International Journal of u- and e- Service, Science and Technology
Vol.9, No. 12 (2016), pp.317-326
http://dx.doi.org/10.14257/ijunesst.2016.9.12.28

Using ARIMA Model to Fit and Predict Index of Stock Price Based on Wavelet De-Noising

Shihua Luo[1], Fang Yan[1*], Dejian Lai[2], Wenyi Wu[1] and Fucai Lu[1]

[1]*School of Statistics, Center of Applied Statistics, Jiangxi University of Finance & Economics, Nanchang, 330013, China;*
[2]*The University of Texas School of Public Health, Houston, TX, 77030, USA*
[*]*helen516@yeah.net*

Abstract

To accommodate non-stationarity and strong noise in the SPI data, the research used wavelet method for de-noising and autoregressive integrated moving average model (ARIMA) for prediction. Seven-day moving averages of closing time SPI data in four Asian stock markets were analyzed. Empirical results show that after de-noising more accurate forecasting results can be obtained in developed markets. More developed market indexes seem more significant improvement; while for less developed market indexes, the improvement of de-noising is less significant. This is in accordance with current situation of market.

Keywords: *stock price index (SPI); ARIMA model; wavelet de-noising; high-frequency signals*

1. Introduction

As the main part of financial markets, stock market plays an important role in national economics. The stock price index (SPI) reflects the overall price level and fluctuations of various stock markets. A SPI is a method of measuring the value of a section or overall of the stock market. It is computed from the prices of selected stocks. Form the macroscopic aspect, analyzing SPI may provide valuable basis for the government's decision-making; while from the microscopic aspect, it may improve investors' investment strategies. During the past several decades, great research effort has been devoted to analysis and prediction of SPI. For example, Saad *et. al.*, exploited time delay, recurrent and probabilistic neural networks to predict stock trend [1]; Watanabe *et. al.*, used neural network and artificial neural network to predict trend of stock market [2].

Despite these advances, analysis and prediction of SPI is not a trivial task. The SPI data are commonly regarded as non-stationary and heteroskedastic processes with strong noise. A good prediction model for SPI should be able to deal with these characteristics. Recently, wavelet analysis has emerged as a popular de-noising method for time series prediction [3-6]. Wavelet method is able to remove low amplitude noise or undesired signal in wavelet domain and desired signal can be retrieved from an inverse wavelet transform with little loss of main details.

Considering these advantages, wavelet method was used to de-noise seven-day moving average data of the SPI to facilitate subsequent
construction of prediction model in this article.

With the de-noised signals obtained, time series model can be used for modeling and prediction of the SPI data. Among different time series models, ARIMA is one of the most popular. ARIMA model is a generalization of the autoregressive moving average (ARMA)

[*]Corresponding Author

ISSN: 2005-4246 IJUNESST

model, which is widely used in time series prediction due to its capability to dealing with the non-stationarity. For example, Conejo *et. al.*, used ARIMA to forecast day-ahead electricity prices [7]; Hou *et. al.*, integrated ARIMA model with wavelet method to predict wind power generation [8].

In this paper, The SPI data of four Asian stock markets, *i.e*, Hong Kong Hang Seng Index (HSI), Taiwan Weighted Index (TAIEX), Shanghai Composite Index (SSE) and Shenzhen Component Index (SZSE) are analyzed and predicted by using the ARIMA model. The HSI was created by Hong Kong banker Stanley Kwan and debuted on November 24, 1969 and the TAIEX was first published in 1967 by Taiwan Stock Exchange. The SSE was calculated using a Paasche weighted composite price index formula and launched on July 15, 1991. The SZSE is an index of 40 stocks that are traded at the Shenzhen Stock Exchange and was launched on May 5, 1995.

First, we use wavelet multi-resolution analysis to conduct wavelet fast decomposition for the SPI time series, and obtain trend coefficient sequences and detail coefficient sequences. After that, we reconstruct the de-noising SPI time series by soft-thresholding at each level. With the de-noised signal obtained, we build ARIMA model to fit it. Our results show that, for more developed market indexes such as HSI and TAIEX, noise can be removed effectively using wavelet method. Combination of ARIMA model with wavelet de-noising has better prediction results than using ARIMA model alone.

2. Wavelet de-Noising Method

Wavelet de-noising method used in this article is based on Mallat algorithm. The Mallat algorithm of binary orthogonal wavelet transform is a rapid algorithm for multi-scale analysis of discrete time series, which has a wide range of applications in many engineering fields [9-12]. Mallat algorithm is consisted of two steps, *i.e*, decomposition and reconstruction.

The decomposed formula for Mallat algorithm is [10]:

$$c_{j,k} = \sum_m h(m-2k)c_{j-1,m} \tag{1}$$

$$d_{j,k} = \sum_m g(m-2k)c_{j-1,m} \tag{2}$$

The reconstruction formula for Mallat algorithm is:

$$c_{j-1,m} = \tilde{c}_{j-1,m} + \tilde{d}_{j-1,m} \tag{3}$$

$$\begin{cases} \tilde{c}_{j-1,m} = \sum_k c_{j,k} h(m-2k) \\ \tilde{d}_{j-1,m} = \sum_k d_{j,k} g(m-2k) \end{cases} \tag{4}$$

Here $\tilde{c}_{j-1,m}, \tilde{d}_{j-1,m}$ are reconstructed coefficients and $c_{j,k}$, $d_{j,k}$ are decomposition coefficients of corresponding wavelet, $m = 2k+n, k, n \in Z$; *h(n)*, *g(n)* are low-pass filter and high-pass impulse response of corresponding wavelet respectively.

c_0 is defined as original signal *X*; by (1) and (2), *X* can be decomposed into $d_1, d_2, \ldots, d_J, c_J$ (where *J* is maximum decomposition level); c_j and d_j are respectively referred to trend signals and detail fluctuation signal in resolution 2^{-j} of the original signal. Using Mallat algorithm for wavelet decomposition, detail signal and approximation signal reduce into a half after decomposition, so that time resolution is reduced, while frequency resolution is doubled.

The Mallat algorithm works as follows [13-14]:

(1) Selection of wavelet function and level of wavelet decomposition

Selecting appropriate wavelet function and determining level of wavelet decomposition

J, then disintegrating J layer wavelet decomposition of the original signal.

(2) Determination of the threshold for high-frequency coefficients

Processing each layer of high-frequency coefficients di (i=1, 2, …, J) with soft-thresholding, then the high frequency components after de-noising can be obtained.

(3) Reconstruction of original signal

Choosing the J-th layer low-frequency coefficients of wavelet decomposition and the first to J-th layer high-frequency coefficients after quantization process to reconstruct signal, which is based on the procedure shown in express (3) and (4). Then the true signal after de-noising can be obtained.

Mallat algorithm is implemented by Matlab7.6 [15].

1. Autoregressive Integrated Moving Average Model（ARIMA）

As generalization of the stationary ARMA model, ARIMA can model non-stationary time series. ARIMA model uses a difference operation first to eliminate non-stationarity. After that, traditional ARMA model is used to fit the differential stationary sequence.

Time series $\{x_t : t = 1, 2, \cdots, N\}$ of ARIMA(p, d, q) model is defined as [16-20]:

$$(1-\phi_1 B-\cdots-\phi_p B^p)\nabla^d x_t = (1-\theta_1 B-\cdots-\theta_q B^q)\varepsilon_t \tag{5}$$

here B is a delay operator, $\nabla^d = (1-B)^d$. $\{\varepsilon_t\}$ is a zero mean white noise sequence.

If $\varphi(B) = \dfrac{(1-\theta_1 B-\cdots-\theta_q B^q)}{(1-\phi_1 B-\cdots-\phi_p B^p)\nabla^d}$, equation (5) can be written as:

$$x_t = \varphi(B)\varepsilon_t \tag{6}$$

Obviously, if d=0, ARIMA(p, d, q) model is reduced to ARMA(p, q) model; if d=1,p=q=0, ARIMA(0, 1, 0) model would become:

$$\begin{cases} x_t = x_{t-1} + \varepsilon_t \\ \varepsilon_t \sim N(0, \sigma_t^{\,2}) \end{cases} \tag{7}$$

Express (7) describes the random walk model.

For a given sequence $\{x_t : t = 1, 2, \cdots, N\}$, construction of ARIMA model involves difference operation, stability test, autocorrelation test, partial autocorrelation test and fitting of the ARMA(p, q) model. With the fitted model, k-th step prediction of future value can be obtained as follows:

$$\hat{x}_t(k) = \psi_k \varepsilon_t + \psi_{k+1}\varepsilon_{t-1} + \psi_{k+2}\varepsilon_{t-2} + \cdots \tag{8}$$

where

$$\begin{cases} \psi_1 = \phi_1 - \theta_1 \\ \psi_2 = \phi_1\psi_1 + \phi_2 - \theta_2 \\ \cdots\cdots \\ \psi_j = \phi_1\psi_{j-1} + \cdots + \phi_{p+d}\psi_{j-p-d} - \theta_j \end{cases}, \ \psi_j = \begin{cases} 0, & j < 0 \\ 1, & j = 0 \end{cases}, \ \theta_j = 0 \ (j > q) \tag{9}$$

Accuracy of the model is often evaluated by relative error, which is defined as follows:

$$P\text{err} = \sum_{n=1}^{L} [x'(n) - x(n)]^2 / \sum_{n=1}^{L} x^2(n) \tag{10}$$

Here *x'(n)* is predictive value, *x(n)* is observed value and *L* is predicted length [21].

4. Predictive Algorithm Framework

First step is to de-noise the SPI data by wavelet method. After selecting appropriate mother wavelet, J-times Mallat wavelet fast decomposition can be obtained. Based on maximum-minimum method for threshold, high frequency signal component is then filtered and de-noised time series reconstructed by Mallat reconstruction. With the de-noised signal, an ARIMA(p, d, q) model is built for prediction. The specific algorithm framework is shown in Figure 1:

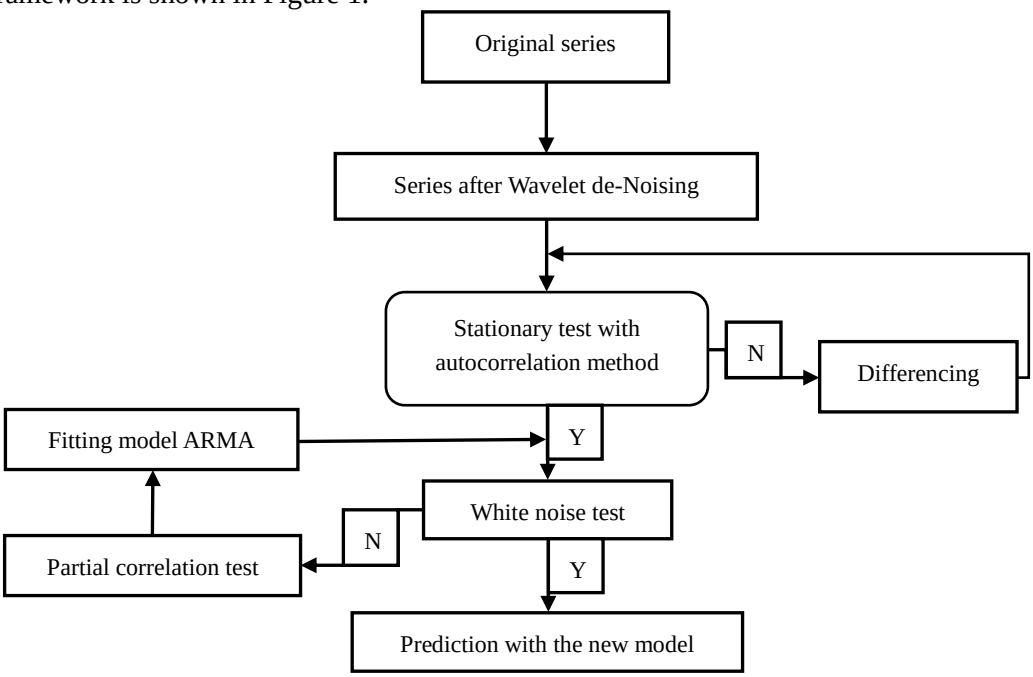


Figure 1. Framework of SPI Prediction

5. Modeling of Four Asian Stock Markets

The paper extracts the seven-day moving average of closing price sequence from July 1, 2013 to June 30, 2014 for four markets as sample space. As more distant historical factors usually affect time series forecasting smaller, too much data may be expand prediction error and the purpose is just to have a short-termpredictions. The moving average data are shown in Figure 2.

For the purpose of de-noising, we perform double-time Mallat wavelet fast decomposition on the SPI data, and select wavelet basis as Db3 wavelet. Figure 3, presents the time series decomposition of HSI moving average closing price (There are the similar processes about other three market samples, the paper does not show these details by space limited in this paper):

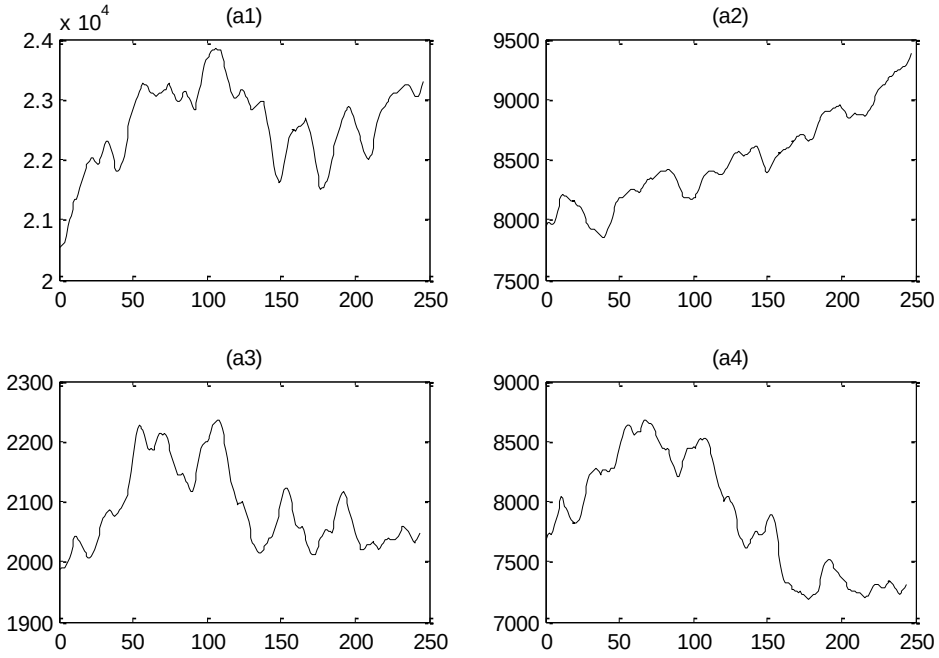


Figure 2. The Original Time Series' Seven-day Moving Average of Four Markets, a1-a4 Represent the HSI, TAIEX, SSE and SZSE Respectively

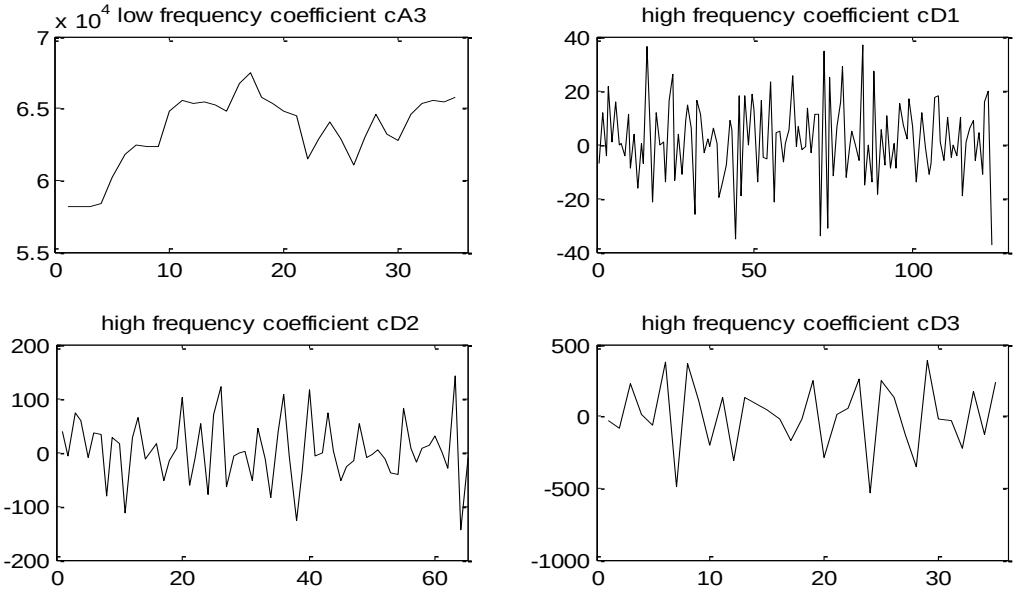


Figure 3. Time Series Decomposition of HSI Moving Average Closing Price

To investigate the characteristics of de-noised signal using Matlab, we inspect the autocorrelogram of HSI as presented in Figure 4.

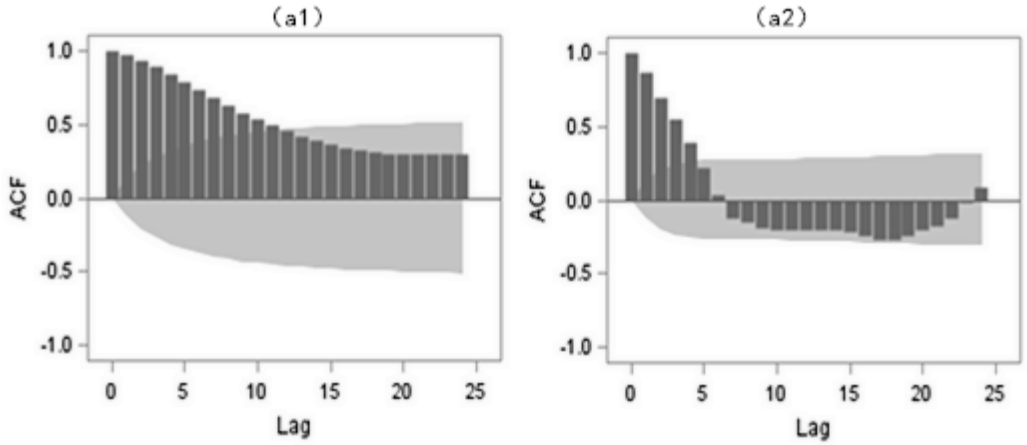


Figure 4. Autocorrelogram of HSI (a1) and Autocorrelogram of HSI First-Order Difference (a2)

Figure 4. (a1), shows that the SPI of HSI is non-stationary. To obtain stationary time series, we perform first order differencing operation of the HSI sample. Autocorrelogram of the HSI data after differencing operation is shown in (a2). (a2) shows the autocorrelation coefficient of differenced sequence has minor fluctuations within confidence band after lag5; so it can be considered as stable. Further test of unit root on the differential serial are shown in Table 1.

Table 1. Augmented Dickey-Fuller Unit Root Tests of HSI Sample First-order Difference Sequence

Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	-29.2182	<.0001	-3.88	0.0001		
	1	-44.9890	<.0001	-4.67	<.0001		
Single Mean	0	-29.9396	0.0015	-3.94	0.0022	7.75	0.0010
	1	-46.1595	0.0015	-4.73	0.0002	11.20	0.0010
Trend	0	-30.1602	0.0066	-3.92	0.0127	7.74	0.0151
	1	-46.8342	0.0006	-4.73	0.0008	11.24	0.0010

Table 1, shows that hypothesis of unit root is rejected. Hence we can conclude that (confidence level more than 99.999%) the first-order difference sequence of HSI sample is stationary. Further the paper conducts tests of white noise, and results are presented in Table 2:

Table 2. Autocorrelation Check for White Noise of the HSI Sample

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	446.82	6	<.0001	0.873	0.712	0.562	0.398	0.223	0.035
12	498.73	12	<.0001	-0.114	-0.157	-0.191	-0.209	-0.202	-0.207

18	587.83	18	<.0001	-0.211	-0.212	-0.217	-0.241	-0.266	-0.268

From Table 2, χ^2 statistics show that hypothesis of white noise is rejected as the p values are small (less than 0.0001). The hypothesis test shows that we can have a great confidence (confidence level more than 99.999%) that the sequence is not white noise.

In all of the model ARIMA(p, 1, q) ($p \le 5, q \le 5$), combined optimal fitting results, BIC(2,5) = 7.103277 is minimum with least squares estimation and AIC is minimum, the final model can fit of ARIMA(2, 1, (2,5)) model:

$$\left(1 - 1.01997B + 0.23547B^2\right)(1-B)x_t = \left(1 - 0.19001\,B^2 + 0.3297B^5\right)\varepsilon_t \quad (11)$$

To compare the effect of two algorithms, the original data are fitted with ARIMA(2, 1, (2,3,4,5)) model:

$$\left(1 - 0.85266B + 0.39095B^2\right)(1-B)x_t$$

$$= \left(1 + 0.52315\,B^2 + 0.4086B^3 + 0.41106B^4 + 0.53116B^5\right)\varepsilon_t \quad (12)$$

To ensure the model can adapt to online evolution of index, the top of index value would be wipe off and latest real data would be added in after each prediction, then re-fitting model and forecasting the next day. In this way, the model can keep up with inherent variation of the system. Table 3, shows forecasting simulation results of seven-day moving average data for Hong Kong stock market. (July 2, 2014 - July 22, 2014).

Table 3. The Forecasting Simulation Results of HSI

Date	Original Data	Original prediction	De-noising prediction
2014-07-02	23396.92	23377.63	23380.90
2014-07-03	23445.99	23458.89	23464.08
2014-07-04	23439.50	23488.25	23487.05
2014-07-07	23446.40	23428.02	23420.90
2014-07-08	23401.23	23429.93	23433.09
2014-07-09	23374.83	23357.15	23360.97
2014-07-10	23362.49	23346.44	23345.74
2014-07-11	23359.97	23338.95	23344.17
2014-07-14	23357.04	23381.27	23384.26
2014-07-15	23396.86	23373.85	23377.16
2014-07-16	23418.02	23435.91	23433.92
2014-07-17	23496.40	23442.42	23443.57
2014-07-18	23585.72	23555.30	23554.44
2014-07-21	23683.08	23651.81	23657.07
2014-07-22	23782.04	23773.88	23775.33

For other three markets by same processes, the paper established the following models ARIMA(2, 1, 0), ARIMA(2, 1, (3,4,5)), ARIMA(2, 1, (2,3,4,5)) for TAIEX, SSE and SZSE respectively :

$$\left(1 - 1.14166B + 0.29692B^2\right)(1-B)x_t = 10.22687 + \varepsilon_t \quad (13)$$

$$\left(1 - 1.11333\mathrm{B} + 0.38359\mathrm{B}^2\right)(1\text{-}\mathrm{B})\,x_t = \left(1 + 0.35895\mathrm{B}^3 + 0.31717\mathrm{B}^4 + 0.13424\mathrm{B}^5\right)\varepsilon_t \quad (14)$$

$$\left(1 - 0.82024\mathrm{B} + 0.47973\mathrm{B}^2\right)(1\text{-}\mathrm{B})\,x_t = \left(1 + 0.63687\mathrm{B}^2 + 0.34926\mathrm{B}^3 + 0.414\mathrm{B}^4 + 0.54244\mathrm{B}^5\right)\varepsilon_t \quad (15)$$

The SPI of the next fifteen days forecast as well. In order to evaluate predictive model reasonably, we use relative error as criterion, which have mentioned in equation (10).

Table 4. The Comparison of Relative Error of the Four Stock Market Samples

Perr	HSI	TAIEX	SSE	SZSE
Original prediction	1.382×10^{-4}	1.324×10^{-4}	$\underline{1.260\times10^{-4}}$	$\underline{3.332\times10^{-4}}$
De-noising prediction	$\underline{1.345\times10^{-4}}$	$\underline{1.269\times10^{-4}}$	2.261×10^{-4}	1.630×10^{-3}

Table 4, shows: (1) Good results (most of the Perrs are as low as 10^{-4} order of magnitude) get by using the prediction model for SPI, and current predictive model has still enough competitive power compared with other more complex models. Main dominance is that only SPI historical data are utilized for building predictive model in this study. (2) For HSI and TAIEX, the accuracy of prediction of de-noising prediction model is higher than the original prediction. However, for SSE and SZSE, the accuracy of prediction of de-noising prediction model is even lower than the original prediction.

It means that high-frequency signals which including in more developed market indexes such as HSI and TAIEX are easier to distinguish from the noise than including in developing market indexes such as SSE and SZSE. When the time series of more developed market index are de-noised, the valuable high-frequency signals will be preserved and generate a positive role in subsequent predictive modeling process. On the other hand, for developing markets, many investors are inexperienced and their behaviors are influenced by many sporadic and random factors. The high-frequency signals which included in such less developed market and noise are usually closely mixed and can be removed together in the process of de-noising. So after de-noised, the hit ratio of prediction may be even lower than the original prediction by losing the valuable high-frequency signals. The simulation results in Table 4 reveal such inherent law.

6. Conclusion

This paper used a combination of wavelet de-noising and ARIMA models to model and to predict the seven-day moving average of closing price data of four stock markets from mainland China, Hong Kong and Taiwan. Simulation results show that: (1) Most of the Perrs are as low as 10^{-4} order of magnitude that only SPI historical data are utilized for building predictive model; (2) The more developed markets, the better such new model is used. The fitted and predicted details show that developed market (HSI&TAIEX) have better regulations, and the market participants have higher level of operation and education, the short-term trading behaviors of investors are more rational. Above all, noise can be separated from high frequency signal more easily in more developed market, and valuable high-frequency signals are preserved and generate a positive role in the subsequent predictive modeling process, relative to the developing markets (SSE&SZSE). More developed markets have better predictability as they have better regulations and more experienced investors; for less developed markets, many investors are inexperienced and their behaviors are more dramatically influenced by many factors. The research is helpful for investors to judge the situation of the stock market and make corresponding

measures correctly.

Acknowledgements

This research is partially supported by the NSFC (61263014, 61563018, 71463020), the China Postdoctoral SF (2013M531553, 2013KY15), the NSF of Jiangxi Province 20132BAB201011, 20133BCB23014) and the Foundation of the Office of Education, Jiangxi Province (KJLD13033).

References

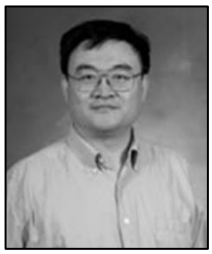
[1] E. M. Saad, D. V. Prokhorov and D. C. Wunch, "Comparative study of stock trend prediction using time delay, recurrent and probabilistic neural networks", IEEE Transactions on Neural Networks, vol. 9, no. 6, **(1998)**, pp. 1456 -1470.
[2] H. Watanabe, B. Chakraborty and G. Chakrabortyj, "Soft computing approach for stock price trend forecasting from multivariate time series", Proceedings of 2nd International Conference on Innovative Computing, Information and Control(ICICIC), Kumamoto, Japan, **(2007)** September 5-7.
[3] B. Alberto, B. Francesco and M. Manfred, "Model predictive control based on linear programming-the explicit solution" , IEEE Transaction on Automatic Control, vol. 47, no. 12, **(2002)**, pp. 1974-1985.
[4] J. S. Zeng, C. H. Gao, X. G. Liu and K. P. Yang, "Using non-linear GARCH model to predict silicon content in blast furnace hot metal", Asian Journal of Control, vol. 10, no. 6, **(2009)**, pp. 632-637.
[5] W. L. Qin, H. S. Yan and Q. L. Da, "Analysis of SH&SZ stock market dependence based on multi-resolution recognition", Application of Statistics and Management, vol. 28, no. 3, **(2009)**, pp. 517-522.
[6] J. L. Zhao, "Stock market analysis based on the wavelets", Journal of Taiyuan Normal University, vol. 6, no. 2, **(2007)**, pp. 5-7.
[7] A. J. Conejo, M. A. Plazas, R. Espinola and A. B. Molina, "Day-ahead electricity price forecasting sing the wavelet transform and ARIMA models", IEEE Transaction on Power Systems, vol. 20, no. 2, **(2005)**, pp. 1035-1042.
[8] Z. S. Hou, Y. V. Makarov, N. A. Samaan and P. V. Etingov,"Standardized software for applications at multiple geographically distributed wind farms", Proceedings of 46th Hawaii International Conference on System Sciences (HICSS), Hawaii, America, **(2013)**, pp. 5005-5011.
[9] J. Tang, "Financial time series research and empirical study based on wavelet analysis", University of Science and Technology of China, Management of Science and Engineering, Hefei, **(2011)**.
[10] D. L. Donoho, "De-noising by soft-thresholding", IEEE Trans on Information Theory, vol. 41, no. 3, **(1995)**, pp. 613-627.
[11] D. F. Zhang, "Wavelet de-noising with MATLAB", China Machine Press, Beijing, **(2009)**.
[12] X. J. Wang, J. J. Lian and S. M. Fei, "Chaos identification of hydrologic system based on wavelet de-noising", System Engineering Theory and Practice, vol. 6, no. 3, **(2008)**, pp. 220-222.
[13] S. H. Luo, J. S. Zeng, "Multi-fractal identification of the fluctuation of silicon content in blast furnace hot metal based on multi-resolution analysis", Acta Physica Sinica, vol. 58, no. 1, **(2009)**, pp. 150-157.
[14] A. Shabri and R. Samsudin, "Daily Crude Oil Price Forecasting Using Hybridizing Wavelet and Artificial Neural Network Model", Mathematical Problems in Engineering, **(2014)**, pp. 1-10.
[15] Z. X. Ge and W. Sa, "Wavelet analysis theory and MATLABR2007 to achieve", Publishing House of Electronics Industry, Beijing, **(2007)**.
[16] C. Javier, E. Rosarion, J. N. Francisco and J. C. Antonio, "ARIMA models to predict next-day electricity prices", IEEE Transaction on Power systems, vol. 18, no. 3, **(2003)**, pp. 1014-1020.
[17] D. J. Lai, "Monitoring the SARS Epidemic in China: A Time Series Analysis", Journal of Data Science, vol. 3, no. 3, **(2005)**, pp. 279-293,.
[18] Y. Wang, "Application of time series analysis", Chinese Renmin University publishing company, Beijing, **(2011)**.
[19] G. S. Zhao, "Research of stock price trend forecast based on time series Analysis", Xiamen University, Xiamen, **(2009)**.
[20] A. A. Adebiyi, A. O. Adewumi and C. K. Ayo, "Comparison of ARIMA and Artificial Neural Networks Models for Stock Price Prediction", Journal of Applied Mathematics, vol. 33, no.1, **(2014)**, pp. 75-81.
[21] J. S. Zeng, C. H. Gao, X.G. Liu, K. P. Yang and S. H. Luo, "Using non-linear GARCH model to predict silicon content in blast furnace hot metal", Asian Journal of Control, vol. 10, no. 6, **(2008)**, pp.632-637.

Authors

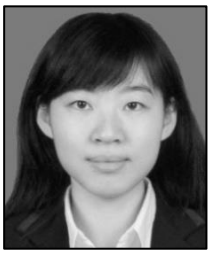
Shihua Luo, received the Ph.D. degree in Operational Research and Control Theory from Zhejiang University, China, in 2007. He is currently a professor of Jiangxi University of Finance and

Economics. His research interests are in the field of modeling and optimization of complex systems, and multivariate analysis based on data-driven.

Fang Yan, received the B.S. degree in Statistics from Jiangxi University of Finance and Economics, China, in 2013. Her major is mathematical statistics in graduate school, and her research interest is multi-scale time series analysis.



Dejian Lai, Ph.D., is professor of statistics at The University of Texas School of Public Health. Dr. Lai received his Ph.D. in statistics from The University of Texas at Dallas. His research areas include time series analysis and statistical analysis of clinical trials.



Wenyi Wu, received the B.S. degree in Economics and Management from Jiangxi University of Finance and Economics, China, in 2013. From then on to this day, she is a postgraduate student. Her major is mathematical statistics, and her research interest is multi-scale time series analysis.

Fucai Lu, received the Ph.D. degree in Industrial Economies from Graduate School of Chinese Academy of Social Sciences in 2000. He is currently vice head of the Jiangxi Provincial Department of Science and Technology, professor of Jiangxi University of Finance and Economics. His research interests are in the field of Industrial organization theory, Ecological Economics, and human resources development.