A Novel Model for Stock Portfolio Based on ARX, RS and a New Grey Relational Grade Theories

Kuang Yu Huang
Department of Information Management
Ling Tung University
#1 Ling Tung Road, Taichung City 408, Taiwan
kyhuang@mail.ltu.edu.tw

Chuen-Jiuan Jane
Department of Finance
Ling Tung University
#1 Ling Tung Road, Taichung City 408, Taiwan
ltc869@mail.ltu.edu.tw

***Abstract*—In this study, the new grey relational grade (GRG) method is combined with moving Average Autoregressive Exogenous (ARX) prediction model, GM(1,N) theory and Rough Set (RS) theory to create an automatic stock market forecasting and portfolio selection mechanism. In the proposed approach, financial data are collected automatically every quarter and are input to an ARX prediction model to forecast the future trends of the collected data over the next quarter or half-year period. The forecast data is then reduced using a GM(1,N) model, clustered using a K-means clustering algorithm and then supplied to a RS classification module which selects appropriate investment stocks by applying a set of decision-making rules. Finally, a new grey relational analysis technique is employed to specify an appropriate weighting of the selected stocks such that the portfolio's rate of return is maximized. The validity of the proposed approach is demonstrated using electronic stock data extracted from the financial database maintained by the Taiwan Economic Journal (TEJ). It is found that the proposed method yields an average annual rate of return, 25.91%, on the selected stocks from 2004 to 2006 in Taiwan stock market.**

***Keywords*—component, formatting, style, styling, forecasting, ARX model, rough set, grey relational analysis, stock portfolio**

I. INTRODUCTION

A number of applications have been proposed in recent decades for predicting market trends. Typical mechanisms include the use of genetic algorithms to choose optimal portfolios [1, 2], the application of neural networks to predict real-world stock trends [3-5], the integration of fuzzy logic and forecasting techniques to create artificial intelligence systems for market tracking and forecasting purposes [6-7], the use of statistical approaches for the forecasting of economic indicators [8-12], the application of rough set (RS) theory to predict the S&P 100 index [13], and so on.

Rough set theory was introduced more than twenty years ago [14] and has emerged as a powerful technique for the automatic classification of objects [15]. Most RS applications are designed to deal with classification problems of one form or another. In constructing such applications, RS theory is generally integrated with other theories such as Grey Systems theory, for example. Typical applications include the self-learning system LERS [16-17], the classification scheme [18], the stock market analysis algorithm [19-21]. In these applications, the stock market is essentially a dynamic process. In general, such processes can be described using a discrete-time series model, such as an autoregressive (AR) model, a moving average (MA) model, or an auto-regressive with exogenous inputs (ARX) model [22-25].

The grey relational analysis theory initiated by Deng [26], which was applied most widely, makes use of grey relational generating and calculates the grey relational coefficient to handle the uncertain systematic problem under the status of only partially known information.

The proposed approach in this study combines the ARX prediction model, the multivariate GM(1,N) model [26], the K-means clustering technique, RS theory, a new grey relational analysis, and the investment guidelines prescribed by Buffett [27] to develop an algorithm for forecasting financial data over a quarter or half-year period and for predicting the stock portfolio which will maximize the rate of return.

The remainder of this paper is organized as follows. Section 2 presents the fundamental principles of the ARX model, RS theory and Grey theory, respectively. Section 3 describes the integration of these concepts to construct an automatic forecasting and portfolio selection scheme. Section 4 compares the performance of the proposed method with that of a conventional GM(1,1) forecasting model. Finally, some brief conclusions and future researches are presented.

II. REVIEW OF RELATED METHODOLOGIES

A. ARX model

In general, the objective of ARX prediction models is to minimize a positive function of the prediction errors. When constructing the ARX model, these parameters are generally estimated using a system identification process. The ARX model has the form

$$\begin{aligned} & y(t)+a_1 y(t-1)+\ldots+a_{n_a} y(t-n_a) \\ & =b_1 u(t-1)+\ldots+b_{n_b} u(t-n_b)+e(t) \end{aligned} \tag{1}$$

where $y(t)$ is the output, $u(t)$ is the input and $e(t)$ is a white-noise term.

In the application considered in the present study, i.e. the forecasting of time-series financial data, the most important aspect of this ARX model is its one-step-ahead predictor, which has the form $\hat{y}(t,\theta)=\theta^T\varphi(t)$, where

$\theta = \left[a_1,...a_{n_a}, b_1,...,b_{n_b}\right]^T$ is obtained using the least squares method and yields $\min_{\theta} \frac{1}{N}\sum_{t=1}^{N}\left(y(t)-\hat{y}(t,\theta)\right)^2$, and

$$\varphi^T(t) = \left[-y(t-1),...,-y(t-n_a),u(t-1),...,u(t-n_b)\right]. \quad (2)$$

As shown, the one-step-ahead predictor is essentially a scalar product of a known data vector $\varphi(t)$ and the parameter vector θ. In statistics, such a model is known as a linear regression model and vector $\varphi(t)$ is the regression vector.

In the ARX model, the prediction error is given by $e(t) = y(t) - \hat{y}(t,\theta)$ and can be computed given a knowledge of $y(t)$ and $\varphi(t)$. For convenience, each instance of $\left(y(t),\varphi(t)\right)$ is generally referred to as a data point.

The corresponding prediction error is given by $e(t) = y(t) - \hat{y}(t,\theta)$. It can be computed given by $y(t)$ and $\varphi(t)$, and we will refer to $\left(y(t),\varphi(t)\right)$ as a data point.

B. Rough set theory

Rough set theory (RST) was introduced by Pawlak [14] in 1982. RST is a powerful mathematical tool for handling the vagueness and uncertainty inherent in many decision-making processes. The underlying principle of RST is the assumption that every object in the universe of discourse has a set of information (i.e. attributes) associated with it. Objects characterized by the same information are regarded as indiscernible. The indiscernibility relationships generated amongst all the objects in the universe of discourse provide the basic mathematical basis for RST. Typical problems amenable to RST processing include classifying sets of objects in terms of their attribute values, checking dependencies (full or partial) between attributes, reducing attributes, analyzing the significance of individual attributes, generating decision rules, and so on.

1) Information systems

In RST, knowledge about the universe of discourse is represented using so-called information systems. A typical information system has the form $S = (U, \Omega, V_q, f_q)$, where U is a non-empty set of finite objects and Ω is a non-empty finite set of attributes describing each object. Here, $\Omega = C \cup D$, in which C is a finite set of conditional attributes and D is a finite set of decision-making attributes. For each $q \in \Omega$, V_q represents the domain of q. Finally, f_q is the information function and is given by $f: U \to V_q$. The elements ($X \subseteq U$) in the information system represent individual cases, states, processes, patients or observations, for example, while the attributes (C & D) can be regarded as the features, variables or characteristic conditions of these elements. The decision-making table (also known as an attribute-value table) is a particular RS information system in which the rows and columns represent elements in the universe of discourse and the attributes of these elements, respectively.

2) Approximation of Sets

In RST, this indiscernability of the elements is handled using the concept of approximate sets.

Assume that $S = (U, \Omega, V_q, f_q)$ is a decision table in which $X \subseteq U$ and $R \subseteq \Omega$. The upper and lower approximates of X are denoted as $R^-(X)$ and $R\_(X)$, respectively, and are defined as

$$R^-(X) = \bigcup\{Y \in U / IND(R) : Y \cap X \neq \phi\} \quad (3)$$

$$R\_(X) = \bigcup\{Y \in U / IND(R) : Y \subseteq X\} \quad (4)$$

where $U / IND(R)$ expresses the equivalence of R and $IND(R)$ denotes the indiscernability of R, i.e.

$$IND(R) = \{(x,y) \in U^2 : \text{for every } a \in R, a(x) = a(y)\} \quad (5)$$

The lower approximate set $R\_(X)$ contains all elements (X) of the same rank when evaluated in terms of the Y decision-making attribute, while the upper approximate set $R^-(X)$ contains the set of all possible same-rank elements (X) when processed in accordance with the Y decision-making attribute. Finally, the set $BN_R(X) = R^-(X) - R\_(X)$ is referred to as the boundary set of X.

C. GM(1,N) model

Imagine a system described by the sequences $x_i^{(0)}(k), i = 1,2,3,...,n$, in which $x_1^{(0)}(k)$ describes the main factor of interest and sequences $x_2^{(0)}(k), x_3^{(0)}(k),...,x_n^{(0)}(k)$ are the factors which influence this main factor. Such a system can be analyzed using the following multivariate GM(1,N) Grey model:

$$x_1^{(0)}(k) + az^{(1)}(k) = \sum_{j=2}^{N} b_j x_j^{(1)}(k) \quad (6)$$

where $k = 2,3,...,n$ in which $x_j^{(1)}(k) = \sum_{i=1}^{k} x_j^{(0)}(i)$ and

$$z_1^{(1)}(k) = 0.5x_1^{(1)}(k) + 0.5x_1^{(1)}(k-1), k \geq 2 \quad (7)$$

Substituting all possible $x_j^{(1)}(k)$ terms into above Equation yields a matrix of the form

$$X_N = \begin{bmatrix} x_1^{(0)}(2) \\ x_1^{(0)}(3) \\ \vdots \\ x_1^{(0)}(n) \end{bmatrix} = \begin{bmatrix} -z_1^{(1)}(2)x_2^{(1)}(2)...x_n^{(1)}(2) \\ -z_1^{(1)}(3)x_2^{(1)}(3)...x_n^{(1)}(3) \\ \vdots \\ -z_1^{(1)}(n)x_2^{(1)}(n)...x_n^{(1)}(n) \end{bmatrix} \begin{bmatrix} a \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = B\hat{a}. \quad (8)$$

Applying the matrix operation $\hat{a} = \left(B^T B\right)^{-1} B^T X_N$, the values of $b_j, j = 2,3,...,N$ can be found. The relative influence exerted on the major sequence by each influencing sequence can then be determined by inspecting the b_j value of the corresponding sequence.

D. Grey relational analysis

In grey relational analysis (GRA), data characterized by the same set of features are regarded as belonging to the same series. The relationship between two series can be determined by evaluating the differences between them and assigning an appropriate grey relational grade. The grey relational grade represents the degree of correlation between two sequences and a new grey relational grade proposed by Huang is defined as

$$\Gamma_{0i} = \frac{\overline{\Delta}_{0i} - \Delta_{\min}}{\Delta_{\max} - \Delta_{\min}} \tag{9}$$

$$\overline{\Delta}_{0i} = \left[\prod_{k=1}^{m} \delta_{0i}(k) \right]^{\frac{1}{m}} \tag{10}$$

$$\delta_{0i}(k) = \frac{|x_i(k)|}{|x_0(k)|} \tag{11}$$

where $i = 1,2,...,m, k = 1,2,...,n$, $x_0(k)$ is the reference value, and $x_i(k)$ is the comparative value. Furthermore, $\Delta_{\min} = \min_{\forall i}\{\overline{\Delta}_{0i}\}$ and $\Delta_{\max} = \max_{\forall i}\{\overline{\Delta}_{0i}\}$.

Having calculated the grey relational grades, the sequences can be ranked using a so-called grey relational ranking procedure. For example, for the case of a reference sequence $x_0(k)$, the grey relational rank of $x_i(k)$ is greater than that of $x_j(k)$ if $\gamma(x_0, x_i) > \gamma(x_0, x_j)$. The corresponding ranking is denoted as $x_i \succ x_j$.

III. The RS Model for Stock Market Forecasting and Portfolio Selection

A. Use of ARX prediction model in preparing data set for rough set processing

Assuming that U is the domain of discourse and R is the set of equivalences of U, then the RS problem can be formulated as

$$X \subseteq U \text{ is } : (R\_(X), R^{-}(X)), BN_R(X) \tag{12}$$

where X is the set of elements; $U / IND(R)$ is the equivalence of R; $IND(R)$ is the indiscernibility of R; ϕ is the zero set; R is the attribute set of X which includes the condition set (C) and the decision-making set (D); $R\_(X)$ is the lower approximate of X; $R^{-}(X)$ is the upper approximate of X; and $BN_R(X)$ is the boundary of X. Every element in the domain of discourse U ($X \subseteq U$) has an attribute set (R), which describes the particular value of X.

In the model developed in this study, every X ($X \subseteq U$) of U is assigned an appropriate set of predicted conditional attribute and decision-making attribute values $R = (C_1, C_2, ..., C_n, D_1, D_2, ..., D_m)$. The resulting attributes are then processed using RST and Grey relational analysis to determine an optimal stock portfolio.

In the model developed in this study, each element (X_i) in U is processed by the ARX prediction model and assigned appropriate values of the conditional attributes ($C_1 \sim C_n$) and decision-making attributes ($D_1 \sim D_m$) based upon their trends over the previous quarter.

B. Detailed processing steps of forecasting and stock selection model

The detailed processing steps in the hybrid forecasting and stock selection model are illustrated in Figure 1. The basic steps in this model can be summarized as follows:

1) Data collection and attribute determination

In the current model, the conditional attributes should reflect the financial quality of a company. Therefore, the proposed model specifies the following attributes: the profitability, the capitalized cost ratio, the individual share ratio, the growth rate, the debt ratio, the operational leverage and all statutory financial ratios. The decision as to which of the selected companies should actually be processed by the RST portfolio selection mechanism is then made by processing the forecast data generated by the ARX model in accordance with the seven decision-making attributes specified in the previous section.

2) Data preprocessing

Having collected the relevant financial data every quarter, a basic pre-processing operation is performed to improve the efficiency of the ARX prediction model.

For example, any data records containing missing fields (i.e. attributes) are immediately rejected. In addition, the problem of data outliers is addressed by using the Box Plots method [28] to establish an inter-quartile range such that any data falling outside this range can be automatically assigned a default value depending on the interval within which it is located.

3) ARX prediction

As discussed earlier, the stock market is a sequentially correlated signal system, and hence the current model uses an ARX forecasting model to predict the future trends of the financial variables of each of the selected companies. In the current ARX model, forecasting is deliberately restricted to a one-step-ahead mode to prevent the accumulation of errors from previous forecasting periods.

4) Information reduction using GM(1,N) multivariate model

To improve the efficiency of the RST / Grey relational analysis process, certain conditional attributes are removed if they are found to have little effect on the decision-making attributes. In the proposed model, this pruning operation is performed by using the GM(1,N) method to identify the top ten influential sequences (i.e. conditional attributes) and then removing the remainder.

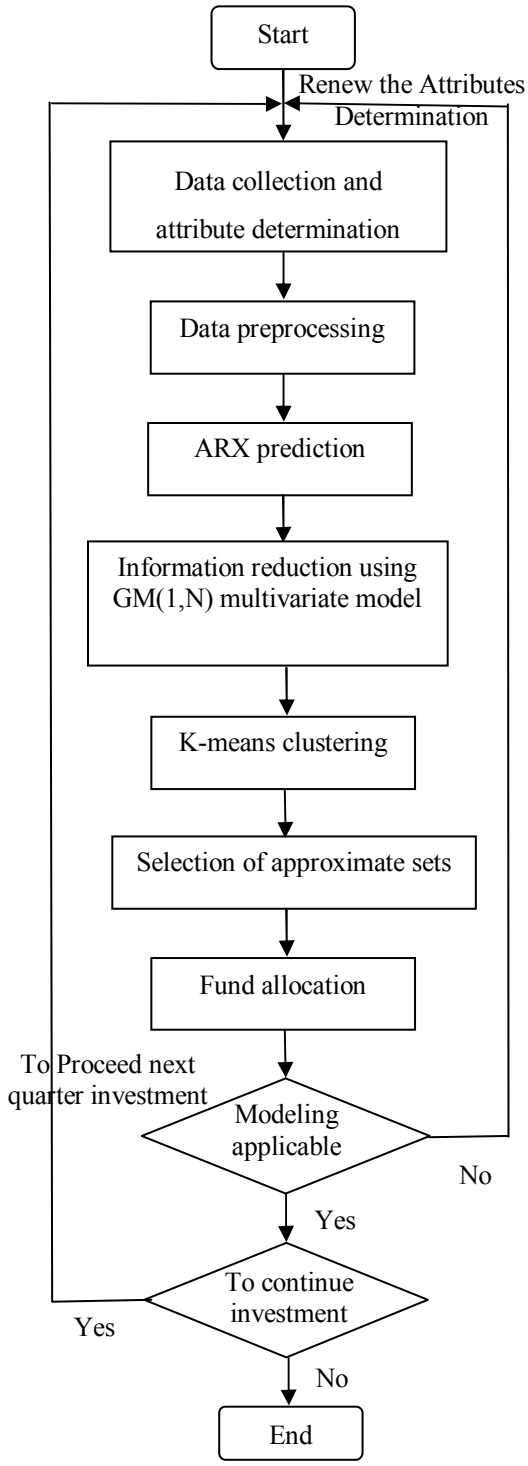

Figure 1. Flow chart of forecasting and stock selection model.

5) K-means clustering

Prior to submission to the RST stock selection mechanism, the forecast data of conditional attribute values ($C_1 \sim C_n$) are clustered into three groups using a K-means clustering algorithm.

6) Selection of approximate sets

Having clustered the forecast data, the rough set method is applied to determine the lower approximate set. The generalized rules extracted by the low approximate set were all recognized rules or relationships in the investment industry.

This indicates that rough set analysis is a feasible means of identifying top stock performers by classifying the contributions of the attributes, and is also helpful in constructing decision rules which might be applied to evaluate new stocks. In other words rough set theory avoids the need for the blind, haphazard stock selection methods commonly employed by investors in the past.

7) Fund allocation

Having identified suitable stocks in which to invest, it is then necessary to determine an appropriate weight of the stocks in the portfolio in order to optimize the overall rate of return. In the current model, this fund allocation problem is processed using a Grey Relation Sequence based on $stock\ weight(i) = \frac{N-i+1}{\sum_{i=1}^{N} i}$, where i is the grey relation order of each stock item, and N is the total number of invested stocks. Having completed all the steps described above, a check is made of the overall rate of return on the investment. If the rate of return is acceptable, a decision is made as to whether or not the model should be run for a further quarter using the existing attributes. However, if the rate of return is unacceptable, the suitability of the conditional attributes is reviewed and amended if appropriate.

IV. Evaluation of Proposed Model Using Electronic Stock Data

A. Data extraction

The feasibility of the proposed forecasting and stock selection model was evaluated using electronic stock data extracted from the New Taiwan Economy database (TEJ). The data collection period extended from the first quarter in 2003 to the fourth quarter in 2006, giving a total of 16 quarters in all. Meanwhile, the forecasting period extended from the first quarter in 2003 to the first quarter in 2007, giving a total of 17 quarters.

In general, financial statements for a particular accounting period are subject to a considerable delay before they are actually published. For example, annual reports are published after four months, half-yearly reports after two months, and first and third quarterly reports (without notarization) after at least one month.

TABLE I. RATES OF RETURN OF THE HYBRID MODE

Second quarter in 2004	-0.09%
Third quarter in 2004	-1.38%
Fourth quarter in 2004	13.43%
Year rate of return in 2004	11.96%
Second quarter in 2005	14.96%
Third quarter in 2005	0.43%
Fourth quarter in 2005	10.56%
Year rate of return in 2005	25.95%
Second quarter in 2006	0.13%
Third quarter in 2006	5.00%
Fourth quarter in 2006	34.70%
Year rate of return in 2006	39.82%
Average year rate of return	25.91%

Since the last quarter data every year can not be acquired until 31st May in the following year, the data can not be used by the ARX model to predict the financial trends over the first quarter of the year. In other words, the forecasting and investing process proposed in this study can only be conducted three times each year, i.e. 5/31~09/22, 9/22~11/15 and 11/15~05/31 next year. In addition, in the decision-making rules used in the rough set stock selection process, the Return on Equity (ROE) and constant EPS indicators are based on the full 12 months of the previous year. Thus, the forecasting period for investment purposes is actually reduced to the second quarter in 2004 to the fourth quarter in 2006.

B. Verification of a hybrid model performance

Table 1 summarizes the quarterly and yearly rates of return obtained over the nine investment periods between 2004 and 2006 using this hybrid model. As shown, this model achieves a significant average yearly rate of return.

CONCLUSIONS AND DISCUSSIONS

This study has presented a fusion model based upon the ARX prediction model, Grey System theory and Rough Set theory for the selection of an optimal stock portfolio. The major findings and contributions of this study can be summarized as follows:

- This study has confirmed the effectiveness of combining different forecasting techniques to improve the efficiency and accuracy of automatic prediction and classification algorithms.
- The process of information reduction usually obtained from the Rough Sets model is based on the principle of eliminating any redundant or unimportant information in the original data.
- The efficacy of the fusion model has been evaluated by the set of history data, the hybrid model provides a highly accurate forecasting performance.

Overall, the results presented in this study have confirmed that the proposed fusion model provides a promising method for stock portfolio management. Furthermore, the structure of the proposed model represents a suitable foundation for a broad range of derivatives in other research fields. In future studies, the use of alternative grey relational analysis techniques will be considered in order to evaluate the potential for further improving the rate of return obtained from the selected stocks.

REFERENCES

[1] R.J. Bauer Jr., Genetic Algorithms and Investment Strategies. New York :Wiley, 1994.

[2] Md. Ra.ul Hassan, Baikunth Nath, Michael Kirley, "A fusion model of HMM, ANN and GA for stock market forecasting," Expert Systems with Applications, vol. 33, pp. 171-180, 2007.

[3] Qing Cao, Karyl B. Leggio, Marc J. Schniederjans, " A comparison between Fama and French's model and artificial neural networks in predicting the Chinese stock market," Computers & Operations Research, vol. 32, pp. 2499-2512, 2005.

[4] R.J. Kuo, C.H. Chen, Y.C. Hwang, "A intelligent stock trading decision support system through integration of genetic algorithm based fuzzy neural network and artificial neural network," Fuzzy Sets and Systems, vol. 118, pp.21-45, 2001.

[5] Aiken, M., & Bsat, M, "Forecasting market trends with neural networks," Information Systems Management, vol. 16, no. 4, pp. 42-48, 1999.

[6] Romahi, Y., & Shen, Q, "Dynamic financial forecasting with automatically induced fuzzy associations," In Proceedings of the 9th international conference on fuzzy systems, San Antonio, TX, USA, pp. 493-498, 2000.

[7] Abraham, A., Nath, B., & Mohanthi, P. K, "Hybrid intelligent systems for stock market analysis," In Vassil N. Alexandrov et al. (Eds.), Computational science, Germany, San Fransisco, Springer-Verlag, USA, pp. 337-345, 2001.

[8] Box, G. E. P., & Jenkins, G. M, Time series analysis: forecasting and control, San Fransisco, CA: Holden-Day, 1976.

[9] Pankratz, A, "Forecasting with Univariate Box-Jenkins models: Concepts and Cases," New York: John-Wiley, 1983.

[10] Engle, R. F., "Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation," Econometrica, vol. 50, pp. 987-1008, 1982.

[11] Raymond, Y. C. Tse, " An application of the ARIMA model to real estate prices in Hong Kong," Journal of Property Finance, vol. 8, no.2, pp. 52-163, 1997.

[12] Kolarik, T., & Rudorfer, G.., "Time series forecasting using neural networks," Time Series and Neural Networks, ACM APL Quote Quad, vol. 25, no. 1, pp.86-92, 1994.

[13] C. Skalko, "Rough sets help time the OEX," Journal of Computational Intelligence in Finance, vol. 4, no. 6, pp. 20-27, 1996.

[14] Pawlak, Z., "Rough sets," International Journal of Information and Computer Sciences, vol. 11, no. 5, pp. 341-356, 1982.

[15] Tsumato, S., Slowinski, S., Komorowsk, J., & Grzymala-Busse, J. W., "Lecture notes in artificial intelligence," The fourth international conference on rough sets and current trends in computing (RSCTC'2004), Uppsala, Sweden, 2004.

[16] Chien-Chung Chan, " A rough set approach to attribute generalization in data mining," Journel of Information Sciences, vol. 107, pp. 169-176, 1998.

[17] Grzymala-Busse, J.W., "LERS– A knowledge discovery system," In Polkowski, L., Skowron, A. (Eds.), Rough Sets in Knowledge Discovery 2 (562-565), Physica-Verlag, Wurzburg, 1998.

[18] Jerzy B aszczyński , Salvatore Greco, Roman Slowiński, " Multi-criteria classification -A new scheme for application of dominance-based decision rules," European Journal of Operational Research, vol. 181, pp.1030-1044, 2007.

[19] Yi-Fan Wang, " Mining stock price using fuzzy rough set system," Expert Systems with Applications, vol. 24, pp. 13-23, 2003.

[20] Lixiang Shen, Han Tong Loh, "Applying rough sets to market timing decisions," Decision Support Systems, vol. 37, no. 4, pp. 583- 597, 2004.

[21] Bazan, J.G., Szczuka, M., "RSES and RSESlib – A collection of tools for rough set computations," In Ziarko, W., Yao, Y.Y. (Eds.),

Proceedings of the Second International Conference on Rough Sets and Current Trends in Computing (RSCTC'2000), Canada: Banff, pp.74–81, 2000.

[22] Ljung, L., System identification: theory for the user. 2nd ed., Englewood Cliffs, NJ: Prentice-Hall, 1999.

[23] So Young Sohn, Michael Lim, "Hierarchical forecasting based on AR-GARCH model in a coherent structure," European Journal of Operational Research, vol 176, pp.1033-1040, 2007.

[24] Geetesh Bhardwaj, Norman R. Swanson, " An empirical investigation of the usefulness of ARFIMA models for predicting macroeconomic and financial time series," Journal of Econometrics, vol. 131, pp. 539-578, 2006.

[25] Massimiliano Marcellino, James H. Stock, Mark W. Watson, " A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series," Journal of Econometrics, vol. 135, pp. 499-526, 2006.

[26] Deng, J.L., "Introduction to Grey System Theory," The Journal of Grey System vol. 1, no. 1, pp. 1-24, 1989.

[27] Robert G. Hagstrom, Bill Miller, and Kenneth L. Fisher , The Warren Buffett Way: Investment Strategies of the World's Greatest Investor, JOHN WILEY & SONS (ASIA) PTE LTD, 2005.

[28] Chakravarti, Laha, and Roy, Handbook of Methods of Applied Statistics, Volume I, John Wiley and Sons, 1967.