Chapter 3

Data and Methodology

3.1 Data

This study uses the daily data of the number of stocks advancing, declining or remaining unchanged in price and the daily closing levels of the stock market indices of the Kuala Lumpur Stock Exchange. The period covered a total of 1233 trading days, from January 1994 to December 1998. All data were obtained from the 'Daily Diary' published by the KLSE. Doubtful data were verified with the Research and Publication Department of the KLSE.

The number of stocks advancing, declining or remaining unchanged in price were from the Main Board, seven main sectors of the Main Board and the Second Board. The seven main sectors of the Main Board were the Consumer Products, Industrial Products, Construction, Trading & Services, Finance, Property and Plantation sectors. Excluded from this study were the data from the Hotel and Mining sectors. The market price indices used were the daily closing indices of the KLSE Emas Index, KLSE Composite Index, KLSE Second Board Index, and the sectoral indices for the Consumer Products, Industrial Products, Construction, Trading & Services, Finance, Property and Plantation sectors of the Main Board.

During the period covered by the analysis, new stocks were added and some disappeared through mergers or were reclassified. Table 3.1 shows the

groups being analyzed indicating how the number of stocks varied over the period investigated.

Table 3.1
Numbers of stocks included in the study

Group	Beginning of study January 2, 1994	End of study December 31, 1998
Consumer Products	55	58
Industrial Products	70	99
Construction	13	33
Trading & Services	54	78
Finance	43	68
Property	45	70
Plantation	43	37
Main Board	393	602
Second Board	85	310

Source: Daily Diary

The raw data of number of stocks advancing, declining or remaining unchanged in price were expressed in proportionate form. The proportions of stocks advancing, declining or remaining unchanged in price at day t are q_{1t}, q_{2t} and q_{3t} respectively, where

$$q_{1t} = \frac{\text{Total number of stocks advancing in price at day t}}{\text{Total number of stocks traded at day t}}$$

$$q_{2t} = \frac{\text{Total number of stocks declining in price at day t}}{\text{Total number of stocks traded at day t}}$$

$$q_{3t} = \frac{\text{Total number of stocks unchanged in price at day t}}{\text{Total number of stocks traded at day t}}$$

Clearly the three proportions sum to one. t =1,2,3,..... stands for consecutive days. These are not always successive calendar days because Saturdays, Sundays and public holidays are non-trading days.

The time series of successive daily difference between proportion of stocks advancing and proportion of stocks declining in price $x_1, x_2, x_3, \ldots\ldots x_t$ is defined by

$$x_t = (q_{1t} - q_{2t})$$

The time series of successive market returns $y_1, y_2, y_3, y_4, \ldots\ldots y_t$ is defined by

$$y_t = \ln V_t - \ln V_{t-1}$$

where V_t is the closing level of the stock market index at period t

3.2 Methodology

3.2.1 Theil-Leenders Test

This statistical test is based on the observed proportions of stocks advancing, declining and remaining unchanged in price over any period. This information theory test serves to examine the extent to which the observed proportions of stocks advancing, declining and remaining unchanged in price

over any trading day can predict the same proportions the following trading day.

Suppose q_{1t}, q_{2t} and q_{3t} are respectively the daily proportions of stocks advancing, declining and remaining unchanged in price over day t and p_{1t}, p_{2t} and p_{3t} are the corresponding predicted proportions. Theil and Leenders (1965) quantified the inaccuracy of the prediction with the "information inaccuracy" measure $I_t(q:p)$ where

$$I_t(q:p) = \sum_{i=1}^{3} q_{it} \log_2 \frac{q_{it}}{p_{it}}$$

The *average information inaccuracy* for the series of prediction in successive periods from t=1 to t=T is obtained as

$$\bar{I}(q:p) = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{3} q_{it} \log_2 \frac{q_{it}}{p_{it}}$$

There are many ways to calculate the predicted proportions. The prediction rule used by Theil and Leenders in specifying p_{it} is given by

$$p_{it} = \alpha q_{i,t-1} + (1 - \alpha)\bar{q}_i$$

where $\bar{q}_i = \frac{1}{T}\sum_{t=1}^{T} q_{it}$ and $0 \le \alpha \le 1$

In this study, the average information inaccuracy $\bar{I}(q:p)$ was calculated for each year from 1994 to 1998 and for the full period of 1994 to 1998 for the Main Board and the Second Board as well as for the main sectors of the Main Board of the Malaysian stock market using values of

α = 0, 0.001, 0.002, ..., 1. The average information inaccuracy $\bar{I}(q:p)$ is a minimum and hence the power of the predictor $p_{i\,t}$ is a maximum at some value of $\alpha = \alpha^*$. If $\alpha^* = 0$, then only the long-run average proportion $\bar{q}_i$ is needed to predict the proportions $q_{i\,t}$. This is consistent with market efficiency. If $\alpha^*=1$, this means that only the last period's proportion $q_{i,\,t-1}$ is needed to predict the present period's proportion q_{it}. This would imply market inefficiency.

3.2.2 Autocorrelation tests

Auto (serial) correlation tests are primarily designed to test the hypothesis that successive values of a variable x_t are random and temporally independent. Temporal independence means that the value the variable assumed in one period is independent of the values it assumed in any other previous period. The runs test, serial correlation test, Ljung-Box-Pierce Q test and von Neumann's ratio test are commonly used to test for serial correlation in a time series.

(i) Runs test

In the runs test, a run is defined as a sequence of successive changes of the same sign of the series. There are three possible types of runs: plus, minus or zero. The length of a run is defined as the number of changes in a run. The actual number of runs in a sequence is then the sum of the number

of runs of changes for plus, minus and zero changes. The actual number of runs is compared with the expected number of runs. If there were too many runs, it would suggest that the signs change frequently thus indicating negative serial correlation. Similarly, if there were too few runs, it would suggest positive autocorrelation. If the hypothesis of no serial correlation is true, the expected number of runs and its variance are given by:

$$E(R) = (n+1) - \frac{\sum_{i=1}^{3} m_i^2}{n}$$

$$\text{Var}(R) = \frac{\sum_{i=1}^{3} m_i^2 [\sum_{i=1}^{3} m_i^2 + n(n+1)] - 2n\sum_{i=1}^{3} m_i^3 - n^3}{n^2(n-1)}$$

where n is the total number of changes and m_i is the number of runs of each type (i=1 for positive changes, i=2 for negative changes, i =3 for no change). The statistical significance of the difference between the actual number of runs and the expected number of runs is tested by using the test statistic Z given by:

$$z = \frac{R + 0.5 - E(R)}{\sqrt{Var(R)}} \qquad \text{if } R < E(R)$$

$$z = \frac{R - 0.5 - E(R)}{\sqrt{Var(R)}} \qquad \text{if } R > E(R)$$

Z is approximately a standard normal distribution. Reject H_o if $z < -z_\alpha$

(ii) Serial Correlation test

Serial correlation coefficient measures the strength of the linear relationship between successive values of the same variable within the same time series. The serial correlation of lag k for the time series $x_1, x_2, \ldots\ldots, x_t$ is given by

$$r_k = \frac{\sum_{t=k+1}^{n}(x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2}$$

where $\bar{x} = \frac{\sum_{t=1}^{n} x_t}{n}$ is the arithmetic mean of the x_t values. If the null hypothesis of no serial correlation is true, the statistic r_k is normally distributed with zero mean and variance of $\frac{1}{(n-k)}$. Reject H_o if $z > z_\alpha$ where $z = r_k\sqrt{n-k}$.

(iii) Ljung-Box-Pierce Q test

To test for no serial correlation at different lags, several separate tests must be done. The significance level of the resulting combinations of tests may be quite different from the significance level for the individual test. A set of m lags may be tested all at once, thereby controlling the significance level by using Ljung-Box-Pierce Q test. Under the null hypothesis of no serial correlation of lags 1 to m, the test statistic Q_m is given by:

$$Q_m = n(n+2)\sum_{k=1}^{m}\frac{r_k^2}{(n-k)}$$

The Q statistic has an approximate chi-square distribution with m degrees of freedom. Reject H_o if $Q_m > \chi^2_{\alpha}(m)$

(iv) von Neumann's ratio test

Another way to test for randomness of the series $x_1, x_2, \ldots\ldots, x_t$ is to use a procedure developed by von Neumann (1941). This test attempts to assess whether a time series has short-run serial correlation. The test statistic for von Neumann's ratio test is given by:

$$V = \frac{\frac{\sum_{t=2}^{n}(x_t - x_{t-1})^2}{(n-1)}}{\frac{\sum_{t=1}^{n}(x_t - \bar{x})^2}{n}}$$

If the null hypothesis of no serial correlation is true, von Neumann has shown that for large n, the ratio V is approximately normally distributed with mean and variance given by:

$$E(V) = \frac{2n}{(n-1)} \qquad Var(V) = \frac{4n^2(n-2)}{(n+1)(n-1)^3}$$

Reject H_o if $z < -z_{\alpha}$ where $z = \frac{V - E(V)}{\sqrt{Var(V)}}$.

3.2.3 Tests for relationship between two series

The difference between the proportion of stocks advancing and proportion of stocks declining in price is an indicator of the market sentiment

and so is the market index returns. The objective here is to examine the short-run relationship between the time series defined by the difference between the proportion of stocks advancing and proportion of stocks declining in price and the time series of the corresponding market index returns.

(i) Cross-correlation Test

The cross-correlation function describes the extent to which two series, x_t and y_t, are correlated. It does exactly the same way the autocorrelation function describes a single series. The cross-correlation function (ccf) estimates the correlation between y_t and the shifted series x_{t-k} or equivalently between y_{t+k} and x_t. The calculation of the cross-correlation function is analogous to that of the autocorrelation function; namely,

$r_{xy}(k)$ = sample cross-correlation coefficient of lag k

$$= \frac{\sum_{t=1}^{n-k}(x_t - \bar{x})(y_{t+k} - \bar{y})}{\sqrt{\sum_{t=1}^{n}(x_t - \bar{x})^2 \sum_{t=1}^{n}(y_t - \bar{y})^2}} \qquad \text{where } k=..-3,-2,-1,0,1,2,...$$

If the null hypothesis that the theoretical cross-correlation function $\rho_{xy}(k)=0$ is true, the statistic $r_{xy}(k)$ is normally distributed with zero mean and variance of $\frac{1}{n-|k|}$

(ii) Granger Causality

Cross-correlation does not necessarily imply causation in any meaningful sense. A number of tests of causality have been developed and one of these is the test proposed by Granger (1969). The Granger approach to whether y_t causes x_t is to see how much of the current x_t can be explained by past values of x_t and to see whether adding lagged values of y_t can improve the explanation. The Granger regression models are as follows:

$$x_t = a_o + \sum_{i=1}^{m} b_i x_{t-i} + \sum_{i=1}^{m} c_i y_{t-i} + \varepsilon_t \qquad (1)$$

$$y_t = a_o + \sum_{i=1}^{m} b_i y_{t-i} + \sum_{i=1}^{m} c_i x_{t-i} + \varepsilon_t \qquad (2)$$

In regression model (1), y_t is said to Granger cause x_t if the coefficients of lagged y_ts are jointly significant in explaining x_t. Similarly, in regression model (2), x_t is said to Granger cause y_t if the coefficients of lagged x_ts are jointly significant in explaining y_t.

The joint significance of the coefficients in each model is tested by means of the F-test. The determination of the lag length, m, of the regression model is based on the minimization of the Schwartz Criterion $\left(\frac{ESS}{T}\right)T^{k/T}$ where T= number of observations, k= number of independent variables including constant and ESS= residual (error) SS. The value of m is restricted

to a maximum of ten and may take different values in model (1) and model (2).

(iii) Dicker-Fuller Test

Before examining the causal relationship between two time series, it is important to pretest each series for stationarity to avoid obtaining results that are spurious when non-stationary variables are used. The Dicker-Fuller (1979) test involves the estimation of an autoregressive equation of the form:

$$R_t = \mu + \beta t + \alpha R_{t-1} + \sum_{i=1}^{p} \theta_i \Delta R_{t-i} + \varepsilon_t \qquad (3)$$

where Δ represents first difference and the error terms, ε_t, are assumed to be normally, identically and independently distributed. R_t represents the series being tested for stationarity. The null hypothesis that $\alpha=1$ can be tested using the Dickey-Fuller t-statistics, τ_α. Under the null hypothesis R_t is non stationary, being a random walk with drift. Under the alternative hypothesis R_t does not contain a unit root and is stationary.