CHAPTER 2

BACKGROUND INFORMATION AND LITERATURE REVIEW

2.1 Random variable, Mean and Variance

A discrete random variable is a variable that takes on only a discrete set of values like 0, 1, 2, 3 .... Examples of discrete random variables are the number of tails when a coin is tossed ten times and the number of children in a family. Suppose that a random variable X takes on N possible values, x_1, ..., x_N, where x_1 denotes the first value, x_2 denotes the second value, etc., and that the probability that X takes on x_1 is p_1, the probability that X takes on x_2 is p_2 and so forth. The expected value of X or its mean, denoted as E(X) is

$$E(X) = \mu_X = x_1p_1 + x_2p_2 + ... + x_Np_N = \sum_{i=1}^{N} x_ip_i$$

The variance of the discrete random variable X, denoted as σ_X^2 is

$$Var(X) = \sigma_X^2 = E[(X - \mu_X)^2] = \sum_{i=1}^{N} (x_i - \mu_X)^2 p_i$$

The standard deviation of X is σ_X, the square root of the variance. If each outcome of X has equal probability (chance), the mean and variance of X become

$$\mu_X = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{and} \quad \sigma_X^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)^2$$

2.2 Autocorrelation and Partial Autocorrelation

A discrete time series is a sequence of data values $\{X_t\}_{t=0}^{N-1}$ observed at particular (say, equally spaced) values of time t. The jth autocovariance of a series X_t is the covariance between X_t and its jth lag, X_{t-j}. That is

$$j^{th}\ \text{autocovariance} = c_k = \text{cov}(Y_t, Y_{t-j}) = \frac{1}{N}\sum_{t=k+1}^{N}(Y_t - \overline{Y})(Y_{t-k} - \overline{Y})$$

$$\text{and } j^{th}\ \text{autocorrelation} = r_k = \text{corr}(Y_t, Y_{t-j}) = \frac{\text{cov}(Y_t, Y_{t-j})}{\sqrt{\text{var}(Y_t)\text{var}(Y_{t-j})}}$$

$$= \frac{\sum_{t=k+1}^{N}(Y_t - \overline{Y})(Y_{t-k} - \overline{Y})}{\sum_{t=1}^{N}(Y_t - \overline{Y})^2}$$

Partial autocorrelation is a measure of the degree of association between X_t and its kth lags, X_{t-k}, when the effect of other time lags (X_{t-1}, X_{t-2}, X_{t-3}, ... $X_{t-(k-1)}$) is removed. The values of the partial autocorrelation coefficients with lag k (ρ_k) are obtained from the iterative solution of Yule-Walker equations.

2.3 Normal Distribution

The most important probability distribution in statistics known as the *Normal* or *Gaussian* distribution has been widely used in modeling random systems. Many random processes occurring in nature actually appear to be normally distributed. In fact, under some moderate conditions, it can be proved that a sum of random variables with any distribution tends toward a normal distribution. The theorem that describes this property is called the central limit theorem. Finally, the normal distribution has some nice properties that make it mathematically tractable and attractive. Given a random process X with mean μ_X and variance σ_X the probability density function of X is given by

$$p(X = x_i) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left[-\frac{1}{2\sigma_x^2}(x_i - \mu_x)^2\right]$$

on the domain $x_i \in (-\infty, \infty)$. The above distribution is fully described by its mean and its variance and any linear function of a normally distributed random process (variable) is also a normally distributed random process.

2.4 Linear Time Series Model Development

The systematic development of modeling and forecasting a time series consists of three main phases namely model identification, model estimation, and model validation. In model identification phase, the input data is first examined and

prepared to ensure that they meet the requirements needed for Box-Jenkins ARMA model. At this stage, a visual inspection of the time plot and the autocorrelation plot of the series would reveal non-stationarity, seasonality, trend and cycle that might be present. Non-stationarity and trend are usually eliminated by differencing once or twice. For seasonality and cycle, once the period is identified, seasonal differencing may remove most or all of the seasonality effect. Information about differencing for non-stationarity and seasonality could actually be incorporated into the general Box-Jenkins ARMA model. When the information is included in the general model, it is known as the ARIMA model and it requires no prior differencing on the time series data before modeling. This is because the differencing operation is already taken care of by the model. However, this inclusion would increase the complexity of the model and it will undeniably affect its accuracy (Hamilton 1994). In fact, removing seasonality prior to modeling may help in the identification of the non-seasonal component of the model. Other methods of removing non-stationarity and seasonality include transformation and decomposition.

Once the series is stationary and free from significant seasonality, its autocorrelation function (ACF) and partial autocorrelation function (PACF) are re-examined. From this examination, the order of the ARMA model (p and q) is estimated. Makridakis et.al (1989) believe that in practice an ARMA model with

values of p and q less than three can cover most of practical forecasting situations.

In the model estimation stage, the values of the unknown parameters are estimated using least squares and maximum likelihood methods. When the model is in simple autoregressive (AR) form, the estimation is essentially a linear regression process. However, when the model is in moving average (MA) form or when it comprises both the autoregressive (AR) and moving average (MA) parts, it requires an iterative non-linear estimation routine. The maximum likelihood estimation technique is generally more preferable due to its desirable mathematical properties.

Finally, in the model validation phase, the accuracy of the estimated model will be tested. First, the autocorrelation function of the simulated series produced by the estimated model can be compared with the autocorrelation function obtained from the original series. If they look markedly different, the validity of the model can be doubtful and a remodeling may be required. If the two autocorrelation functions are similar, an analysis of the residuals of the model can be executed to verify that it is indeed normally distributed and independent. The autocorrelation function of the residuals should be close to 0 for displacement (lag) greater than or equal to one. Obviously, there will usually be more than one model that could be used to model a time series. Measures like adjusted R

square and Akaike Information Criterion (AIC) can be used to select the best model.

It is important to point out that in this study, the analysis has been made using linear forecasts; that is forecasts which are linear combinations of present and past values of the series. Evidently, not all forecasts are linear. There are two main reasons for restricting the analysis to linear forecasts. First, the theory is relatively simple and there are multitudes of softwares that can perform regression for linear forecast for a stationary time series. Second, if the time series is Gaussian (i.e., normally distributed) as assumed, then the best linear forecast is in fact the best of all possible forecasts. No nonlinear forecast can do better in terms of mean squared prediction error. Thus, as long as the series is Gaussian, we need look no further than the linear methods (Hamilton 1994).

2.5 Kalman Filtering Model Development

The Kalman filter is a set of mathematical equations that provides an efficient computational (recursive) solution of the least squares method (Kalman 1960). The filter is very efficient that it can be applied to estimate past, present and future states or values. In the development of Kalman filter model, the same three phases of modeling are applicable. In model identification phase, after the

input data is examined and prepared, a suitable Kalman filter model is selected. If the process that generates the data is assumed to be linear, the standard Kalman filter can be used. If it is nonlinear, the extended Kalman filter (EKF) or other variants of Kalman filter would be more appropriate. Usually, researchers would start their analysis with the standard Kalman filter and if they discover that the results are unsatisfactory, then the extended Kalman filter can be employed. Next, the nature of a number of parameters has to be determined whether they are static or dynamic. In most instances, the initial assumption made is that the parameters are static since this assumption greatly simplifies the filter algorithm unless there is compelling evidence that suggests otherwise.

In the model estimation stage, the values of some unknown parameters are estimated prior to operation of the filter. For instance, the variance of measurement noise can be estimated from some off-line sample measurements. The variance of the process noise is more heuristic and tuning can be done to change its initial value to improve model accuracy. The values of other parameters are assumed to be known either from the laws of physics or from other logical deductions.

Once the values of the parameters are identified, the accuracy of the estimated model will be tested in the model validation phase. A series of state predictions are estimated through the time update and measurement update equations. The

accuracy of these predictions is compared against the actual observations and residuals are calculated. If the values of the residuals are too big (unacceptable), fine tuning is done by changing the values of the parameters of the model until a satisfactory residual level is achieved.

2.6 Survey of Past Work

Makridakis (1976) believes that stock prices follow the random walk or Brownian Motion model and it is nearly impossible to track the direction of the stock market or individual stocks. This idea was proposed by Osborne (1959), Van Horne and Parker (1968). Hong (1978) uses daily data of Stock Exchange of Singapore between October 1975 and April 1976 to show that there is no reason to reject the random walk model. On the other hand, according to Mui (1974), there is a growing tendency among small investors to use technical analysis or charting to forecast movements of share market in Singapore.

In the working paper entitled Neural Network Approach in Predicting KLSE Composite Index, Sanugi (1996) applies a type of neural network called Radial Basic Function (RBF) to predict the KLSE composite index. The author claims that one of the main advantages of employing neural network is its ability to discover the patterns in data which is imperceptible to human visual inspection or

standard statistical methods. Chee (1998) establishes a composite model of the KLSE composite index between 1992 until 1997 by combining multiple linear regression model and the ARIMA time series model. He claims that a slight improvement is obtained by the composite model over the normal Box-Jenkins ARIMA model.

Chang (1990) compares alternative housing investment growth models in Japan, Taiwan, Korea and United States of America over period 1953-1983. It is found that the ARIMA models provide significantly more accurate growth rate forecasts than the traditional models. On the average, mean absolute growth rate forecast error is approximately 10%.The results also indicate that the growth rate time series model for one country can generally be used to forecast another country's growth rate with reasonable accuracy.

Mc Cluskey (1988) studies the different forecasting methods in real property analysis. Qualitative or judgmental methods are useful for short term forecasts and for planning purposes, they assume that the forecaster can determine projections based on extensive knowledge of the business and its environments. Some qualitative techniques deemed useful are executive opinion, formal surveys and market research based methods.

Chong (1992) models Singapore property market using transfer function approach. He discovers that the accuracy of this transfer function model is more superior to the univariate stochastic model. Another merit of the transfer function model is that it requires less data compared to the simultaneous equation modeling. It is concluded that transfer function models give smaller residual standard errors then the univariate models. However, the reduction in standard errors is not very large because of the uncertainty involved in the forecast of the cumulative demand.

Tabak (2003) tests a set of daily Brazilian stock data given by the Sao Paulo Stock Exchange Index (IBOVESPA) in the period of 1986-1998 using random walk hypothesis. It is concluded that prior to 1994 the random walk hypothesis is rejected but after that it cannot be rejected. Institutionally maturing markets, increasing liquidity and the openness of Brazilian markets for international capital are thought to have increased the efficiency of the Brazilin stock market. Evidence suggests that the release of foreign capital control is one of the main factors that increase the efficiency of the Brazilian equity market. Indeed, after 1994 there was a huge inflow of foreign portfolio capital after 1994.

Mcmillan, Speight and Gwilym (2000) analyze a comparative evaluation of the ability of a variety of statistical and econometric models to forecast the volatility of

the UK FTA All Share and FSTE100 stock indices on monthly, weekly and daily basis under symmetric and asymmetric loss functions. A total of ten volatility mean, moving average, random walk, exponential smoothing, exponentially weighted moving average, simple regression, GARCH, TGARCH, EGARCH and component-GARCH models were tested. It was found that under symmetric loss, the random walk model provides superior monthly volatility forecast while random walk, moving average and recursive smoothing models provide better weekly volatility forecasts. GARCH, moving average and exponential smoothing models provide slightly better daily volatility forecasts under the same assumption. When asymmetric loss is considered, the ranking of forecasting methods is dependent on the series, frequency and direction of that asymmetry.

Seddighi and Nian (2004) examine the efficiency of the Chinese stock exchange market operations in the wake of China's entry into the WTO. Tests have been carried out to ascertain whether the share prices in the Chinese stock market follow a random-walk process as required by market efficiency. The empirical results suggest that the prices of some of the shares follow the random walk model while others violate it. It is also found that the Shanghai Stock Exchanges index also follows the random walk pattern. Compared with the developed stock markets in the USA and the UK, the Chinese stock market has a short history of only ten years. It is stipulated that as the Chinese stock market becomes more

integrated with the global economy, the market will become more mature and efficient, playing a more significant role in the global economy.

Chan and McAleer (2003) investigate several empirical issues regarding quasi-maximum likelihood estimation of smooth transition autoregressive (STAR) models with GARCH errors (STAR-GARCH) and the STAR models with smooth transition GARCH errors (STAR_STGARCH). It is shown that different algorithms produce substantially different estimates for the same model. Consequently, the interpretation of the model can differ according to the choice of algorithm. Convergence, the choice of different algorithms for maximizing the likelihood function, and the sensitivity of the estimates to outliers and extreme observations, are examined using daily data for S&P 500, Hang Seng and Nikkei 225 for the period January 1986 to April 2000. The difficulties in evaluating some of these models because of the absence of structural and/or statistical properties, particularly, the regularity conditions for consistency and asymptotic normality, were emphasized.

Chaudhuri and Smiles (2004) present the evidence of long-run relationships between real stock price and real activity which include real price of oil, real GDP, real private consumption and real money supply. The results from the error correction mechanism indicate that real returns are in general related to

changes in real macroeconomic variables along with the deviations from the observed long-run relationships. It is also revealed that other sources of stock return variations such as term spread, future GDP growth rates do not provide substantial additional information, to that which is already contained in the error correction term. However, exploiting other countries' stock return variation, especially that of US market, help explain the stock return in Australian market significantly. Therefore, a proper modeling of long-run stock market dynamics along with the influence of other countries' return variation provide better tracking of the stock price movements in Australian market.

Kalman filter has found widespread applications in aeronautics, navigation and engineering since its introduction in 1960 (Welch 2001). The main advantage of Kalman filter is its ease of application even for large data set. It can handle unobservable variables and can be used for linear as well as non-linear models even when the models have no obvious analytical solution (Maybeck 1979). Harvey (1990) provides an interesting direct comparison between Kalman filters and a number of ARIMA models. However, in deducing the Kalman filters that represent some of the ARIMA models, the meaning of some of the Kalman filter parameters are obscured. Lately, Kalman filter has been used for forecasting in the field of financial economics. Claessans et. al (1995) and Sarno and Taylor (1999) utilized Kalman filter to model the flow of capital from US into several

emerging markets in Asia and Latin America. The model was designed to capture the intrinsic characteristics of the time series without explicit reference to the underlying economic determinants. Mody, Taylor and Kim (2001) employed Kalman filters in analyzing temporary and permanent components of international capital flow to 32 emerging markets around the globe. Statistical models were developed to separate permanent and temporary components for time series for flow of bond, equity and syndicated loans. Results of the Kalman based forecasts were compared to those obtained using fundamentals-based approach. They found that bond, equity and loan flows to emerging markets are largely temporary and reversible in nature and thus renders emerging market assets as relatively risky. Lautier and Galli (2004) compared the performances of the simple and extended Kalman filters in modeling the term structures of commodity prices. They found that the extended Kalman filter produced less accurate estimates than its simple counterpart.