CHAPTER 3
METHODOLOGY

3.1 Data

This study uses daily sectoral stock indices for the KLSE from 29 March 1993 - 30 June 1999. The indices included are the Finance Index, Industrial Index, Plantation Index, Mining Index and Property Index. We explain the selection of these indices in Chapter 4. The relationship among the five sectors of KLSE is examined for the entire sample period and also for different sub-periods. The sub-periods are selected such that each of them represents a different market condition. For this selection, an indicator of the stock market performance is needed. The Composite Index and Emas Index are used as the market performance indicator.

Data were downloaded from the Financial Times databases by using the software Sequencer, available in the Main Library of the University of Malaya. To check for accuracy of the data, they were compared with the data downloaded from the Bank Negara Malaysia's website (http://www1.bnm.gov.my). For any discrepancy found, we checked the daily stock market report published in The Star of recent years. For the earlier years, the data were checked against those reported in the Daily Diary published by the KLSE.[1] The discrepancies we found were mainly due to the Financial Times databases, not taking into consideration the non-trading days. Indices were recorded for some non-trading days, and they

[1] We thank Professor Dr Kok Kim Lian from the Department of Applied Statistics, Faculty of Economics and Administration, University of Malaya, for making this source available to us

are actually readings for the previous day's indices. Other than this, the sectoral indices available from Sequencer are generally accurate.

3.2 Methodology

3.2.1 Tests on Mean and Variances of Daily Returns

Let I_t denote the sectoral index for period t. The sectoral returns are computed as the natural logarithms of index relatives as follows:

$$R_t = \ln\left(\frac{I_t}{I_{t-1}}\right) \qquad (3.1)$$

Chapter 4 describes the characteristics of the data used in this study. The mean and variance of the returns are examined. These statistics are compared for the different sub-periods. Suppose, there are k sub-periods.

We use the t-test to check if the mean return in a sub-period is different from zero. For example, our null hypothesis is that the mean return is equal to zero, and the alternative hypothesis is that the mean return is greater than zero. If we hypothesize that the sub-period has a negative mean return, then our alternative hypothesis is less than zero. The hypotheses are as follows:

$$H_0 : \mu_i = 0 , i = 1, 2, \ldots, k.$$

$$H_1 : \mu_i > 0 \quad \text{or} \quad H_1 : \mu_i < 0$$

The test statistic is given by

$$t = \frac{\bar{x}_i}{\frac{s_i}{\sqrt{n_i}}} \qquad (3.2)$$

where $\bar{x}_i$ is the mean return for sub-period i,

$$s_i = \sqrt{\frac{\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}{n_i - 1}}$$

s_i is the standard deviation of the returns of sub-period i and n_i is the sample size for sub-period i. The t-statistic follows a student t distribution with n_i-1 degrees of freedom under H_0.

In order to determine whether the mean returns of the k sub-periods are equal, we use the F-test. The null and alternative hypotheses are as follows:

$H_0 : \mu_1 = \mu_2 = \ldots\ldots = \mu_k$

H_1 : At least one of the above equations is not true

The F statistics for the test is given by:

$$F = \frac{MSB}{MSW} \qquad (3.3)$$

where

$$MSB = \frac{\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2}{k-1},$$

$$MSW = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}{N-k},$$

$N = \sum_{i=1}^{k} n_i$ and $\bar{x}$ is the mean for the entire sample. The F statistic follows a $F_{k-1,N-k}$ distribution under H_0.

To test for equality among variances of different sub-periods, the Bartlett test is performed. The purpose of this analysis is to compare the sub-period volatility behaviour. The hypotheses continue to be tested is as follows:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$$

H_1 : At least one of the above equations is not true

The test statistic is given by

$$M = 2.3026 \frac{q}{c} \qquad (3.4)$$

where

$$q = (N - k)\log_{10} S_p^2 - \sum_{i=1}^{k}(n_i - 1)\log_{10} S_i^2 ,$$

$$c = 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i - 1} - \frac{1}{N-k}\right),$$

and

$$S_p^2 = \frac{\sum_{i=1}^{k}(n_i - 1)S_i^2}{N - k} .$$

If the sub-period sample variances S_i^2 differ greatly, q will be large. On the other hand, when all S_i^2 are equal, q is equal to zero. The M statistic follows a χ^2 distribution with k-1 degrees of freedom under H_0.

3.2.2 Short and Long-Run Dynamics

The long-run relationship among the different sectors is examined in Chapter 5 and the short-run relationship in Chapter 6. This section describes the methodology involved.

3.2.2.1 Unit Root Tests

Before modeling the sectoral relationship, the first step is to determine the presence of unit roots in each of the sectoral indices involved. This is to examine the stationarity properties of the series. A time series X_t that is stationary is said to be integrated of order zero, or denoted as $X_t \sim I(0)$. A series that requires first order differencing to achieve stationarity is said to be I(1). A series that is I(1) contains a unit root.

A widely applied test for the presence of unit roots is the augmented Dickey-Fuller (ADF) test. The test was proposed by Dickey and Fuller (1979). The test involves testing for presence of a unit root as indicated in the following hypotheses.

$$H_0 : \alpha = 0$$

$$H_1 : \alpha < 0$$

in the equation $$\Delta X_t = \mu + \beta t + \alpha X_{t-1} + \sum_{i=1}^{m} \theta_i \Delta X_{t-i} + \varepsilon_t \quad t = 1,2,\ldots, N \qquad (3.5)$$

where

Δ is the difference operator,

t is the trend term,

m is the number of lags of ΔX_t included and

$\varepsilon_t \sim IN(0, \sigma^2)$.

Lags of ΔX_t are included to account for higher-order serial correlation in the series. It makes a parametric correction by assuming that the X_t series follows an autoregressive process. The test statistic used is the Dickey-Fuller t_α statistic.

The t_α statistic does not follow a student-t distribution under H_0, but its empirical distribution is tabulated by MacKinnon (1991).

We fit equation (3.5) with X_t = $\ln I_t$ to test for stationarity in $\ln I_t$. If H_0 is not rejected, this shows that the series $\ln I_t$ is non-stationary and contains at least one unit root. The series will then have to be differenced before it is tested for stationarity again. We proceed to test for presence of unit roots in $\Delta \ln I_t$ by estimating equation (3.5) with $X_t = \Delta \ln I_t$. If H_0 is rejected, then $\ln I_t \sim I(1)$. Otherwise, we repeat the process with $X_t = \Delta^2 \ln I_t$ and so on until H_0 is rejected. Note also that equation (3.5) contains a drift and deterministic time trend which may possibly exist in the sectoral indices.

The next step is to decide on the size of m, the lag order of the ADF test. We fit equation (3.5) for different lag orders (We used m=1 to m=12). The lag order, which yields the smallest Schwarz criterion, is selected as the optimal lag length. The Schwartz criterion for a single equation is given by

$$SC_1 = \frac{W \log N}{N} + \log\left(\frac{1}{N}\sum_{t=1}^{N} e_t^2\right)$$

where

W is the number of regressors, including constant and

e_t is the residual for the regression estimated for equation (3.5).

The assumption of independently and identically distributed error term underlying the ADF test may not be true. The Phillips-Perron (Phillips and Perron, 1988) procedure offers an alternative for testing presence of unit roots that is invariant to a wide class of serially dependent and heteroscedastic innovations. It uses nonparametric corrections to improve on the ADF test statistic such that it is robust to heterocedasticity and autocorrelation.

The test involves fitting

$$\Delta X_t = \mu + \alpha X_{t-1} + \varepsilon_t \qquad (3.6)$$

The Phillips-Perron t-statistic for testing $H_0: \alpha = 0$ in equation (3.6) is given by

$$t_{pp} = \frac{\gamma_0^{\frac{1}{2}} t_\alpha}{w} - \frac{(w^2 - \gamma_0) N S_\alpha}{2w\hat{\sigma}} \qquad (3.7)$$

where

$$w^2 = \gamma_0 + 2\sum_{j=1}^{m}\left(1 - \frac{j}{m+1}\right)\gamma_j, \quad \gamma_j = \frac{1}{N}\sum_{t=j+1}^{N} e_t e_{t-j},$$

m is the truncation lag, e_t is the residual of the estimated regression for equation (3.6), t_α is the usual t-statistics for testing $H_0: \alpha = 0$, S_α is the estimated standard error of the OLS estimate of α and $\hat{\sigma}$ is the standard error of the regression.

The t_{pp} statistic has a similar limiting distribution to the ADF test statistic in the absence of a deterministic time trend. The critical values for the test can be obtained from MacKinnon (1991).

3.2.2.2 Vector Autoregression (VAR) Model

We consider a VAR model that is used to explain the relationship among the five sectors of the KLSE. The model contains five variables $X_{1t}, X_{2t}, \ldots\ldots,$ and X_{5t}. Suppose that $X_{it} \sim I(1)$, i = 1, 2, ..., 5. The VAR (p) model for five variables is given by

$$\Delta \mathbf{x}_t = \mathbf{a}_0 + \mathbf{a}_1 \Delta \mathbf{x}_{t-1} + \ldots\ldots + \mathbf{a}_p \Delta \mathbf{x}_{t-p} + \varepsilon_t \qquad (3.8)$$

where

$$\Delta \mathbf{x}_t = \begin{bmatrix} \Delta X_{1t} \\ \Delta X_{2t} \\ \vdots \\ \Delta X_{5t} \end{bmatrix} \qquad \mathbf{a}_0 = \begin{bmatrix} a_{01} \\ a_{02} \\ \vdots \\ a_{05} \end{bmatrix} \qquad \varepsilon_t = \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \vdots \\ \varepsilon_{5t} \end{bmatrix}$$

$$\mathbf{a}_j = \begin{bmatrix} a_{11,j} & a_{12,j} & \cdots & a_{15,j} \\ a_{21,j} & a_{22,j} & \cdots & a_{25,j} \\ \vdots & \vdots & & \vdots \\ a_{51,j} & a_{52,j} & \cdots & a_{55,j} \end{bmatrix}, j = 1, 2, \ldots, p \text{ and } \varepsilon_t \sim \text{i.i.d. } N(0,\Omega)$$

The model can be rewritten as:

$$\Delta X_{1t} = a_{01} + \sum_{i=1}^{p} a_{11,i}\Delta X_{1,t-i} + \sum_{i=1}^{p} a_{12,i}\Delta X_{2,t-i} + \sum_{i=1}^{p} a_{13,i}\Delta X_{3,t-i} + \sum_{i=1}^{p} a_{14,i}\Delta X_{4,t-i} + \sum_{i=1}^{p} a_{15,i}\Delta X_{5,t-i} + \varepsilon_{1t}$$

$$\Delta X_{2t} = a_{02} + \sum_{i=1}^{p} a_{21,i}\Delta X_{1,t-i} + \sum_{i=1}^{p} a_{22,i}\Delta X_{2,t-i} + \sum_{i=1}^{p} a_{23,i}\Delta X_{3,t-i} + \sum_{i=1}^{p} a_{24,i}\Delta X_{4,t-i} + \sum_{i=1}^{p} a_{25,i}\Delta X_{5,t-i} + \varepsilon_{2t}$$

$$\vdots$$

$$\Delta X_{5t} = a_{05} + \sum_{i=1}^{p} a_{51,i}\Delta X_{1,t-i} + \sum_{i=1}^{p} a_{52,i}\Delta X_{2,t-i} + \sum_{i=1}^{p} a_{53,i}\Delta X_{3,t-i} + \sum_{i=1}^{p} a_{54,i}\Delta X_{4,t-i} + \sum_{i=1}^{p} a_{55,i}\Delta X_{5,t-i} + \varepsilon_{5t}$$

(3.9)

The VAR model explains the short-run relationship among the five sectors. To determine the lag order p in equation 3.8, we use the Schwarz (1978) criterion for a system of equations given by

$$SC_2 = \frac{-2\ell}{N} + \frac{W \log N}{N}$$

where W is the number of parameters estimated in the VAR model and ℓ is the value of the log-likelihood function evaluated at these W estimates. As this model involves 5 equations, the full system log likelihood is used to compute SC_2. Assuming a multivariate normal distribution,

$$l = \frac{-5N}{2}(1+\log 2\Pi) - \frac{n}{2}\log|\hat{\Omega}|$$

where $|\hat{\Omega}| = \det(\sum e_t e'_t/N)$ and e_t is the vector of residuals for period t.

Using the VAR model, we can test for presence of lead-lag relationship between two different sectors. To test if sector j leads sector i, we use the F-test to test

$$H_0 : a_{ij,1} = a_{ij,2} = \ldots\ldots = a_{ij,p} = 0$$

H_1 : at least one of the above equations is not true.

If H_0 is rejected, sector j is said to Granger cause sector i.

3.2.2.3 Variance Decomposition

Based on the results from Hamilton (Chapter 10, 1994), the forecasting error of the VAR model defined in (3.8) for s periods ahead is given by

$$\Delta \mathbf{x}_{t+s} - \Delta \hat{\mathbf{x}}_{t+s/t} = \varepsilon_{t+s} + \psi_1 \varepsilon_{t+s-1} + \psi_2 \varepsilon_{t+s-2} + \ldots\ldots + \psi_{s-1} \varepsilon_{t+1}$$

where ψ_j is a matrix with elements made up of 0,1, and elements in $\mathbf{a}_1, \mathbf{a}_2, \ldots\mathbf{a}_p$. This matrix is defined after equation (10.1.14) of Hamilton (1994).

Since Ω is a symmetric positive definite matrix, there exists a matrix B and a unique diagonal matrix D with positive entries along the principal diagonal such that

$$\mathbf{\Omega = BDB'}$$

where

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ b_{21} & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ b_{51} & b_{52} & b_{53} & \cdots & 1 \end{bmatrix}$$

Using matrix **B**, we can construct a vector $\mathbf{u}_t$ from $\mathbf{u}_t = \mathbf{B}^{-1}\varepsilon_t$.

We can write

$$\varepsilon_t = \mathbf{B}\mathbf{u}_t = \mathbf{b}_1 u_{1t} + \mathbf{b}_2 u_{2t} + \ldots\ldots + \mathbf{b}_5 u_{5t}$$

where $\mathbf{b}_j$ denotes the j-th column of matrix **B**.

The elements in $\mathbf{u}_t$ are known as orthogonalized innovations. Following equation (11.5.6) of Hamilton (1994), the mean squared error (MSE) of the s period ahead forecast is

$$MSE(\Delta \mathbf{x}_{t+s/t}) = \sum_{j=1}^{5} \{var(\mu_{jt})[\mathbf{b}_j\mathbf{b}'_j + \psi_1\mathbf{b}_j\mathbf{b}'_j\psi_1 + \psi_2\mathbf{b}_j\mathbf{b}'_j\psi_2 + \ldots + \psi_{s-1}\mathbf{b}_j\mathbf{b}'_j\psi_{s-1}]\} \tag{3.10}$$

The contribution of the j-th orthogonalized innovation to the MSE of the s period ahead forecast given by (3.10) is

$$\sum_{j=1}^{5} \{var(\mu_{jt})[\mathbf{b}_j\mathbf{b}'_j + \psi_1\mathbf{b}_j\mathbf{b}'_j\psi_1 + \psi_2\mathbf{b}_j\mathbf{b}'_j\psi_2 + \ldots + \psi_{s-1}\mathbf{b}_j\mathbf{b}'_j\psi_{s-1}]\} \tag{3.11}$$

Thus, the variance decomposition allows us to examine the portion of the total variance of ΔX_t that is due to the innovations of u_t. In other words, we can examine the impact of an innovation in a particular sector on the other sectors in the KLSE. Another way of looking at this technique is that it provides the out of sample causality test.

3.2.2.4 Cointegration and Vector Error Correction (VEC) Model

A linear combination of more than one non-stationary series may result in a stationary relationship. The series are said to be cointegrated which means that there is a long-run equilibrium relationship among the series. For example, all the sectoral returns might be bound by a long-run relationship, which is also known, as the cointegrating equation.

A maximum likelihood test procedure that estimates the multiple cointegrating vectors in a multivariate framework was introduced by Johansen (1991). To discuss the procedure, consider a VEC model given by

$$\Delta x_t = \mu + \Pi x_{t-1} + \Gamma_1 \Delta x_{t-1} + \Gamma_2 \Delta x_{t-2} + \ldots\ldots + \Gamma_p \Delta x_{t-p} + \varepsilon_t \qquad (3.12)$$

where

$$x_t = \begin{bmatrix} X_{1t} \\ X_{2t} \\ \vdots \\ X_{5t} \end{bmatrix} \qquad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_5 \end{bmatrix} \qquad \Pi = \begin{bmatrix} \Pi_{11} & \Pi_{12} & \cdots & \Pi_{15} \\ \Pi_{21} & \Pi_{22} & \cdots & \Pi_{25} \\ \vdots & \vdots & & \vdots \\ \Pi_{51} & \Pi_{52} & \cdots & \Pi_{55} \end{bmatrix}$$

$$\Gamma_j = \begin{bmatrix} c_{11,j} & c_{12,j} & \cdots & c_{15,j} \\ c_{21,j} & c_{22,j} & \cdots & c_{25,j} \\ \vdots & \vdots & & \vdots \\ c_{51,j} & c_{52,j} & \cdots & c_{55,j} \end{bmatrix}, j = 1, 2, \ldots, p$$

Alternatively, the model can be expressed as

$$\Delta X_{1t} = \mu_1 + \sum_{i=1}^{5} \Pi_{1i} X_{i,t-i} + \sum_{i=1}^{p} c_{11,i} \Delta X_{1,t-i} + \sum_{i=1}^{p} c_{12,i} \Delta X_{2,t-i} + \ldots\ldots + \sum_{i=1}^{p} c_{15,i} \Delta X_{5,t-i} + \varepsilon_{1t}$$

$$\Delta X_{2t} = \mu_2 + \sum_{i=1}^{5} \Pi_{2i} X_{i,t-i} + \sum_{i=1}^{p} c_{21,i} \Delta X_{1,t-i} + \sum_{i=1}^{p} c_{22,i} \Delta X_{2,t-i} + \ldots\ldots + \sum_{i=1}^{p} c_{25,i} \Delta X_{5,t-i} + \varepsilon_{2t}$$

$$\Delta X_{5t} = \mu_5 + \sum_{i=1}^{5} \Pi_{5i} X_{i,t-i} + \sum_{i=1}^{p} c_{51,i} \Delta X_{1,t-i} + \sum_{i=1}^{p} c_{52,i} \Delta X_{2,t-i} + \ldots\ldots + \sum_{i=1}^{p} c_{55,i} \Delta X_{5,t-i} + \varepsilon_{5t}$$

(3.13)

The presence of level terms in the equations of the VEC model enables one to test for cointegration by examining the rank of the estimated Π matrix. If the rank of Π is r, where $r < 5$, then there exists r linear independent cointegrating vectors. In general, if we have j series and each series has one unit root, there can be from 0 to j-1 linear cointegrating relations.

Initially, we test for H_0: $r = 0$ or no cointegrating equation against a general alternative of H_1: $r > 0$. If H_0 is rejected, this means that there is at least one cointegration. We will continue to test for H_0: $r = 1$ against H_1: $r > 1$ for existence of one cointegrating equation. If we do not reject the null hypothesis, it means that the system has one cointegrating equation. If we reject the null hypothesis, then we test for H_0: $r = 2$ against H_1: $r > 2$. This process is repeated until a non-rejection is found. If the null hypothesis is rejected all the way, it means Π has full rank and a VAR model should be used on $\mathbf{x}_t$. Once r is known, the number of common stochastic trends in the system is given by j - r.

The statistic used for the cointegration test is the likelihood ratio trace test statistic given by

$$Q_r = -N \sum_{i=r+1}^{5} \log(1 - \lambda_i)$$

where r is the hypothesized number of cointegrating vector under H_0, and λ_i is the i-th largest eigenvalue for C = 0 where

$$C = | \lambda S_{11} - S_{10} S_{00}^{-1} S_{01} |$$
$$S_{00} = N^{-1} \Sigma r_{0t} r'_{0t}$$
$$S_{01} = N^{-1} \Sigma r_{0t} r'_{1t}$$
$$S_{10} = N^{-1} \Sigma r_{1t} r'_{0t}$$
$$S_{11} = N^{-1} \Sigma r_{1t} r'_{1t} \text{ and}$$

r_{0t} and r_{1t} are the residuals from the regression of Δx_t and x_{t-1} on μ and the lags of Δx_t, respectively.

The critical values for the trace test can be obtained from Osterwald-Lenum (1992). If the test indicates existence of r cointegrating vectors, the j x r matrix of eigenvectors corresponding to the r largest eigenvalues give the long-run relationship. The relationship is included in the model (3.12) as Πx_{t-1}. The term is also known as the error correction term (ECT).

The VEC model given by (3.12) implies that changes in the dependent variable are a function of the level of disequilibrium in the cointegrating relationship (captured by the ECT) as well as changes in the other explanatory variables. The VEC model shows the long-run dynamics of the adjustment process among sectoral indices. One ECT is obtained if one cointegrating equation were to be found. If there are two cointegrating equations, then there are two ECTs.

As in the VAR model, we can determine the Granger causality relationship between two different sectors based on the VEC model. The usual F-test can be used to test

$$H_0 : c_{ij,1} = c_{ij,2} = \dots = c_{ij,p} = 0$$

H_1 : at least one of the above equations is not true.

f H_0 is rejected, then sector j is said to lead or Granger cause sector i.

.2.2.5 Forecasting

'he usefulness of the VAR and VEC models is evaluated by examining their orecasting ability. The forecasts for the daily sectoral indices for the month of uly 1999 are obtained. This is carried out for all the sectors. Among the neasures used to evaluate forecast performance are the Mean Absolute)eviation (MAD), Root Mean Squared Error (RMSE), Mean Absolute Percent rror (MAPE) and Theil's U.

The MAD measures the forecast accuracy by averaging the magnitudes of the forecast errors. It is interpreted as the absolute error per day in the forecasting period. Suppose there are s days in the forecasting period. As for the RMSE, it also measures the average error per day in the forecasting period. These measures for sector i are given by

$$MAD_i = \frac{\sum_{t=N+1}^{N+s} \left| X_{it} - \hat{X}_{it} \right|}{s} = \frac{\sum_{t=N+1}^{N+s} |e_{it}|}{s} \quad , i = 1, 2, \dots, 5$$

$$RMSE_i = \sqrt{MSE_i}$$

where

$$MSE_i = \frac{\sum_{t=N+1}^{N+s}\left(X_{it} - \hat{X}_{it}\right)^2}{s} = \frac{\sum_{t=N+1}^{N+s} e_{it}^2}{s}.$$

The MAPE is the average of percentage of relative error. It gives us an indication of how large the forecast errors are in comparison to the actual values of the series. The MAPE for sector i is

$$MAPE_i = \frac{\sum_{t=N+1}^{N+s}\left|\frac{e_{it}}{X_{it}}\right|}{s} \times 100$$

The Theil's U for sector i is given by

$$U_i = \frac{\sqrt{\sum_{t=N+1}^{s} e_{it}^2}}{\sqrt{\sum_{t=N+1}^{s}\left(X_{it} - X_{i,t-1}\right)^2}}$$

If $U_i < 1$, this implies that the model used for forecasting performs better than the no change model. The no change model assumes that the values of the time series are relatively stable from one period to another. Thus, the current value of the series is used as the forecast for the next period, $\hat{X}_{i,t+1} = X_{it}$. $U_i = 0$ implies that this model forecasts perfectly, and it is better than the no change model. $U_i > 1$ implies that the model performs worse than the no change model.