Chapter 2

Methodology

This chapter describes the data selection, data structure, and methods used in the study. The data selection comprises the variables selected, the source of data, the frequency of data and the period of the study that the data were collected. The data structure comprises the program and methods used in the study.

Data Selection and Sources of Data

The data comprise economic indicators is collected from the weekly magazine *Far Eastern Economic Review*, which is released every Thursday. This magazine contains regional, art & social, business and regular features. The data was taken from a regular feature in this magazine, called *prices & trends*. The data consist of exports, imports, trade balance, international reserves, money supply and economic growth.

This study involves the data for every last week of the month in Thailand and Malaysia. The period of the study is from 1983 to 1996.

Data Structure

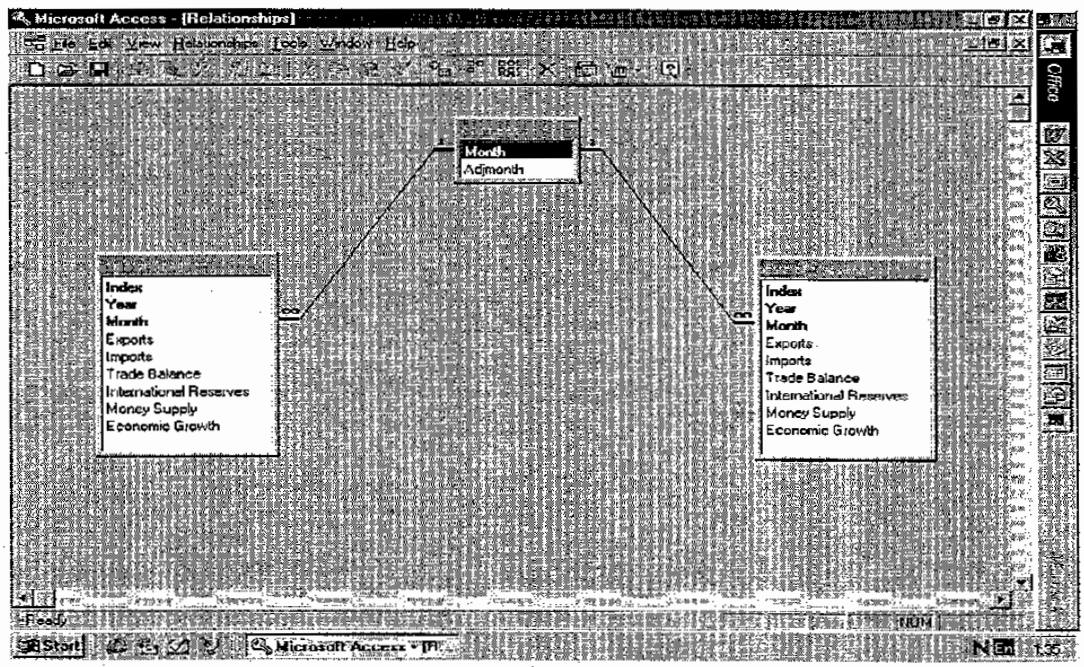
This study used Microsoft Access to store and retrieve the data. The economic indicators data are recorded in a Microsoft Access database file called *dataindicators.mdb*. The structure and relations for the economic indicator data are given in Figure 2.1.

There are three tables in the database, which are *data of Thailand, data of Malaysia* and *calmonth*

The tables data of Thailand and data of Malaysia comprise the variables, index, year, month, exports, imports, international reserves, money supply and economic growth. The table calmonth contain the variables month and adjmonth.

A query is a question that is asked about the information stored in the tables. The way to questions about this information is through use of the query tools. Therefore queries were used to create the data structure appropriate for statistical analysis in this study.

Figure 2.1 Relationships between tables in database



Statistical Methods

The two-sample t-test (see McNeil, 1998b) and time series analysis (see McNeil, 1998a and DeLurgio, 1998) are used in this study. The statistical analysis of the two-sample t-test and time series are also described in many texts such as McNeil (1996), Rakpao (1996), Losunthor (1995), McNeil (1994), Lunn and McNeil (1991), and Box and Jenkins (1976). These methods are summarised as follows.

1. Two-sample t-test

The two-sample t-test takes the form

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{1}$$

In this formula, s is the pooled sample standard deviation, defined by subtracting the sample mean from each sample to give a set of $n_1 + n_2$ residuals, dividing the sum of the squares of these residuals by $n_1 + n_2 - 2$, and taking the square root of the result. If s_1 and s_2 denote the standard deviations of the two samples, respectively, it may be shown that the pooled sample standard deviation is given by the formula

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{2}$$

A p-value is now obtained from the table of two-tailed t distribution with $n_1 + n_2 - 2$ degrees of freedom.

The two samples comprise continuously varying responses and the observations are independent of each other. There are two assumptions. The first assumption is that the two populations from which the samples are drawn have the same spread. The second assumption that is the two populations are normally distributed.

These assumptions may be assessed graphically. Box plots are useful for comparing spreads because they explicitly show the midspread (the distance between the quartiles) for each sample.

The *Asp* function *compar* (See McNeil, 1998) is used to provide a graphical analysis with the two-sample t-test using Matlab Version 5. This command gives a p-value in the (anova) table, confidence intervals for means, as well as box plots.

2. Time Series Analysis

Economic data are appropriate for time series analysis. The data constitute a set of numerical data measured sequentially in time with trends, cycles, seasonal variation and random disturbance or irregular movements. So the statistical methods used for the data analysis are based on time series.

2.1 Time Series Methods

A time series is a continuous set of numerical data measured sequentially in time. The measurements are often equispaced in time or nearly so. Time series data arise in economics and marketing, the physical sciences, engineering, biology and demography, and in many other applications areas.

There are four important objectives of time series analysis. These are (1) forecasting future values of a series, (2) estimating the trend or overall character of a time series, (3) modelling the dynamic relations between two or more time series, and (4) summarising characteristic features of a time series.

A crucial assumption underlying many of the methods used in time series analysis is stationarity, meaning that the statistical properties of the series do not change with time. This means that the mean of the series should be approximately constant and the variability should be homogeneous, or unrelated to its level. If a time series is not stationary, making a transformation may help.

2.2 Common Time Series Patterns

A trend is a general increase or decrease in a time series that persists. Trends are caused by long-term population changes, growth during product and technology introductions, changes in economic conditions, and so on. Trends are not necessarily linear because there are a large number of nonlinear causal influences that yield nonlinear series.

Seasonal series result from events that are periodic and recurrent (e.g., monthly changes recurring each year). Common seasonal influences are climate, human habits, holidays, repeated promotions, new-product announcements, and so on. Seasonality can occur in many different ways, for example, by week of the year,

month of the year, day of the month, day of the week. When seasonal influences are present, seasonal forecasting models should be used.

Cyclical patterns, economic and business expansions (increasing demand) and contractions (recessions and depressions) are the most frequent causes of cyclical influences on time series. These influences most often last for two to five years and recur, but with no known period. In the search for explanations of cyclical movements, many theories have been proposed, including sunspots, positions of the planets, stars, long-wave movements in weather conditions, population life cycles, growth and decay of new products and technology (e.g., phonograph records, tape cassettes), product life cycles, and the economy.

Random time series are the result of many influences that act independently to yield nonsystematic and nonrepeating patterns about some average value. Purely random series have a constant mean and no systematic patterns.

Another pattern that is often seen in time series is a concept called *autocorrelation*. Correlation measures the degree of dependence or association between two variables. The term autocorrelation means that the value of a series in one time period is related to the value of itself in previous periods. With autocorrelation, there is an automatic correlation between observations in a series. Highly positive autocorrelated series without trends or seasonality are often random-walk series. When the mean of the series is always changing, the series is called a nonstationary series.

2.3 Removing a Trend

Many time series have a trend. In these situations it may be useful to fit a straight line, or possibly a quadratic function, and use the residuals as a basis for further statistical analysis. Least squares regression may be used to simply fit a linear or quadratic trend to time series data.

2.4 Periodogram 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 are moves 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 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 characterised by the property that taking a linear transformation of its argument only shifts its frequency and its phase or position relative to some origin.

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, $\cos\{2\pi j(t-1)/n+\phi\}$, $t = 1, 2, \ldots, n$

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

$$Y_t = a_t + \sum\left[a_j \cos\{2\pi j(t-1)/n\} + b_j \sin\{2\pi j(t-1)/n\}\right] + a_m \cos\{\pi(t-1)\} \qquad (3)$$

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

$$y_t = a_0 + \sum A_j \cos\{2\pi j(t-1)/n + \phi_j\} \qquad (4)$$

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 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_j on the corresponding harmonic. The formulas for these coefficients (for n even) are as follows.

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

$$a_j = (2/n)\sum y_t \cos\{2\pi j(t-1)/n\},\ b_0 = (2/n)\sum y_t \sin\{2\pi j(t-1)/m\}$$

These formulas show 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 $\{I_j, j = 1, 2, \ldots, m\}$ is defined in terms of the amplitudes of the harmonics in the Fourier representation as

$$I_j = (n/2)\left(a_j^2 + b_j^2\right) \tag{5}$$

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 thehsinusoidal terms are uncorrelated with each other, it follows that

$$\sum \left(y_t - \sum y_t / n\right)^2 = \sum \left(I_j\right)^2 \tag{6}$$

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 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. A function tsplot in the Asp library (McNeil, 1998a) may be used to show a periodogram of a time series.

2.5 Decomposition of a Time Series

A time series may be written in the form

$$y_t = p_t + s_t + z_t \tag{7}$$

where p_t is a trend (usually linear or quadratic), s_t is a stationary signal having the Fourier series representation given by Equation (6), and z_t is the residual, or noise series. In classical time series analysis, it is assumed that z_t has a normal distribution. In the simplest case, the terms in the process z_t are mutually uncorrelated, in which case the noise is called white noise.

Provided the noise is normally distributed, it may be shown that the periodogram coefficients are exponentially distributed. Now an exponential distribution has the property that its standard deviation is equal to its mean. However, the logarithm of an exponential distribution has approximately constant standard deviation. For this reason, it is useful, when analysing time series data, to plot the logarithm of the periodogram.

2.6 Autoregressive Models

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 spacings. Thus the (auto) correlation coefficient at spacing (or lag) s may be estimated from the formula

$$r_s = \frac{\sum_{t=1}^{n-s} \{y_t - \bar{y}\}\{y_{t+s} - \bar{y}\}}{\sum_{t=1}^{n} \{y_t - \bar{y}\}^2} \tag{8}$$

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.

Due 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 white noise process the theoretical correlation between observations at different spacings is zero, so we 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 periodogram values of a white noise process, being exponentially distributed with constant standard deviation, do not settle down as the length of the series increases. Instead they become more densely packed. Ljung & Box (1978) suggested using the statistic

$$Q = n(n+2)\sum_{s=1}^{m} \frac{r_s^2}{n-s} \tag{9}$$

where m is a specified integer substantially less than the series length n, to test 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.

Now let us consider more general models for describing a noise process z_t. A simple model, involving just a single parameter, takes the form

$$z_t = a_1 z_{t-1} + w_t \tag{10}$$

where w_t is a white noise process. This process is called a simple Markov process, and is characterised by the fact that the best forecast of its next value, z_{t+1} is based only on the current value, z_t. Note that this process reduces to white noise when the parameter is 0. This leads us to consider introducing a second parameter, extending the simple Markov process to the *second-order autoregressive* model, which takes the form

$$z_t = a_1 z_{t-1} + a_2 z_{t-2} + w_t \tag{11}$$

the general auoregressive process of order p takes the form

$$z_t = \sum_{j=1}^{p} a_j z_{t-j} + w_t \tag{12}$$

where the summation goes from j=1 to j=p. It may be shown that an autoregressive process is stationary if and only if all of the roots of the *characteristic* polynomial

$$P(z) = 1 - \sum_{j=1}^{p} a_j z^j \tag{13}$$

are *outside* the unit circle |z|=1 in the plane of complex numbers z. In particular, this means that a simple Markov process is stationary if $|a_1| < 1$. The condition for a second-order autoregressive process is rather more complicated, but it may be shown that necessary and sufficient conditions are

$$a_1 + a_2 < 1,\ a_2 - a_1 < 1,\ |a_2| < 1.$$

Autoregressive moving average (arma) models. The Asp function tsplot used to analyse of a univariate time series. It shows the periodogram, the base 10 logarithm of the periodogram, and thc autocorrelation function of a time series. It also has the capability of removing a linear or quadratic trend, fitting specified harmonics terms, and estimating autoregressive coefficients at specified lags.