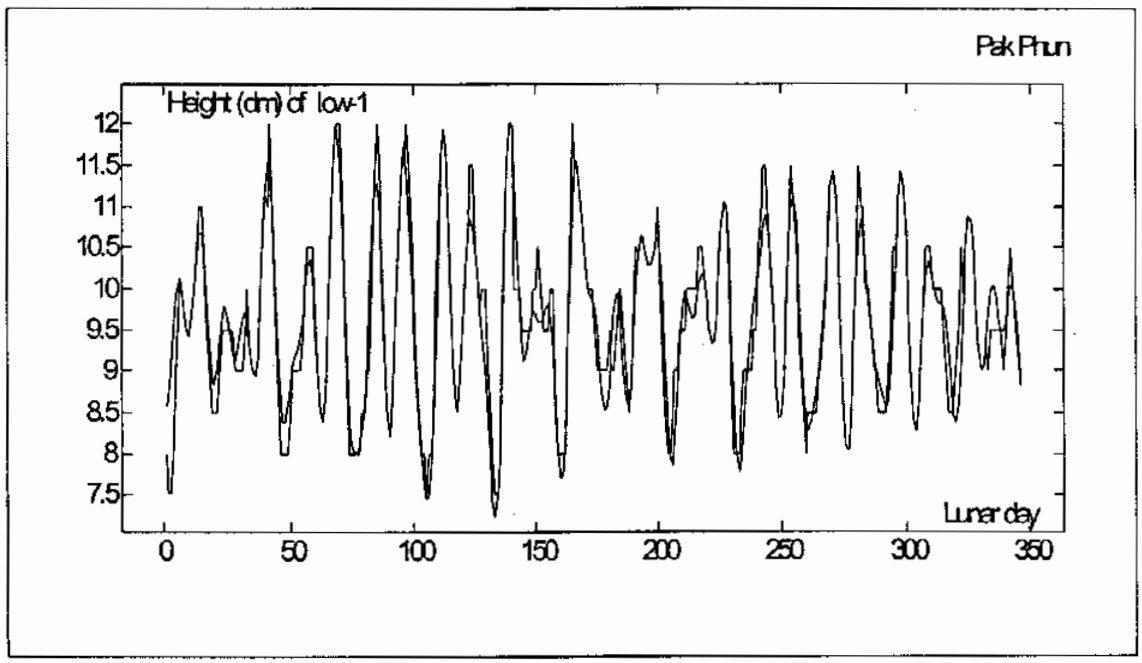


Figure 2.11 Time series of heights of low-1 tide at Pak Phun where missing data were replaced by fitted model

3.2 Time series : harmonic 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 assumed to be 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

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

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 mathematical theory of Fourier analysis any function defined at n equispaced points on a finite interval may be represented exactly by a constant plus 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

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

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

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

where the amplitude $A_j = \sqrt{(a_j^2 + b_j^2)}$ and the summation is from 1 to m (see, for example, Chatfield, 1989).

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 number of data values, there is no residual error, the regression model provides to the perfect data. Moreover, it may be shown that the sum of products of 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 follow.

$$a_0 = \sum y_t /n,\ a_m = \sum (-1)^{t-1} y_t/n,$$

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

It can be seen 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

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

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 a sinusoidal function with frequency j/n. Since the sinusoidal terms are uncorrelated with each other, it follows that

$$\sum(y_t - \sum y_t/n)^2 = \sum(I_j)$$

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 the various harmonics. For this reason it is 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.

Often a periodogram has a large component which dominates the graph to such an extent that it is difficult to discern any pattern in the rest of the periodogram. For this reason, it is better to show the periodogram on a logarithmic scale. Lines on a graph of the base 10 logarithm of the periodogram give 95% confidence intervals for individual periodogram values. Another, possibly more important, reason for graphing the logarithms of the periodogram values is that the variance of the periodogram is stabilized by taking logarithms, making it easier to compare values of the periodogram at different frequencies.

The model fitted to a time series is called the *signal*, and the residual series after subtracting this signal is called the *noise*. To gain some understanding of the periodogram as a graphical method for analysing time series data, it is instructive to look at the periodogram of a purely random series. A series that is purely random, in the sense that future values are completely unpredictable, is called a *white noise* series.

If significantly more than 5% of the periodogram values fall outside the theoretical 2.5 and 97.5 percentile limits, there is evidence that the time series does not behave like a white noise process. If the time series is not a white noise process the height of the periodogram should be different at different frequencies, but the values will still be exponentially distributed, with different means at different frequencies. An exponential distribution has the property that its variance is equal to its mean, whereas the logarithm of an exponential distribution has a constant variance. This is the main reason why graphing the logarithm of the periodogram makes it easier to detect the nature of the harmonic pattern in a time series, which in turn facilitates the modelling process.

So far time series models containing a signal is considered, which could comprise a linear term and one or more harmonics at different frequencies, and a residual or noise series, which could be white noise. If the residual series looks like white noise, it is easy to forecast the future values of the series, provided it is stationary, as follow:

(a) future values of the signal may be forecasted simply by extrapolating the linear and harmonic functions for values of extending beyond the range of the data;

(b) since the (white) noise series is completely unpredictable, it does not contribute to the forecast.

Thus the forecasts of the time series are simply obtained by extrapolating the fitted signal. If the residual series does not resemble white noise, it is of interest to describe its properties, and to have a model that may be used for forecasting (McNeil et al., 1997).

4. Reconstruction of water heights

Having fitted appropriate simple harmonic models separately to the heights and time of occurrence during the lunar day, the tide table may be reconstructed using the model by recombining the components from all four tide types. The steps in this synthesis method are as follow.

(a) Fit the harmonic models. Suppose that $H_i^{(j)}$ is the fitted value of the height of tide type j on lunar day i, and that $T_i^{*(j)}$ is its corresponding time of occurrence during the lunar day.

(b) Convert the fitted times of occurrence to times in hours after midnight on December 31, 1993. Suppose that $T_i^{(j)}$ is the converted time of occurrence for the tide of type j on lunar day i. Since a lunar day is 0.841 hours longer than a solar day, it follows that

$$T_i^{(j)} = T_i^{*(j)} + 0.841\ i$$

(c) Create a vector with two columns of the form (T, H) where T contains all the times of occurrence of all four tides, and H contains the corresponding heights.

(d) These data may now be plotted to give the reconstructed heights (on the vertical axis) versus the time of occurrence.

Note that where tides are missing, as in the case of mixed diurnal/semidiurnal tides, it will be necessary to omit the records in the array (T, H) where these tides are absent.