International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM)
ISSN: 2150-7988 Vol.1 (2009), pp.170-176
http://www.mirlabs.org/ijcisim

Combination of Forecasting Using Modified GMDH and Genetic Algorithm

Ruhaidah Samsudin
Department of Software Engineering
Faculty of Computer Science and Information System
University Technology of Malaysia, 81310 Skudai,Johor, Malaysia.
Tel:607-5532228, fax: 607-5565044
ruhaidah@utm.my

Puteh Saad
Department of Software Engineering
Faculty of Computer Science and Information System
University Technology of Malaysia, 81310 Skudai, Johor, Malaysia
607-5532344, fax: 607-5565044
puteh@utm.my

Ani Shabri
Department of Mathematic
Faculty of Science
University Technology of Malaysia, 81310 Skudai, Johor, Malaysia.
607-5534238
ani@utm.my

Abstract

Many studies have demonstrated that combining forecasts improves accuracy relative to individual forecasts. In this paper, the combing forecasts is used to improve on individual forecasts is investigated. A combining approach based on the modified Group Method Data Handling (GMDH) method and genetic algorithm (GA), is called as the GAGMDH model is proposed. Four time series forecasting techniques are used as individual forecast, namely linear regression, quadratic regression, exponential smoothing and ARIMA models. The forecasted results of individual forecasting models are used as the input of combining forecasting, and the outputs are the results of combination forecasting. To assess the effectiveness of the GAGMDH model, we used the time series yearly cancer death rate in Pennsylvania. The empirical results with a real data set clearly suggest that the GAGMDH model can improve the forecasting capability of the model compared with optimal simple combining forecasting methods and neural networks combining forecasting methods.

1 Introduction

There are numerous models available for forecasting time series data. However, when building a forecasting model, it is not an easy task to choose a suitable model, because no one model is powerful and general enough to outperform the others for all types of time series data, and every model has some degree of uncertainty, including structure uncertainty and parameter uncertainty. After the seminar paper of Bates and Granger [24], many combination methods have been proposed such as a simple average method, weighted average method [12], regression based methods [6], a neural network method, [11] and the self-organizing data mining algorithms [1].

The main purpose of combining forecasts is to better use the useful information provided by different forecasting models in order to improve the forecast accuracy. For a given forecast problem, different forecasting models can provide different types of information. If we only choose one forecasting model and simply discard the rest, then a lot of useful information will be lost. However, by combining individual models in a proper way, we can fully use the information from each model and improve the overall performance of forecasts. Many empirical evidences have shown that a combination of forecasts often outperforms individual forecasts [4, 14].

Combing forecasting methods can be generally be classified into two categories: the linear ones and the nonlinear ones. The linear combination methods are the weighted combination of forecasting models, including the optimal combining method and variable weighting methods [6]. These models are based on the assumption that there exits a linear relationship among the combined forecasting models. However, when some of the combined forecasting methods are derived from nonlinear models, or the conditional expectation, on which each individual forecasting model is based, is a nonlinear function of the information set, then a linear combination of individual forecasting methods is, on the whole, not the optimal method.

The second category is composed of nonlinearly combined forecasting method. Combing forecasting based on artificial neural network (ANN) has recently attracted the interest of researchers [5,18]. Experimental results have shown that this kind of combination can increase the fitting accuracy and the reliability of forecasting.

He in [1] and Sun & Zhang in [17] introduced self-organizing data mining (SODM) algorithm as the new combination forecasting model. Compared with ANN combining forecasting methods, the new method can improve the forecasting capability of the model. SODM algorithm is based on the Group Method Data Handling (GMDH) method and utilizes a class of polynomials such as linear, modified quadratic, cubic, etc.

In order to overcome the above-mentioned shortcomings of linear and nonlinear combining forecasting models, we introduce the GAGMDH method that combining the Genetic Algorithm (GA) and modified GMDH model. The GA is used to find the optimum parameters for improving the GMDH model of the parameter estimation method. The new method, like artificial neural network methods, is a nonlinear combination method, have several advantages compared with ANN. It has the ability of self-selecting useful input variables. Also, useless input variables are eliminated and useful input variables are selected automatically. The structure parameters and the optimum GMDH architecture can be organized automatically [15].

2. Forecasting Methodology

2.1 The Single Forecasting Models

The four the single forecasting models that were employed are the linear regression, quadratic regression, exponential smoothing and autoregressive integrated moving average (ARIMA) methodology.

2.1.1 Simple Regression Model. Simple linear regression analysis is one of the most widely used techniques for modeling a linear relationship or trend between two variables (y_t and time). The simple linear regression model involves a single predictor variable and is written as [2]

$$y_t^L = \beta_0 + \beta_1 t + \varepsilon_j \tag{1}$$

where β_0 and β_1 are unknown parameters.

2.1.2 The Quadratic regression. Most mathematical curves can be closely approximated by n-degree polynomials. However, the quadratic or the second-degree polynomial is the specification most commonly used. The model can be expression as in [3].

$$y_t^Q = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon_j \tag{2}$$

where β_0 , β_1 and β_2 are unknown parameters

2.1.3 Exponential Smoothing. In the application of the exponential smoothing model, there are three types of models that are widely used in different time series [7]. Simple exponential smoothing is used when the time series has no trend, double exponential smoothing for handling a time series that displays a slowly changing linear trend and the Winters' method which is an exponential smoothing approach to predicting seasonal data. In this study, the double exponential smoothing is suitable for this time series forecasting for the specified time period. The model employs a level component and a trend component at each period. It uses two weights, or smoothing parameters to update the components at each period. The double exponential smoothing equations are

$$y_t^E = L_{t-1} + T_{t-1}, \quad t = 1, 2, ..., n \tag{3}$$

$$L_t = \alpha Y_t + (1-\alpha)(L_{t-1} + T_{t-1})$$
$$0 < \alpha \le 1$$

$$T_t = \gamma(L_t - L_{t-1}) + (1-\gamma)T_{t-1}$$

where L_t is the level at time *t*, α is the weight for the level, T_t is the trend at time *t*, γ the weight for the trend, Y_t is the data value at time *t* and $\hat{Y}_t$ is the fitted value.

2.1.4 The ARIMA model. One of the most popular and widely used time series models is ARIMA model. The popularity of the ARIMA model is due to its statistical properties as well as the well-known Box-Jenkins methodology in the model building processes [23]. The general non-seasonal model is known as ARIMA (p, d, q) can be written as the linear expression

$$y_t^{BJ} = \sum_{i=1}^{p} \phi_i y_{t-1} + \sum_{j=1}^{q} \theta_j a_{t-j} + \varepsilon_j \quad (4)$$

where ϕ_i and θ_j are the autocorrelation and the moving average coefficients respectively.

2.2 Forecasting Combination Models

Several forecast combination methods have been developed in the literature. In this study, three combination forecast methods (the simple combination, artificial neural network and GAGMDH) are used to test the performance of the different forecasting models.

2.2.1 The simple combination method. In combining the forecasts generated by two or more models, it is important to decide the weights which will be assigned to each the models. In the simple combination (SC) method, the combination weight is assigned equally to each of the forecast. The combination forecast is given by

$$y_t^S = \frac{1}{m}\sum_{i=1}^{m} y_i^F \quad (5)$$

where y_i^F is the ith single forecast and y_t^S is the combined forecast generated by the m single forecasts y_i^F.

2.2.2 Artificial Neural Network. ANN approaches gained popularity as a tool in time series and widely applied to a variety of practical problems. ANN is flexible computing frameworks for modeling a broad range of nonlinear problems. The ANN with single hidden layer feed forward network is the most widely used model for modeling and forecasting. The model can be written has following mathematical as:

$$y_t = a_0 + \sum_{j=1}^{q} a_j f(b_{0j} + \sum_{i=1}^{p} b_{ij} y_{t-i}) + \varepsilon_t \quad (6)$$

where a_j and b_{ij} are the model parameters [4]. The logistic function is used as the hidden layer transfer function, that is

$$f(x) = \frac{1}{1+\exp(-x)}$$

The general form of ANN model in Eq.(6) can be expressed as

$$y_t^{ANN} = f(y_t, y_{t-1}, y_{t-2}, y_{t-3})$$

where f is a function determined by the network structure and connection weights.

2.2.3 The GAGMDH model. General connection between inputs and output variables can be expressed by a complicated polynomial series in the form of

$$y = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{j=1}^{n}\sum_{i=1}^{n} a_{ij} x_i x_j + \sum_{k=1}^{n}\sum_{j=1}^{n}\sum_{i=1}^{n} a_{ijk} x_i x_j x_k + \dots \quad (7)$$

which is known as the Volterrs-Kolmogorov-Gabor (VKG) polynomial [19]. In the GMDH algorithm, the VKG series is estimated by the partial quadratic polynomial using only pairs of variables. The form of mathematical description of this polynomial is

$$y = a_0 + a_1 x_i + a_2 x_j + a_3 x_i x_j + a_4 x_i^2 + a_5 x_j^2$$

In this study, the form of mathematical of the GMDH algorithm is modified and the computation process of modified GMDH comprises three basic steps:

Step 1: First n observations of regression-type data are taken. These observations are divided into two sets: the training set and testing set. The first layer model is obtained in every column of the training sample of observations. In each layer, the modified quadratic polynomial

$$z = a_0 + a_1 x_i + a_2 x_j + a_3 x_i x_j + a_4 x_i^2 + a_5 x_j^2 + a_6 x_j x_i^2 + a_7 x_i x_j^2 + a_8 x_i^3 + a_9 x_j^3 \quad (8)$$

is adopted as the transfer function. The coefficients $a_0\ a_1\ a_3 ... a_9$ are evaluated using GA. The GA is a global optimization approach based on the Darwinian principles of natural selection. Developed from the concept of Holland [20], it seeks the extreme of complex function efficiently – see [21] for a detailed description.

The evolution starts from generate an initial population of parameters ($a_0\ a_1\ a_3 ... a_9$) for each model by the real random number. In each generation, the fitness of the whole population is evaluated, multiple individuals are stochastically selected from the current population based on their fitness and modified (mutated or crossover) to form a new population (offsprings). The new population is then used in the next generation of the algorithm. After many generations, only those individuals (solutions) with the best fitness survive. The best individuals provide an optimum solution to the problem [22].

The GA approach presented in this work is employed to find the optimum values of the parameters that minimize the sum square error of the actual data, y and prediction data, *z*. The fitness function, E adopted is

$$\mathrm{E} = \frac{1}{n}\sum_{t=1}^{n}(y_t - z_t)^2$$

Step 2: Construct $M' = M(M-1)/2$ new variables in the training data set for all possibilities of connection by each pair of neurons in the layer. A small number of variables that give the best results in the first layer, are allowed to form second layer candidate models using as the Eq. (8).

Step 3: Select the best neuron out of these M' neurons, according to the value of mean square error (MSE). The MSE is defined by the formula:

$$\mathrm{MSE} = \frac{1}{nt}\sum_{i=1}^{nt}(y_i - z_i)^2$$

where *nt* is the number of observations in the testing set and z is the estimated output values. Once the best neurons, L are selected, the MSE in each layer is further checked to determine whether the set of equations of the model should be further improved within the subsequent computation. The lowest value of the selection criterion obtained during this iteration is compared with the smallest value obtained at the previous one. If an improvement is achieved, then set new input $\{z_1 = x_1, z_2 = x_2, ..., z_L = x_L\}$, $M' = L$ and repeat steps 2 and 3. Otherwise the iterations terminate and a realization of the network has been completed.

3 An Application

The yearly cancer death rate (all forms, per 100 000 population) of Pennsylvania between 1930 and 2000 published in the 2000 Pennsylvania Vital Statistics Annual Report taken from [10] has been considered to demonstrate the effectiveness of the GMFHGA model. The data series is regarded as linear trend and is used to evaluate the effectiveness of combination linear and nonlinear models. The data of this time series are plotted in Figure 1, which show a non-stationary with an increasing linear trend. The data set has 71 observations, with 66 observations from 1930 to 1995 was selected for training and 5 observations from 1996 to 2000 was selected for testing the forecasting.

The performances of the each model for both the training data and forecasting data are evaluated according to the mean absolute error (MAE) and root-mean-square error (RMSE), which are widely used for evaluating results of time series forecasting. The MAE and RMSE are defined as

$$\mathrm{ARE} = \left|\frac{y_t - y_t^F}{y_t}\right|$$

$$\mathrm{MAE} = \frac{1}{nt}\sum_{t=1}^{nt}\mathrm{ARE}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{nt}\sum_{t=1}^{nt}(y_t - y_t^F)^2}$$

where y_t and y_t^F are the observed and the forecasted data series at the time *t*. The criterions to judge for the best model are relatively small of MAE and RMSE in the modeling and forecasting.

4. Results

In this study, the linear regression model, quadratic regression models, exponential smoothing model and ARIMA model is implemented via MINITAB package, while ANN and GAGMDH models are built using the MATLAB package. The results of different individual forecasting models are shown in Table 1.

In this study the outputs y_t^L, y_t^Q, y_t^E, and y_t^{BJ}, of the above four individual forecasting models are used as the input nodes. By linearly combining with the same weight of the above four individual forecasting models, we have the following optimal combination forecasting of the simple combination model

$$y_t^S = \frac{1}{4}(y_t^L + y_t^Q + y_t^E + y_t^{BJ})$$

Table 1. The Individual forecasting model

Model	Formula
Linear Trend	$y_t^L = 104.292 + 2.23014t$
Quadratic Trend	$y_t^Q = 107.463 + 1.9503t + 0.00418t^2$
Double Exponential Smoothing	$y_t^E = a_t + b_t\, t\,;\quad \alpha = 1.23344$
ARIMA	$y_t^{BJ} = \text{-}0.2289 y_{t-1} + 2.8062$

For the artificial neural network combination forecasting model, three values (2, 4, and 8) for the number of neurons in the single hidden layer were tried in order to investigate the effects of model complexity on forecasting performance. Among the 3 neural network architectures, the model with 4 input nodes and 8 hidden nodes has the best test results, and is selected as the ANN model for subsequent analysis. The ANN model has the form

$$y_t^{ANN} = f(y_t^L, y_t^Q, y_t^E, y_t^{BJ}) \text{ (No explicit form)}$$

The GAGMDH forecasting model considered in this study is developed using the outputs y_1, y_2, y_3 and y_4 of the individual forecasting model. For the parameters estimation experiments, a MATLAB package is employed. Each generation has 100 chromosomes. The crossover probability is set to 0.80, implying that 80% of the chromosomes in a generation are allowed to crossover in an average sense. The maximum and minimum mutation probability is set to 0.05 and 0.005, respectively. By using the technique of GAGMDH, the computer automatically generates the following model

$$y_t^{GAGMDH} = f(y_t^L, y_t^Q, y_t^E, y_t^{BJ})$$

4.1 Comparisons among the seven forecasting models

A comparison of fitting error, forecast errors of combination models and the individual models is shown in Table 2. Table 2 reports the forecast results, ARE, MAE and RMSE for each of the individual forecasts and each combined forecast for the yearly cancer death rate from 1996 to 2000. It should be noted that, the ARE (%) values in bold style indicate the corresponding model behaves best among all models/methods for the specific leading time. A comparison between the actual value and the forecast value for the GAGMDH model is given in Figure 1. The forecast performance of each model was evaluated using graphical and statistical comparison as shown in Figure 2. Based on the MAE and RMSE, it is clear that the combined forecasts based on ANN and GAGMDH are more accurate than any of the individual forecasts. The GAGMDH method is superior to other individual forecasting methods or combination forecasting methods. Figure 2 graphically supports these results.

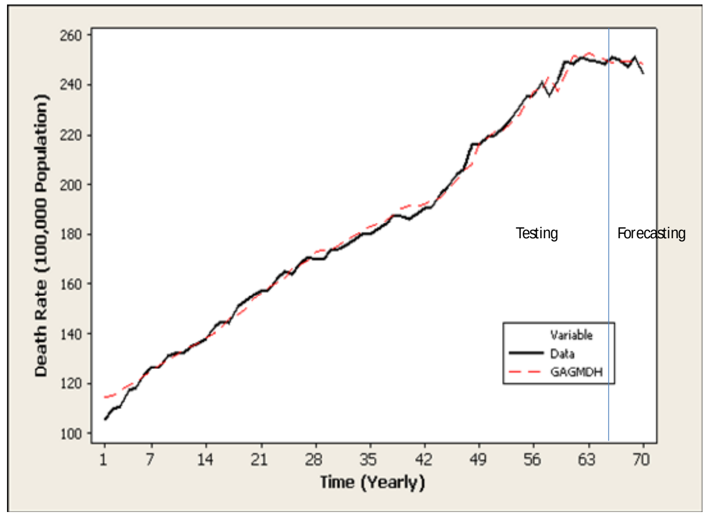

Figure 1: The yearly cancer death rate data and the GAGMDH prediction of the data series.

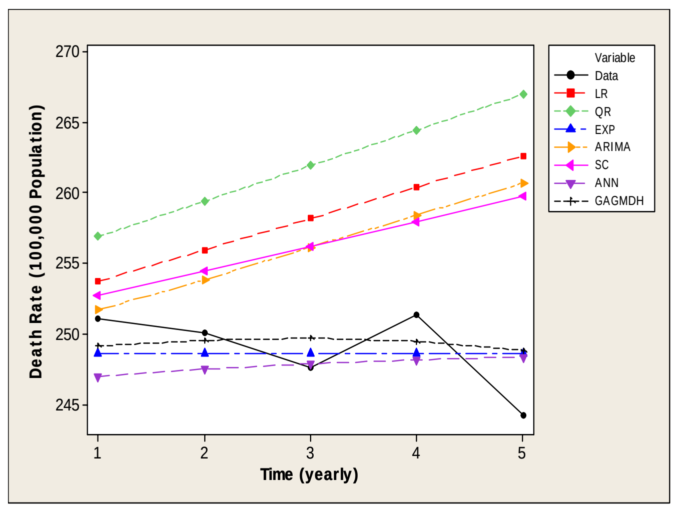

Figure 2: Forecasting comparison for the yearly cancer death rate data from 1996 to 2000

5. Conclusion

In order to improve the forecasting capability of the combination forecasting, this paper proposes the GAGMDH forecasting method. Three combination methods have been used to compare the forecasting accuracy of combined forecasts against single model forecasts. Compared with optimal linear combining forecasting methods and neural networks combining forecast methods, the GAGMDH method can improve the forecasting capability of the model. Our study of the GMDH and real applications suggests that the GMDH method proposed is a powerful and vital method for simulation, analysis, forecast and combination forecast of time series data. It is a model with good prospects and we expect further developments and applications in the field of forecasting.

6. Acknowledgment

The authors thank Ministry of Science, Technology and Innovation, Malaysia for granting this research and Soft Computing Research Group for continuous support.

7. References

[1] He, C. & Xu, X., 2003, "Combination Forecasts Using Self-Organizing Algorithms", *Journal of Forecasting* , 24: 269-278.

[2] Kutner, M.H., Nachtsheim, C.J. & Neter, J., 2004, *Applied Linear Regression Models*, McGraw-Hill, Inc.

[3] Gaynor, P.E & Kirkpatrick, R.C., 1994, *Introduction to Time-Series Modeling and Forecasting in Business and Economics*, McGraw-Hill, Inc.

[4] Zhang, G.P., 2003, "Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model", *Neurocomputing*, 50: 159-175.

[5] Cho, S. & Yoon, S., 1997," Reliable Roll Force Prediction in Cold Mill Using Multiple Neural Networks", *IEEE Transactions on Neural Networks*, 8:874-882.

[6] Granger, C.W. J. & Ramanathan, R., 1984, "Improved method of Forecasting", *Journal of Forecasting*, 3: 197-204.

[7] Hyndman, R.J., Snyder, A.B. and Grose, S. ,2002, "A State Space Framework for Automatic Forecasting Using Exponential Smoothing Methods", *International Journal of Forecasting*, 18: 439-454.

[8] Lai, K.K., Yu, L., Wang, S. & Huang, W., 2006," Hybridizing Exponential Smoothing and Neural Network for Financial Time Series Predication", *ICCS 2006*, Part IV, LNCS 3994: 493-500.

[9] Zou, H.F., Xia, G.P., Yang, F.T. & Wang, H.Y., 2007, " An Investigation and Comparison of Artificial Neural Network and Time Series Models for Chinese Food Grain Price Forecasting", *Neurocomputing*, 70: 2913-2923.

[10] Wei, 2006. *Time Series Analysis : Univariate and Multivariate Methods (2nd Edition).* Pearson Addison Wesley.

[11] Sfetsos, A. and Siriopoulos, C., 2004, "Combinatorial time series forecasting based on clustering algorithms and neural networks", *Neural Computing and Applications* 13(1): 56-64.

[12] Winkler, R.L., 1984, "Combining probability distributions from dependent information sources", *Management Science* **27** (1981), pp. 479–488.

[13] Gunter, SI., 1989, "N-step combinations of forecasts", *Journal of Forecasting,* 8:253-267.

[14] Clemen, R.T. and Winkler, R.L., 1986, "Combining economic forecasts", *Journal of Business and Economic Statistics*, 4: 39-46.

[15] Kondo,T., 2008, " Feedback GMDH-type neural network algorithm using prediction error criterion for self-organization", *SICE Annual Conference 2008*, 1045-1049.

[16] Wang, W., Gelder, P.V and Vrijling, J.K. , 2005, "Improving Daily Stream Flow Forecasts By Combining ARMA and ANN Models", *International Conference on Innovation Advances and Implementation of Flood Forecasting Technology*.

[17] Sun, W. and Zhang, X., 2007, "Application of self-organizing combination forecasting method in power load forecast", *International Conference on Wavelet Analysis and Pattern Recognition, ICWAPR '07*. Vol: 2, p: 613-617.

[18] S. Ruhaidah, . S. Ani and S. Puteh., 2008, " A Comparison of Neural Network, ARIMA model and Multiple Regression Analysis in Modeling Rice Yield", *International Journal of Soft Computing Applications*. 3:113-127.

[19] Farlow, S.J., 1984, *Self-organizing methods in modeling, GMDH-type algorithms*, Marcel Dekker, Inc, New York.

[20] Holland, J., 1975, *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, MI, pp:183.

[21] Goldberg, D.E., 1989, *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, Reading, MA, pp:432.

[22] A. Cobo, A. Galvez, J. Puig-Pey, A. Iglesias and J. Espinola. Bio-Inspired Metaheuristic Methods for Fitting

Points in CAGD. *International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM).* First Special Issue on "Computer Graphics and Geometric Modelling ", 2008.

[23] Box, G.E.P and Jenkins, G., 1970, "Time Series Analysis", *Forecasting and Control*, Holden-Day, San Francisco, CA.

[24] Bates, J.M. and Granger, C.W.J., 1969, "The Combination of Forecasts", *Oper. Res. Quart.* Vol. 20, pp. 451–468.

About the authors

S. RUHAIDAH is a lecturer at the the Faculty of Computer Science and Information System, Universiti Teknologi Malaysia, Malaysia. She received her Bachelor of Computer Science and Master of Computer Science degrees from Universiti Teknologi Malaysia, Skudai, Malaysia in 1995 and 1998, respectively. She is currently a PhD student at the same institute. Her research interests include soft computing and time series forecasting.

S. PUTEH is an Prof. at the Faculty of Computer Science and Information System, Universiti Teknologi Malaysia, Malaysia. She obtained her Bachelor of Science (Chemistry) in 1981 from Universiti Malaya, Malaysia, Master of Computer Science from Universiti of Stratchlyde, Glasgow, Scotland in 1987. She Then she received her PhD in Computer Science from Universiti Teknologi Malaysia in 2003. currently teaches Object Oriented Programming and Artificial Intelligence. She has collaborated with government and industry in the past 20 years. Among the project are Smart Calculator for Water Use Efficiency for MARDI from 2000 to 2001; System for the Processing of Teacher Applicants, from June 1983 to Nov 1983, Bahagian Latihan dan Perguruan, Ministry of Education, Malaysia; Registration and Updating System for IEM (Institute of Engineers Malaysia). Her research interest are pattern recognition and soft computing.

S. ANI is a senior lecturer at Department of Mathematic, Faculty of Science, Universiti Teknologi Malaysia, Skudai, Malaysia. He obtained his Bachelor and Master degrees in Industrial Science (Mathematics) from Universiti Teknologi Malaysia in 1992 and 1995 respectively. Then he received his PhD in Mathematics from Universiti Kebangsaan Malaysia in 2007.He has published several papers on flood frequency analysis, soft computing and time series. His research interests are the analysis of flood frequency and time series.

Table 2 Comparison forecasting precision among the seven models

Year	Data	y_t^L	y_t^Q	y_t^E	y_t^{BJ}	y_t^S	y_t^{ANN}	y_t^{GAGMDH}
1996	251.1	253.71	256.88	248.55	251.69	252.71	246.99	249.13
ARE(%)		(-1.04)	(-2.30)	(-1.02)	**(-0.23)**	(-0.64)	(-1.64)	(-0.78)
1997	250.1	255.94	259.4	248.55	253.81	254.43	247.48	249.53
ARE(%)		(-2.34)	(-3.72)	(-0.62)	(-1.48)	(-1.73)	(-1.05)	**(-0.23)**
1998	247.6	258.17	261.92	248.55	256.13	256.19	247.86	249.69
ARE(%)		(-4.27)	(-5.78)	(-0.38)	(-3.45)	(-3.47)	**(-0.11)**	(-0.84)
1999	251.3	260.4	264.45	248.55	258.41	257.95	248.15	249.45
ARE(%)		(-3.62)	(-5.23)	(-1.09)	(-2.83)	(-2.65)	(-1.25)	**(-0.74)**
2000	244.2	262.63	266.99	248.55	260.69	259.72	248.37	248.81
ARE(%)		(-7.55)	(-9.33)	(-1.78)	(-6.75)	(-6.36)	**(-1.71)**	(-1.89)
RMSE		10.726	14.264	7.259	9.049	8.720	3.197	**2.580**
MAE		0.038	0.0527	0.099	0.030	0.030	0.012	**0.009**

The minimal errors are highlighted in bold.