Personalised Modelling for Multiple Time-Series Data Prediction: A Preliminary Investigation in Asia Pacific Stock Market Indexes Movement

Harya Widiputra[1*], Russel Pears[1], Nikola Kasabov[1]

[1] *Knowledge Engineering and Discovery Research Institute, Auckland University of Technology, Private Bag 92006, Auckland 1020, New Zealand*

Abstract

The behaviour of multiple stock markets can be described within the framework of complex dynamic systems (CDS). Using a global model with the Kalman Filter we are able to extract the dynamic interaction network (DIN) of these markets. The model was shown to successfully capture interactions between stock markets in the long term. In this study we investigate the effectiveness of two different personalised modelling approaches to multiple stock market prediction. Preliminary results from this study show that the personalised modelling approach when applied to the rate of change of the stock market index is better able to capture recurring trends that tend to occur with stock market data.

Keywords: complex dynamic systems, global model, dynamic interaction networks, multiple time-series data prediction, personalised model.

1. Introduction

Prediction of multiple time-series data is a challenging task. In this paper we study the interactions that occur within several stock markets within the Asia Pacific region. Previous research has shown that a given stock market index is influenced by movement in the indices of certain other markets (Antoniou 2003), (Collins 2003). With the use of techniques such as multiple linear regression a global model can be built that captures long-term trends to some degree of accuracy. However, such global models are not suitable for short-term predictions, particularly in the case of stock market index data. This task serves as a motivation for our work, and we apply methods adopted in the bioinformatics domain to this problem. Some recent studies in the bioinformatics domain used a personalised modelling approach to predict the behaviour of a patient by looking for similar conditions (by comparing features of the patient) from past patient data (Chan 2006). In this study we associate the patient condition with a current stock market index value and then search for similar conditions based on selected features from historical stock market index data. The similar patterns found from the past are used to predict the movement of multiple stock market indexes.

In our previous work we revealed interactions between stock markets in Asia Pacific which captured the global trend of interdependence between stock markets on a long-term basis. In this study we apply a personalised model to multiple time-series financial data collected on a daily basis (shorter period). We analysed the performance of personalised modelling by using raw values of stock market indexes in Asia Pacific as well the velocity of change between stock market over a period of time.

The study is focused on markets in the Asia Pacific region and includes the stock indexes of Hong Kong, New Zealand, Australia, Japan, Indonesia, Malaysia, Singapore, Korea, Taiwan, and China. In section 2, we briefly describe global model and personalised models that can be used for multiple time-series prediction. Section 3 presents the results from the application of both global and personalised models. Finally, conclusions and directions for future research are outlined in section 4.

2. Global Model and Personalised Models

A Global model is an implementation of inductive reasoning approach, where the approach is concerned with the creation of a model from all available data, representing the entire problem space. Such models are effective in capturing global trends which can be used to provide generalized solutions over the whole problem space. Transductive inference, introduced by Vapnik (1998) and it used by Kasabov (2007a) is defined in contrast as a method used to estimate the value of a potential model only for a single point of space by utilizing additional information related to that point. This type of inference is suitable for building personalised models.

While the inductive approach is useful when a global solution of the problem is needed in an approximate form, the transductive approach is more appropriate for applications where the focus is

*Corresponding author: email address: wpj6371@aut.ac.nz.

not on the model, but rather on a specific case or vector. This is very relevant to the movement of stock market indexes value where the changes of values tend to be specific from time to time. Using a transductive approach fits the common sense principle which states that to solve a given problem one should avoid solving a more general problem as an intermediate step.

2.1. Dynamic Interaction Networks as Global Models

Existing research in bioinformatics (Chan, 2006 and D'haeseleer, 1999) outlines the steps in extracting gene regulatory networks from time-course gene expression data, and shows that genetic network inference can be used successfully to reverse engineer the underlying regulatory interactions from the gene expression data. In our previous work we extracted dynamic interaction networks from interrelated time-series data representing closing stock index values in markets form the Asia Pacific region. Through capturing dynamic influences among markets, we were be able to predict the behaviour of multiple stock indexes. The selected ten markets in the Asia Pacific region include Australia (AORD), China (SSE), Hong Kong (HSIX), Indonesia (JSX), Japan (N225), Malaysia (KLSE), New Zealand (NZ50), Singapore (STI), South Korea (KOSPI), and Taiwan (TSEC).

The algorithm to extract the interaction networks in our previous work is based on the Kalman Filter. The algorithm targets the general problem of estimating the state $x \in \Re^n$ of a discrete-time controlled process governed by the linear stochastic difference equation (1),

$$x_t = Ax_{t-1} + Bu_{t-1} + w_{t-1}, \tag{1}$$

The $n \times n$ matrix A in equation (1) relates the state at time $t-1$ to the state at time t, in the absence of either a driving function or process noise. The Kalman Filter (KF) estimates a process by using a form of feedback control (Welch, 2006). Transition matrix A is used to extract the DIN model from the analysed interrelated time-series. The influence matrix is extracted from Matrix A by setting a threshold value that classifies interactions into three categories: positive (+), negative (-) or negligible (0). The dynamic interaction network extracted from the transition matrix is shown in Figure 1 (c). The three markets illustrated in the DIN are: China, Taiwan, and Hong Kong.

(a)

$$\begin{bmatrix} 0.53 & -0.23 & 0.36 \\ -0.01 & 0.87 & 0.43 \\ 0.52 & -0.02 & 0.31 \end{bmatrix}$$

(b)

$$\begin{bmatrix} + & - & + \\ 0 & + & + \\ + & 0 & + \end{bmatrix}$$

(c)

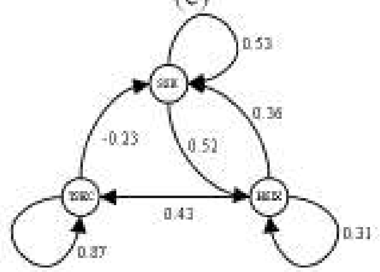

Figure 1: *Constructing a DIN model*
(a) the transition matrix *A*;
(b) the corresponding influence matrix;
(c) the network diagram.

In this study we implement an online learning process, whereby the transition matrix and the DIN model are updated at any time step. Each time step contains the average weekly stock market for each of the 10 markets that we selected. As shown in Figure 2, a model is trained over a certain number of weeks, e.g. 50 weeks of data. This model is then used to predict the next week's index values. The model is then updated with the actual data, and the updated model is used to predict index values for the next week. This incremental process continues into the future, with training and testing takes being interleaved with each other on newly arriving data.

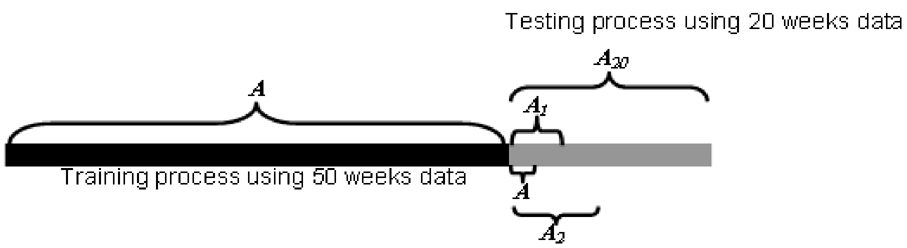

Figure 2: *Prediction with online learning*

2.2. K-Nearest Neighbour (*K*-NN) and Weighted K-Nearest Neighbour as Personalised Model

A personalised model is created on the fly for every new input vector and this individual model is based on the closest data samples to the new sample taken from a data set. One of the simplest and most widely used approaches to personalised modelling is the *K*-NN (*K*-nearest neighbour) method, where for every new sample, the nearest *K* samples are derived from a data set using a distance measure to define similarity, which is usually based on the Euclidean distance. In the *K*-NN method, the output value y_i for a new vector x_i is calculated as the average of the output values of k nearest samples from data set D_i.

Another method that can be used to increase the performance is known as the weighted *K*-NN (WKNN), where the output y_i is calculated based not only on the output values (e.g. class label) y_j of the *K*-NN samples, but also on a weight w_j, that depends on the distance of them to x_i

$$y_i = \frac{\sum_{j=1}^{N_i} w_j y_j}{\sum_{j=1}^{N_i} w_j}, \tag{2}$$

where y_j is the output value for sample x_j from D_i and w_j are their weights calculated based on the distance from the new input vector

$$w_j = [\max(d) - (d_j - \min(d))] / \max(d). \tag{3}$$

The vector $\boldsymbol{d} = [d_1, d_2, \ldots, d_{Ni}]$ is defined as the distances between the new input vector x_i and the N_i nearest neighbours (x_i, y_j) for j = 1 to N_i

$$d_j = sqrt[sum_{l=1\,to\,V}(x_{i,l} - x_{j,l})^2], \tag{4}$$

where V is the number of the input variables defining the dimensionality of the problem space; $x_{i,l}$ and $x_{j,l}$ are the values of variable x_l in vectors x_i and x_j, respectively. The parameters max($\boldsymbol{d}$) and min($\boldsymbol{d}$) are the maximum and minimum values in $\boldsymbol{d}$ respectively. The weights w_j have the values between min($\boldsymbol{d}$)/max($\boldsymbol{d}$) and 1; the sample with the minimum distance to the new input vector has the weight value of 1, and it has the value of min($\boldsymbol{d}$)/max($\boldsymbol{d}$) in case of maximum distance.

3. Experiments and Results

Our dataset consist of weekly averages for use with global modelling, and daily data for the personalised model and spans 70 weeks or 720 days from June 2005 to June 2007. We normalize the data to lie in the range [-1,1], and for personalised modelling we further pre-process the data by removing the linear trend. Figures 3 and 4 illustrate the trajectories produced from these two datasets.

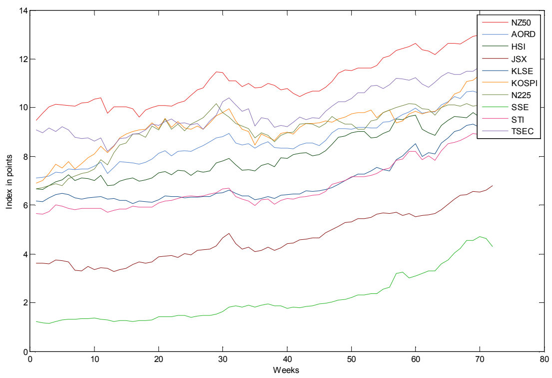

Figure 3: *Normalized weekly stock market*

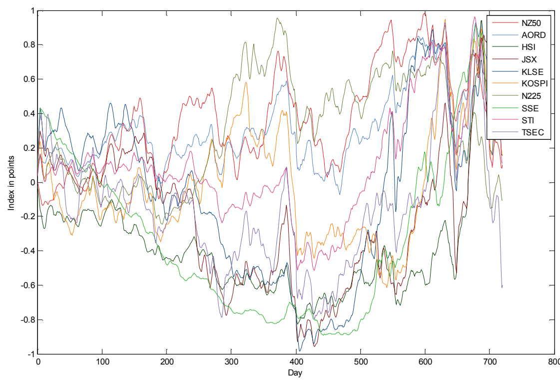

Figure 4: *Normalized and trend-removed daily stock*

3.1. Dynamic Interaction Network Model Analysis

To extract a reliable model for Dynamic Interaction Network, we run 10 different trials on the training dataset and average each of the values in the transition matrix before constructing the dynamic interaction network. The results of the modelling process are: (1) the transition matrix, and (2) the DIN diagram of the interactive stock markets.

Figure 5 presents the results of the DIN modelling process. We test the extracted DIN model to predict future values of multiple stock indexes. The test dataset covers twenty weeks, following on chronologically from the 50 weeks of training data. First, we consider the goodness of fit of the DIN model at all 20 points of the test dataset, plotting actual test trajectories and simulated trajectories. Second, we calculate the root mean squared error to evaluate the performance of DIN prediction. Table 1 clearly shows that the estimated trajectories, based on the extracted DIN, closely track the actual trajectories. We can thus infer that the first-order Kalman Filter difference equations are a good approximator for predicting long trends in stock market indexes..

(a)

$$
\begin{bmatrix}
0.77 & 0.22 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.24 \\
0 & 0 & 0.35 & 0 & 0 & 0 & -0.37 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.96 & -0.27 & -0.24 & 0 & 0 & 0 & 0 \\
0 & 0 & 0.26 & 0 & 0.41 & 0 & 0 & 0 & 0.24 & 0.34 \\
0 & 0 & 0.24 & 0 & -0.24 & 0.69 & 0 & 0 & 0 & 0 \\
0 & 0 & -0.4 & 0 & 0 & 0 & 0.54 & 0 & 0 & 0 \\
0.42 & 0 & 0.21 & 0.27 & -0.8 & 0.59 & -0.22 & 0.81 & -0.32 & 0.49 \\
0 & 0 & 0.22 & 0 & 0 & 0 & 0 & 0 & 0.44 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.21 & 0 & 0 & 0 & 0.89
\end{bmatrix}
$$

(b)

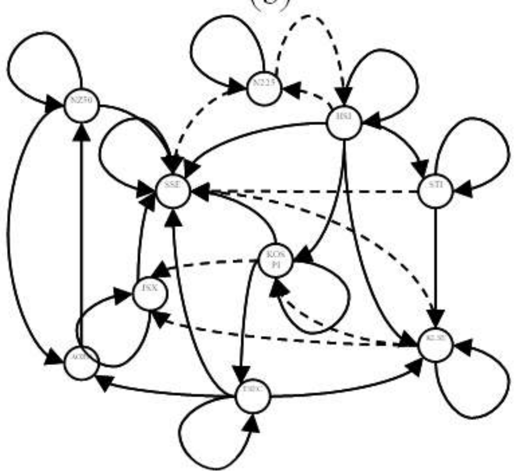

Figure 5: *Results of the DIN modelling process*
(a) transition matrix; *(b)* DIN diagram;

Table 1: *Error rates for prediction with online learning*

[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	0.0143	[illegible]	[illegible]	[illegible]	[illegible]	0.0142

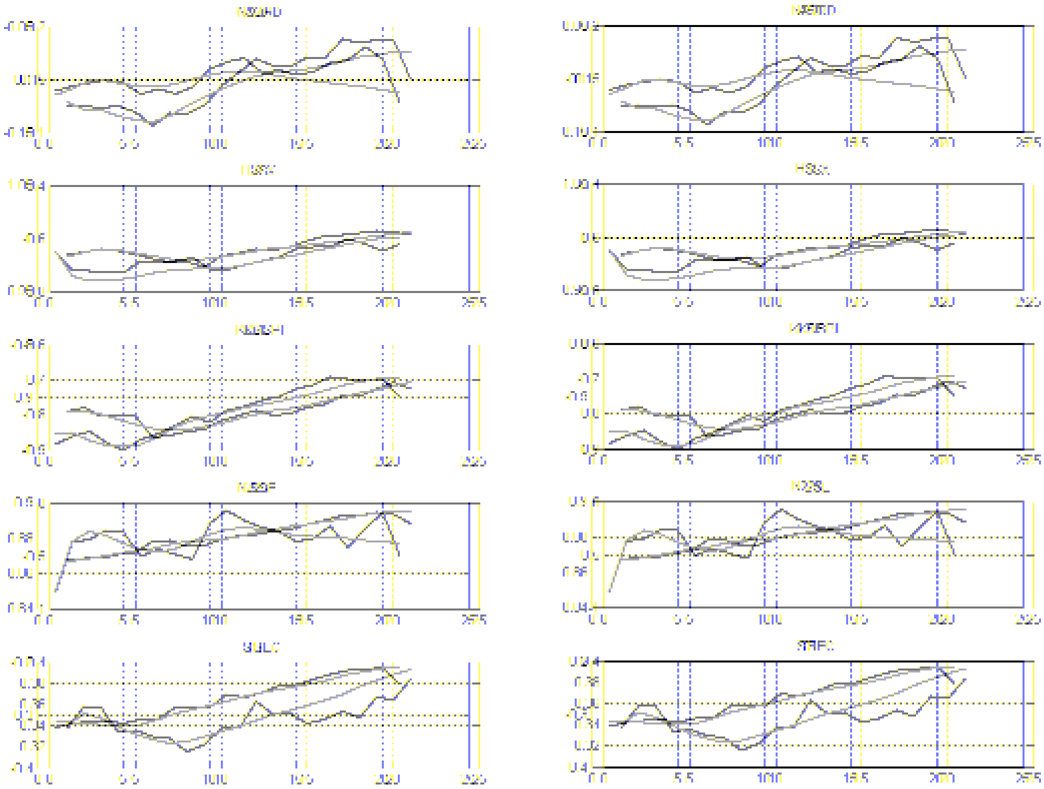

Figure 6: *Comparison between actual (-) and predicted (--) trajectories with online learning over the 20-week test dataset*

3.2. Personalised Modelling with *K*-NN and WKNN

The first implementation of personalised modelling that we use is the basic K-Nearest Neighbour (*K*-NN) to daily data. We apply *K*-NN to data which has been normalized and detrended so that the values lie between -1 and 1. Using *K*-NN we try to predict the value of all stock market indexes at time step *t+1* given the values from previous time steps 1, 2….,t. Results of the application of *K*-NN are shown in Figure 7.

As we can see from the graphic and calculated RMSE (0.079), *K*-NN actually shows good predictive performance. However, a deeper analysis reveals that the nearest neighbours found are actually points which are close to current vector in terms of time. Based on this we can assume that the basic *K*-NN tends to be highly localised in its pattern extraction and does not take into consideration relevant patterns in the more distant past. Past trends are also important when deviations in behaviour occur that cannot be explained by recent events alone. This means that we cannot place absolute confidence in the basic form of the *K*-NN in this domain of stock market prediction.

This motivated us to use the rate of change in the indeed rather, then the value of the index itself. The rate of change has more potential to extract patterns that are outside the immediate time locality. Suppose that we have a state S_i in the system which matches most closely to states S_j, S_k and S_l from the immediate past. However if the rate of change of the current state S_i is very different from S_j, S_k and S_l then the next state from S_i will likely be quite different from each of S_i, S_j or S_k. Having obtained a prediction of the rate of change (hereinafter referred to as velocity) the actual stock index value can be obtained by a combination of the previous value (which is known at time t-1) and the predicted velocity at time t.

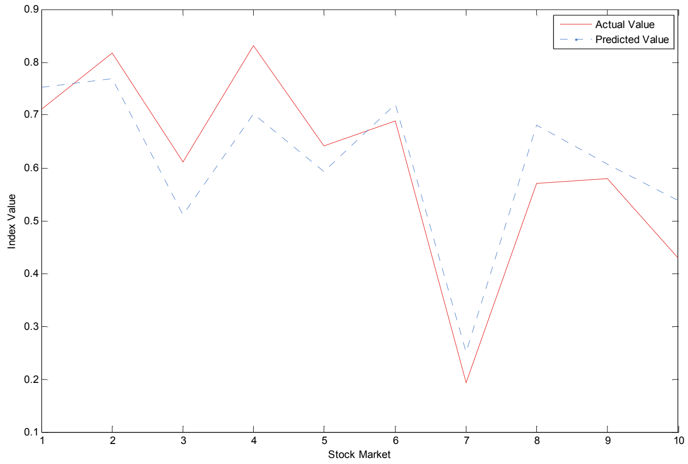

Figure 7: *Prediction of t+1 values of 10 stock market indexes with K-NN; t represents time point 701; K=10*

Table 2: *10 nearest neighbours with the KNN approach based on normalized index value*

Time Point	700	699	698	693	697	694	696	663	660	662
Distance	0.1288	0.3078	0.3203	0.3733	0.3780	0.3783	0.5149	0.5593	0.5660	0.5733

We use 45 features which consist of differences of index values between pairs of stock markets. The results of the velocity prediction process are shown in Figure 8. We can see from the results that the *K*-NN (RMSE=0.1057) performs quite well in predicting the velocity, and we could see that nearest neighbours found are not limited to those points that are close to the current point. Some of the nearest neighbours are points from older data points. Based on this we can assume that patterns of stock movement tend to recur over a period of time. However, it should be noted that the RMSE value is much higher than prediction with the raw index value. This motivated us to investigate the utility of the WKNN approach. We hypothesise that in stock market scenario some neighbours have much more importance then others in the prediction process and the WKNN approach is well suited to weigh the relative importance of each neighbour over the others.

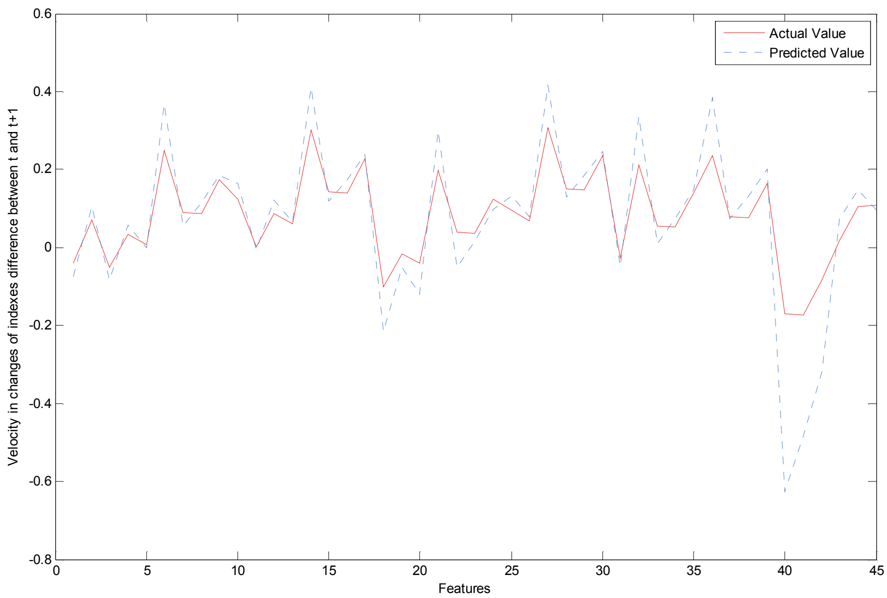

Figure 8: *Predictions based on velocity and KNN*

Table 3: *10 nearest neighbours based on the basic KNN approach applied on velocity*

Time Point	93	654	398	697	700	658	696	291	692	690
Distance	0.1288	0.3078	0.3203	0.3733	0.3780	0.3783	0.5149	0.5593	0.5660	0.5733

Note: *Feature 1 is the velocity of changes in index difference between NZ50 and AORD, feature 2 is the velocity of changes in index difference between NZ50 and HSX, and so on and so forth.*

Results of the WKNN approach can be seen in Figure 9. We can see from Figure 9 and from the RMSE (RMSE=0.0367) value that WKNN performs much better when compared to the basic *K*-NN approach, thus supporting our hypothesis. Thus we can see that personalised modelling can be used not only for prediction of multiple time-series financial data but also could to reveal new knowledge, such as information that interactions between stock markets viewed from the velocity perspective tend to recur over time.

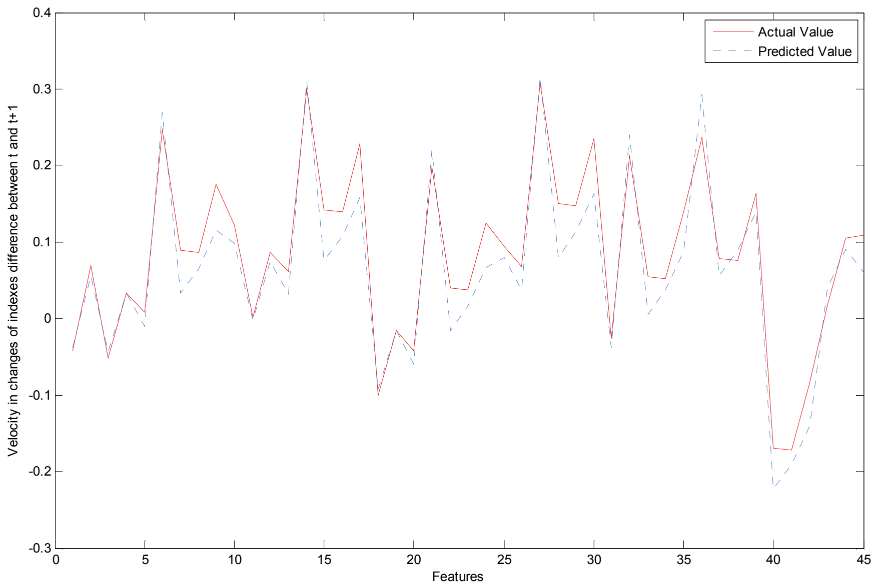

Figure 9: *Predictions based on velocity and WKNN*

Table 4: *10 nearest neighbours based on the WKNN approach*

Time Point	93	654	398	697	700	658	696	291	692	690
Distance	0.2404	0.2449	0.2667	0.3110	0.3272	0.3544	0.3775	0.4239	0.4280	0.4389

4. Conclusion and Further Research

Our study shows that extraction of dynamic interaction networks reveals important and complex interdependencies between stock markets in the Asia Pacific region. These dynamic interaction networks have the potential to predict multiple stock market indexes movement in the long term. Preliminary results from this study using *K*-NN and WKNN show that personalised models predicts stock index values in the shorter term with a reasonable degree of accuracy.

As future work we would like to investigate the effectiveness of the WWKNN approach (Kasabov 2007a) since we believe that interactions between different stock markets have differing degrees of influence, so that more weight should be given to those markets that have a higher degree of interaction with each other. The challenge here is to identify a method that will rank the degree of influence between pairs from the stock markets. We would also like to extend the model to be able to capture global, local and personal trends into one single cohesive model for financial time-series data, and therefore improve prediction. The global model will be used to capture the trend in the whole problem space, the local model will extract local regression or specific behaviour from different sub-spaces of the whole problem space, and finally the personalised model will be used to cope with changes with current data or state of the system. Integrating the three models will provide a better framework for pattern recognition, adaptation and prediction for financial time-series data.

Acknowledgements

We would like to thank Dr Antoaneta Serguieva from Centre for Empirical Finance, Brunel University, and Brunel Business School for her contributions in our previous work of building a global model to extract dynamic interaction between stock markets in Asia Pacific.

References

[1] A. Antoniou, G. Pescetto, and A. Violaris, Modelling international price relationships and interdependencies between the stock index and stock index future markets of three EU countries: A Multivariate Analysis, *Journal of Business Finance and Accounting* 30: 645-667, 2003.

[2] G.M. Caporale, A. Serguieva, H. Wu, A mixed-game agent-based model for simulating financial contagion. In *Proceedings of the 2008 Congress on Evolutionary Computation*, 3420-3425, IEEE Press, 2008.

[3] Z. Chan, N. Kasabov, and L. Collins, A two-stage methodology for gene regulatory network extraction from time-course gene expression data, *Expert System with Applications* 30: 59-63, 2006.

[4] T.C. Chiang, S. Doong, Empirical analysis of stock returns and volatility: Evidence from seven Asian stock markets based on TAR-GARCH model, *Review of Quantitative Finance and Accounting* 17: 301–318, 2001.

[5] D. Collins, N. Biekpe, Contagion and interdependence in African stock markets, *The South African Journal of Economics*, 71(1): 181-194, 2003.

[6] P. D'haeseleer, S. Liang, and R. Somogyi, Gene expression data analysis and modelling. In *Proceedings of the Pacific Symposium on Biocomputing*, Hawaii, 1999.

[7] S. Goldfeld, R. Quandt, A Markov model for switching regressions, *Journal of Econometrics* 1(1): 3-16, 1973.

[8] N. Kasabov, *Evolving connectionist systems: Methods and applications in bioinformatics, Brain Study and Intelligent Machines*, Springer, 2002.

[9] N. Kasabov, Z. Chan, V. Jain, I. Sidorov, D. Dimitrov, Gene regulatory network discovery from time-series gene expression data: A computational intelligence approach Presented at the: International Conference on Neural Information Processing, Calcutta, 2003. Published in *Lecture notes in computer science* 3316: 1344-1353, 2004.

[10] N. Kasabov, Global, local and personalised modelling and pattern discovery in bioinformatics: An integrated approach, *Pattern Recognition Letters* 28: 673-685, 2007a.

[11] N. Kasabov, *Evolving Connectionist Systems: The Knowledge Engineering Approach*, Springer, 2007b.

[12] A. Masih, R. Masih, Dynamic modelling of stock market interdependencies: An empirical investigation of Australia and the Asian NICs, *Working Papers* 98-18: 1323-9244, University of Western Australia, 1998.

[13] G. Welch, G. Bishop, An Introduction to the Kalman Filter, *Computer Science Working Papers* TR95-041, University of North Carolina at Chapel Hill, 2006.

[14] A. Serguieva, T. Kalganova, T. Khan, An intelligent system for risk classification of stock investment projects, *Journal of Applied Systems Studies* 4 (2): 236-261, 2003.

[15] A. Serguieva, T. Khan: Knowledge representation in risk analysis, *Business and Management Working Papers* March'04: 1-21, Brunel University, 2004.

[16] A Serguieva, H. Wu, Computational intelligence in financial contagion analysis. Presented at the *Seventh International Conference on Complex Systems*, Boston, Massachusetts, 2007. Published in *InterJournal on Complex Systems* 2229: 1-12, 2008.

[17] Q. Song, N. Kasabov, Dynamic evolving neuro-fuzzy inference system (DENFIS): On-line learning and application for time-series prediction, *IEEE Transactions of Fuzzy Systems* 10: 144-154, 2002.

[18] V.N. Vapnik, Statistical Learning Theory, Wiley Inter-Science, 1998.