CHAPITRE 3

PROPOSED RESEARCH METHODOLOGY

As indicated in the earlier sections, in deriving IV the daily volatility is converted from a latent variable into an observable one. Traditional financial analytical methods based on parametric models such as the GARCH model family, have been shown to have difficulty to improve the accuracy in volatility forecasting due to their rigid as well as linear structure (Harvey, 1999, Christoffersen *et al.*, 2004). The requirement of a particular distribution assumption further hinders the accuracy of forecasting (Andersen *et al.*, 2001). At the current time, there is just very rare number of publicly available, effective and systematical analytical method in the open literature to deal with the non-linearity inherent in the volatility series of financial indices (Kinlay, 2001, Christoffersen *et al.*, 2000, Kaboudan, 2005, Lee, 2005, Gavrishchaka, 2005). In this regard, we believe that the recent development in financial time series analysis could be beneficial to the forecasting problem, which is the focus of this thesis. Work conducted in (Ma *et al.*, 2004b) has clearly established that by using a GA method, the one-step-ahead moving direction/range of volatility of selected underlying securities could be forecasted at an average accuracy of up to 75%. See references cited in APPENDIX 7 of our own publication list about this work.

3.1 Research Objectives

Given the importance and market opportunity associated with the volatility forecasting, the research efforts covered in this thesis intend to establish a fresh systematic approach and eventually propose an analysis based software tool for financial analysts to forecast more precisely the direction, range and eventually real value of future volatility of financial indices as well as different equities.

In other words, this volatility forecasting method should be capable of dealing with complex signal attributes such as non-linearity and also possess the following advantages, *i.e.* to be:

a) assumption free – no need to assume normal or any other statistical distributions associated with the time series and its estimation errors;
b) more flexible – not limited by the parametric structure;
c) more accurate on the current and hence the forecasted volatility – the conversion of IV transforms volatility from a latent variable into an observable variable;
d) more efficient – data pre-processed by means of wavelet transform.

More specifically, the goals to be attempted would be that

a) to establish, from an EA perspective the rationale for the conversion of a typical time series into a four-lag recursive data set by employing a recently available data mining technique;
b) to establish the rationale for the use of GA in forecasting time series by applying a Markov chain based discrete stochastic optimization method;
c) by using wavelet transform to verify the hypothesis that volatility clustering also occurs at the scale levels;
d) to verify that wavelet transform can improve the efficiency of GA/GP in search of different patterns hidden in noise;
e) to verify that EA can effectively simulate/forecast the volatility time series by minimizing the inaccuracy caused by non-linearity;
f) to demonstrate that the current IV-wavelet-EA method is at least as accurate as the proprietary approach shown in Section 2.2.2 in the volatility forecasting.

3.2 Research Plan and Procedure

As the volatility time series being pre-processed into IV and then applying the wavelet transformation, by treating the wavelet coefficients as a 4-lag recursive data set, we can then apply simple IF/THEN rules with GA's in order to capture the typical patterns most frequently found in the resulting data set. Since wavelet coefficients are far smaller in size compared with the original time series, certain degree of calculation efficiency could be achieved. A hundred rules in the wavelet domain would cover a much broader range than the same number of rules in the time domain alone. Moreover, regularities, if any in the frequency domain could be detected, which adds considerably to the prediction power.

To account for non-linearity in the time series *i.e.,* to represent any abrupt changes or discontinuities, appropriate non-linear functions can be introduced when applying GP to forecast the one-step-ahead values of volatility. Thus, the general strategy is to judiciously use a combined EA approach to obtain a better forecasting volatility model by incorporating non-linearity and abrupt structural changes in the time series. The block diagram in Fig. 1 summarizes the proposed forecasting architecture to solve the volatility problem.

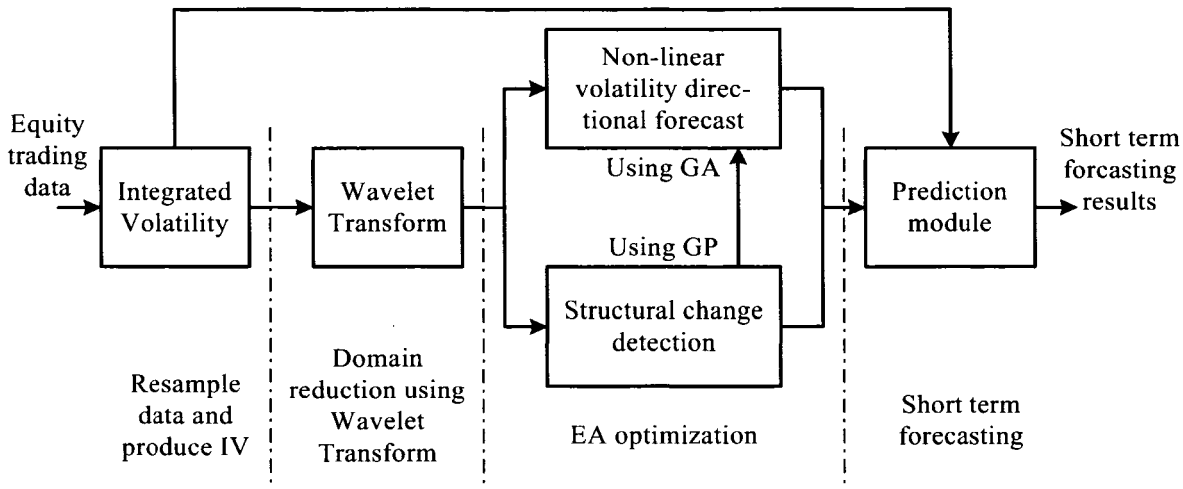

Figure 1 Proposed volatility directional forecasting system architecture

More specifically, the overall methodology for this thesis research can be represented in the form of Fig. 2:

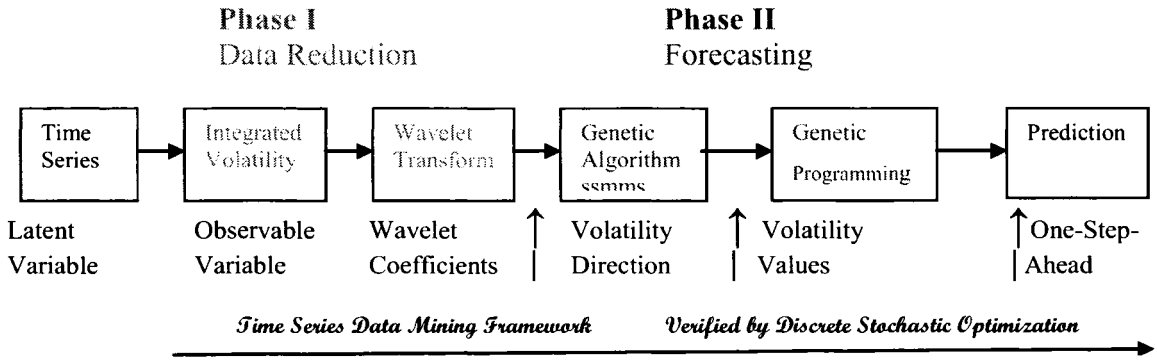

Figure 2 Research plan for this thesis work

The above figure characterizes the rationale of the method implemented throughout the current research work *i.e.*, the application of a combination of data reduction and analysis techniques which converts a difficult volatility forecasting problem into a classical signal analysis problem. And this represents one of the main contributions made in this thesis research effort.

It is apparent that most of the contemporary researchers use EA in time series analysis simply as a tool without much rigorous analytical substantiation. EA, be it GA or GP, typically lacks rigorous mathematical proof and justification even though it is a powerful tool for optimization applications. In order to lay the foundation for applying EA to construct the IV-wavelet-EA time series forecasting approach, the IV time series is first converted into a four-lag recursive series in the TSDM framework. The resultant series is in fact a Markov chain and could be analyzed with a discrete stochastic optimization method in conjunction with GA. Since GP is introduced in the TSDM framework, its operation is better illustrated, whereas its use to forecast IV is also built on a more solid analytical foundation.

The S&P100 and S&P500 indices are selected here for the investigation, not only because of the availability of VIX, hence the convenience of a direct reference, but also because of the popularity of equity and the associated option applications which affords the ease of and usefulness in future portfolio design and trading (Kinlay, 2005). One of the unique characteristics of equity trading is that they are traded only during the prescribed working hours of the working days. An alternative would be to use S&P100 futures (VXB), which is traded both day and night. Martens (2002) actually found that volatility models perform better when index futures instead of indices themselves are used due to the relatively more continuous nature of the data set. However, in this research work, we are more interested in finding ways for dealing with jumps and drops in time series. As a result, it is preferable to use the daily estimation of IV based on indices themselves. In other words, we will calculate IV based on the intra-day data and then use wavelet transform and the EA approach to conduct forecast estimates for the next time step. In the future, Barndorff-Nielsen's working paper (2004) could help select sampling frequencies to avoid market microstructure problems. The research covered in this thesis intends to forecast not only directions and ranges, but as well as values of the volatility movement. For validation purposes, the S&P500 index is obtained from a

different source and analyzed using a similar approach. The results are compared with those derived with the S&P100 index and conclusions are then drawn for suitability of application of the developed methods.

3.3 Data Mining of the IV Time Series

Data mining is the term often employed to represent the process of analysis of data with the goal of uncovering hidden patterns especially those complex relationships in large data sets (Povinelli, 1999). Weiss *et al.* (2004) define data mining as *"the search for valuable information in large volumes of data. Predictive data mining is a search for very strong patterns in big data that can generalize to accurate future decisions."* Similarly, Cabena *et al.* (1994) define it as *"the process of extracting previously unknown, valid, and actionable information from large databases and then using the information to make crucial business decisions."* Data mining evolves from several scientific and computational fields, including machine learning, statistics, and database design. It uses techniques such as clustering, association rules, visualization, decision trees, nonlinear regression, and probabilistic graphical dependency models to identify novel, hidden, and useful structures in large databases (Weiss *et al.,* 2004).

The approach adopted in contemporary literature often requires *a priori* knowledge of the types of structures or temporal patterns to be discovered and represents these temporal patterns as a set of templates. Their use of predefined templates completely prevents the achievement of the basic data mining goal of discovering useful, novel, and hidden temporal patterns. Prompted by shortcomings of parametric models, new class of methods has been created so that they do not rely on pre-assumed models but instead try to uncover/induce the model, or a process of computing values from vast quantities of historic data. Many of them utilize learning methods of CI. Non-parametric approaches are particularly useful when parametric solution either lead to bias, or are too complex to use, or do not exist at all. Many recent publications attempted to use GA's and/or GP's

to forecast financial time series such as stocks, indices and options. Refer to Table 2.1 for a summary of important related works (Chen and Lee, 1998, Goldber, 1998, Chen, 2003, Iba, 1999).

In search of a relatively more general approach for data mining of financial time series, Szpiro's (1997) method permits the discovery of equations of the data-generating process in a *symbolic* form. The GA that is described there uses parts of equations as building blocks to breed ever better formulae. Apart from furnishing a deeper understanding of the dynamics of a process, the method also permits global predictions and forecasts.

Povinelli and Feng (1999) took the subject one step further by introducing the TSDM framework, which differs fundamentally from most of the approaches mentioned above. Povinelli (1999) formulated the TSDM framework that reveals hidden temporal patterns that are characteristic and predictive of time series events. This contrasts with other time series analysis techniques, which characterize and predict all observations.

From a critical review of the above literature, instead of using just one single formula to represent the entire time series, a better idea would be to use multiple GA's and GP's exploring sequentially in the search space to obtain an overall estimation represented by a set of formula rules. For example, the GP's try to find the best fitting rules based on the input time series. The best formula rules are then combined as suggested by Chan and Stolfo (1996) and later used to forecast the future IV values. The same can be applied to rules in the GA approach.

3.3.1 Rule-based Evolutionary Algorithm Forecasting Method

In this section, the necessary background information regarding Povinelli's (1999) TSDM methods is introduced to evoke a good understanding of our research direction.

The TSDM methods create a new structure for analyzing time series by adapting concepts from data mining, time series analysis, EA, and nonlinear dynamics system (Iba, 1999). They are tactfully designed to predict non-stationary, non-periodic and irregular time series, and not restricted by the use of predefined templates. More specifically, they help discover hidden temporal structures predictive of sharp movements in time series, using a time-delay embedding process that reconstructs the time series into a phase space that is topologically equivalent to the original system under certain assumptions (Iba, 1999). The TSDM methods are developed and applicable to yield one-step predictions for time series data sets (Povinelli, 2000). In order to extract non-stationary temporal patterns, a specific TSDM method could be used to address quasi-stationary temporal patterns *i.e.*, temporal patterns that are characteristic and predictive of events for a limited time window Q. It is called the Time Series Data Mining evolving temporal pattern (TSDMe2) method, which uses a fixed training window and a single period prediction window. The TSDMe2 method differs from the other TSDM methods in how the observed and testing time series are formed. The TSDMe2 method creates the overlapping observed time series :

$$\theta = \{\theta_t, t = j, \mathrm{K}\ , j+N\}. \tag{6}$$

The testing time series is formed from a single observation:

$$\eta = \{\theta_t, t = j+N+1\}, \tag{7}$$

Where θ_j is the time series value at time $t=j$, while N is the size of the window. The TSDM*e2* method was created for discovering multiple temporal pattern clusters in a time series (Povinelli, 1999).

In characterizing different embedded patterns hidden in the time series, there are two key factors to consider, number of pattern types and size of the patterns (or windows Q). By parsimony, the simplest characterization of events possible is desired *i.e.*, as small a dimensional phase space Q as possible and as few characterization patterns as necessary. However, the following modifications have been made to the typical TSDM in order to implement the proposed data mining procedure,

a) to increase the pattern characterizations by involving as many as 100 different arithmetic expressions to describe a windowed time series;
b) to use a 4-lag recursive memory as the size of the patterns Q. For definitions of some related concepts, refer to APPENDIX 4 at the end of this thesis and the paper by Povinelli (1999).

There will be many different patterns present in financial time series, both linear and nonlinear types. A financial series is a dynamic entity which is affected by many variables, economic, financial, politics, psychological, legal, *etc.* It is philosophically unwise to use one fixed model, linear or nonlinear to estimate such a process, let alone for forecasting. In using statistical models to estimate time series, wherever abrupt structural changes the model will need to adjust and change its parameters. Furthermore, the application of volatility estimation in option trading deem necessary to extract also the non-event, so that one could capitalize on the time value of the option. In our case, we used 100 different formula/rules to match the frequently appearing events and to extract different patterns buried in noise. Therefore, there is a high probability to extract the patterns and further to forecast the one step-ahead activity. Note that in each of the 100 rules used to characterize different patterns, the value of δ could be considered as the margin of accuracy the rules match the points in the window. In case of GA's, since the data will be divided into four ranges δ is not applicable. Details could be found in the following paragraphs as well as in Section 5.4.

The patterns are determined by the four previous points to make a prediction for the value at the next time interval. The 4-lag recursive system is used due to the potential weekly seasonality of the IV time series as well as the convenience of weekly option trading. Using 4-lag approach could save time and memory and is particularly useful in dealing with volatility forecasting because of past research showing that most of the information is contained in the most recent lags, resulting the popularity of GARCH(1,1)

or other short memory models. One of the contributions of the current research is in dealing with characterizing the non-events. This is especially important in the forecast of volatility. For example, let us assume that one could accurately predict that no major change in the volatility of S&P100 index in the following week. The investor could sell a number of contracts on the strike or near the strike price of the S&P100 option with the shortest expiration available in order to earn the time value.

The current method involves matching clustering patterns, which include sharp fluctuations in the IV series. To find these temporal patterns the time series is embedded into a reconstructed phase space with a time delay of one and a dimension of four (Povinelli, 1999). Once the data is embedded, temporal structures are located using a GA/GP search. Clusters are made of points within a fixed distance of the temporal structures δ. In case of using GP, the event characterization function $g(t)= \theta_{t+1}$, determines the value given to the prediction made from the clustering using the temporal structures. This value is the IV value for the next time interval. The temporal structures are next ordered by how well each predicts the IV movements. A ranking function is defined as the average value within a temporal structure, and it is used to order the structures for optimization. The optimization is a search to find the best set of temporal structures and is done with EA that finds fitness value parameters that maximize the ranking function – the frequency of the correctly guessed patterns in case of GA or minimize the fitness function $f(P)$ – the difference compared with the guessed patterns in case of GP. This leads to the Eq. (8) given below :

$$\min_{P,\delta} f(P)$$
$$f(P) = \sum_{t=5}^{N} (\sigma_t - \theta_t), \quad t = 5,\mathrm{K}\ ,N \tag{8}$$

where σ_t is the forecasted volatility and θ_t is a value in the series to be forecasted. The EA uses a combination of Monte Carlo search for population initialization with a fixed percentage selection, crossover and mutation to find the optimal P^*, and a limited number of generations to halt the genetic programming (Povinelli, 1999, 2000).

When GA is used, the IV window $\theta = \{ \theta_{t+j}, j=0, 1, 2, 3\}$ will be converted into a set of numbers {1, 2, 3, 4, *} by classifying the range as $\{(-\infty, -a], (-a, b], (b, c], (d, \infty), *\}$, where '*' means "don't care". Therefore, all data will become a sequence of numbers. The rules will take the form of <IF [(($\theta_t = I$) AND/OR ($\theta_{t+1} = J$) AND/OR (($\theta_{t+2} = K$)) AND/OR ($\theta_{t+3} = L$))], THEN ($\theta_{t+4} = M$)>, where the event characterization function $g(t) = \theta_{t+4}$ will be a number that predicts the range of the subsequent IV value. And δ will become obsolete. The key difference between using GP and GA is the form of the rule. More details could be found in Chapter 5 where both techniques are applied respectively to the current problem.

In general, such combination can potentially eliminate the erroneous predictions caused by noise in the data. During the rule learning process, independent trials of the GA/GP allow many rules simultaneously to explore different parts of the search space, thereby learning different types of patterns for yielding a prediction. As a result, at a given time some rules generate better predictors than others, thus making them ideal candidates as base predictors to achieve increased predictive accuracy.