THE GMDH MODEL AND ITS APPLICATION TO FORECASTING OF RICE YIELDS

Ruhaidah Samsudin[1], Puteh Saad[1], Ani Shabri[2]

[1]Faculty of Computer Science and Information Systems
University Teknologi Malaysia
81300 Skudai, Johor

[2]Department of Mathematic, Science Faculty, University Technology of Malaysia

Email: {ruhaidah, puteh, ani}@utm.my

Abstract: In this paper, the group method of handling (GMDH) model and their application to the forecasting of the rice yields time series are described. The use of such GMDH leads to successful application in broad range of areas. However, in some fields, such as rice yields forecasting, the use GMDH is still scare. Artificial neural networks (ANN) have been shown to be powerful tools for system modeling. This study addressed the question of whether GMDH could be used to estimate more accurate in modeling and forecasting compared with the ANN model. To assess the effectiveness of these models, we used 9 years of time series records for rice yield data in Malaysia from 1995 to 2001. The results demonstrate that GMDH model is superior to the ANN for rice yield forecasting.

Keywords: ***GMDH, Artificial Neural Network (ANN), rice yields, autocorrelation function, partial autocorrelation function.***

1. INTRODUCTION

The accuracy of time series forecasting is fundamental to many decisions processes ([21]. One of the most important and widely used time series model is the artificial neural network (ANN). The ANN provides an attractive alternative tool for both forecasting researchers and has shown their nonlinear modeling capability in data time series forecasting because of its flexibility in building models without explicit physical representations which may not be well described in most complex non-linear characteristics from inputs which consist of pattern, noise and irrelevant data [15]. A number of investigations have been conducted to explore the ability of ANNs in mapping nonlinear relationships of non-linear systems [3, 20, 16, 4, 5, 2, 12, 18]. However, the selection of an optimal network structure (layers and nodes) and training algorithms still remains a difficult issue in ANNs applications [9].

Recently, the group method of data handling (GMDH) algorithm has been successfully used to deal with uncertainty, linear or nonlinearity of systems in a wide range of disciplines such as economy, ecology, medical diagnostics, signal processing and control systems [14, 8]. Some simplified approximations, such as the two-direction regressive GMDH [17] and the revised GMDH algorithms [1] have been introduced to model dynamic systems in flood forecast and petroleum resource prediction with some success.

In this paper, we investigate the applicability and capability of the GMDH compared with the ANN methods for modeling of rice yields time-series forecasting. To verify the application of this approach, the rice yields data form 27 stations in Peninsular Malaysia is chosen as the case study.

2 THE NEURAL NETWORK FORECASTING MODEL

The ANN with single hidden layer feedforward network is the most widely used model for modeling and forecasting. The model is characterized by a network of three layers of simple processing units connected by a cycle links. The relationship between the input observations $(y_{t-1}, y_{t-2}, \ldots, y_{t-p})$ and the output value (y_t) has following:

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

where a_j ($j = 0, 1, 2, \ldots, q$) is a bias on the jth unit, and w_{ij} ($i = 0, 1, 2, \ldots, p; j = 0, 1, 2, \ldots, q$) is the connection weights between layers of the model, $f(\bullet)$ is the transfer function of the hidden layer, p is the number of input nodes and q is the number of hidden nodes (Lai et al. 2006).

Training a network is an essential factor for the success of the neural networks. Among the several learning algorithms available, back-propagation has been the most popular and most widely implemented learning algorithm for all neural network paradigms [23]. In this paper algorithm of back-propagation is used in the following experiment. A major advantage of neural networks is their ability to provide flexible nonlinear mapping between inputs and outputs. They can capture the nonlinear characteristics of time series well.

3. The Group Method of Data Handling Model

The algorithm of Group Method of Data Handling (GMDH) was first proposed by Madala and Ivakhnenko [8] to produce mathematical models of complex systems by handling data samples of observations. The GMDH method was originally formulated to solve for higher order regression polynomials specially for solving modeling and classification problem. General connection between inputs and output variables can be expressed by a complicated polynomial series in the form of the Volterra series, known as the Kolmogorov-Gabor polynomial:

$$y_n = a_0 + \sum_{i=1}^{M} a_i x_i + \sum_{i=1}^{M}\sum_{j=1}^{M} a_{ij} x_i x_j + \sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{M} a_{ijk} x_i x_j x_k + \ldots \qquad (3)$$

In this case, x represents the input to the system, M is the number of inputs and a are coefficients or weights.

The structure of the GMDH algorithm is illustrated in Figure 1. The computation process comprises three basic steps:

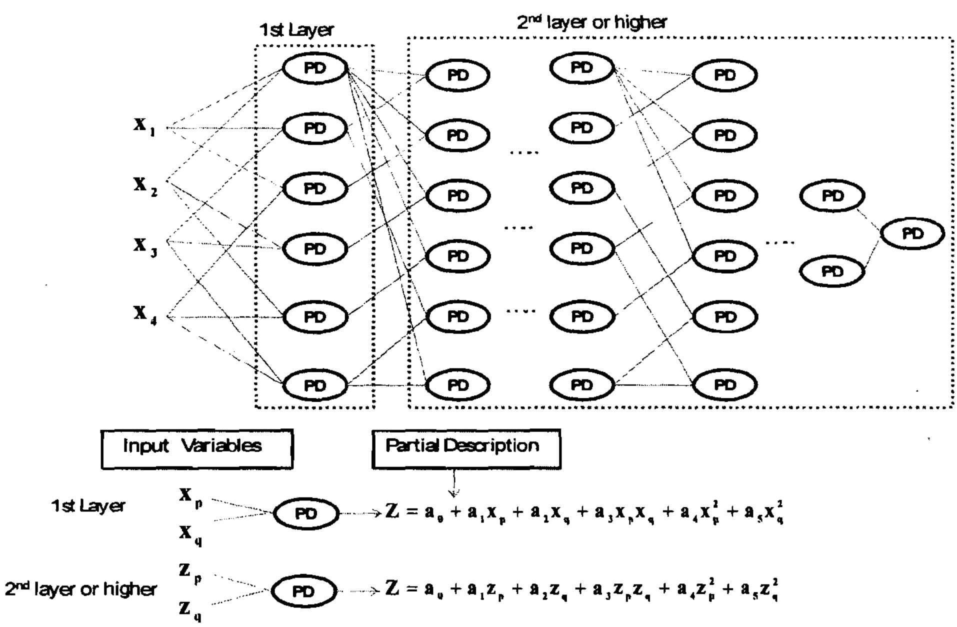

Figure 1: Basic Structure of GMDH

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. The candidate models for first layer have the form:

$$z = a_0 + a_1 x_p + a_2 x_q + a_3 x_p x_q + a_4 x_p^2 + a_5 x_q^2$$

To obtain the value of the coefficients a for each m models, a system of Gauss normal equations is solved. The coefficient a_i of nodes in each layer are expressed in the form

$$\mathbf{A}_i = (\mathbf{X}_i^T \mathbf{X}_i)^{-1} \mathbf{X}_i^T Z$$

where $\mathbf{Z} = [z_1 \ z_2 ... z_{tr}]^T$, $A_i = [a_0 \ a_1 \ a_3 ... a_5]^T$, $\mathbf{X}_i = [x_{1i} \ x_{2i}, ... x_{ki}, ... x_{tri}]^T$ and tr is the number of observations in the training set.

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 of the same form:

$$Z = a_0 + a_1 z_p + a_2 z_q + a_3 z_p z_q + a_4 z_p^2 + a_5 z_q^2$$

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

$$\text{MSE} = \frac{1}{nv} \sum_{i=ntr+1}^{n} (y_i - z_i)^2$$

where nv is the number of observations in the testing set, n is the total number of observation, z is the estimated output value and s is the model whose fitness is evaluated. Once the single best neuron, x' is 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 $\{x_1, x_2, \ldots, x_m, x'\}$, $M' = M' + 1$ and repeat steps 2 and 3. Otherwise the iterations terminate and a realization of the network has been completed.

The performances of the each model for both the training data and forecasting data are evaluated and is selected 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

$$\text{MAE} = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{y_t - z_t}{y_t}\right|$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(y_t - z_t)^2}$$

where y_t and z_t are the observed and the forecasted rice yields 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. EMPIRICAL RESULTS

In our study, the data were collected from Muda Agricultural Development Authority (MUDA) Kedah, Malaysia ranging from 1995 to 2001 is used to validate the GMDHM algorithm for rice yields modeling. The results are compared with those the ANN. These time series come from different location and have different statistical characteristics. The rice yields data contains the yields data from 1995 to 2001, giving a total of 351 observations. Given a set of 351 observations made at uniformly spaced time intervals, the locations of rice yield are rescaled to the time axis becomes the set of integers $\{1, 2, \ldots, 432\}$. For example the first location in 1995 is written as time 1, the second location in 1995 as time 2 and so on. The time series plot is given in Figure 2.

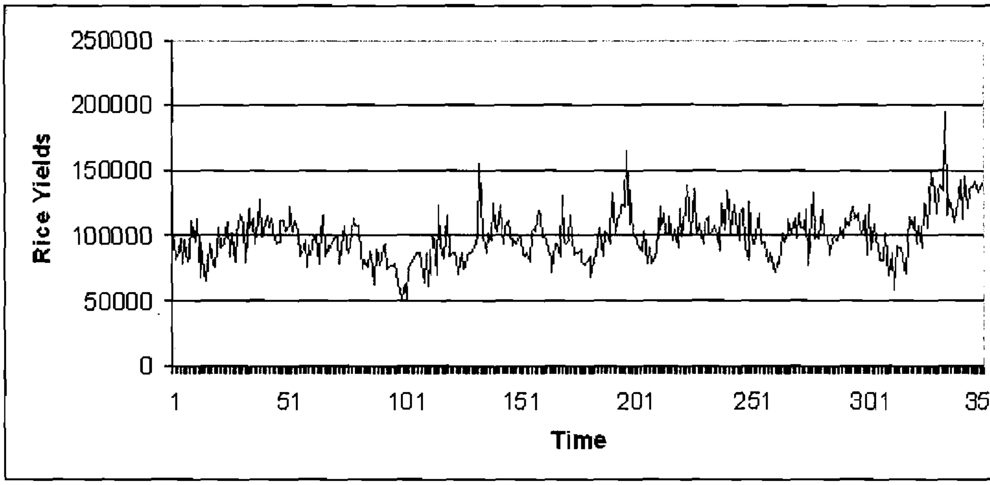

Figure 2. Rice Yields Series (1995-2001)

To assess the forecasting performance of different models, each data set is divided into two samples. The first series was used for training the network (modeling the time series) and the remaining were used for testing the performance of the trained network (forecasting). We take the data from 1995 to 2001 producing 351 observations for training purpose and the remainder as the output sample data set with 27 observations for forecasting purpose.

5 FITTING NEURAL NETWORK MODELS TO THE DATA

In this investigation, we only consider the situation of one-step-ahead forecasting with 27 observations. Before the training process begins, data normalization is often performed. The linear transformation formula to [0, 1] is used

$$x_n = \frac{y_0}{y_{max}}$$

where x_n and y_0 represent the normalized and original data; and y_{max} represent the maximum values among the original data. In order to conform the neural network used in the forecast, autocorrelation function (ACF) and partial autocorrelation function (PACF) were used to determine the maximum number of input neurons used during the training (Cadenas

& Rivera, 2007). Figure 3 presents the ACF and PACF of data sets for the rice yields time series.

Based on these analyses, the maximum number of lags, 27, was identified suitable to use as inputs for the proposed ANN. The one only neuron in the output layer represented being modeled. All the data were normalized in the range 0 and 1. After the input and out variables were selected, the ANN architecture of 27-H-1 was explored for capturing the complex, non-linear and seasonality of rice yields data. The network was trained for 5000 epochs using the back-propagation algorithm with a learning rate of 0.001 and a momentum coefficient of 0.9. Table 2 shows the performance of ANN during training with varying the number of neurons in the hidden layer (H).

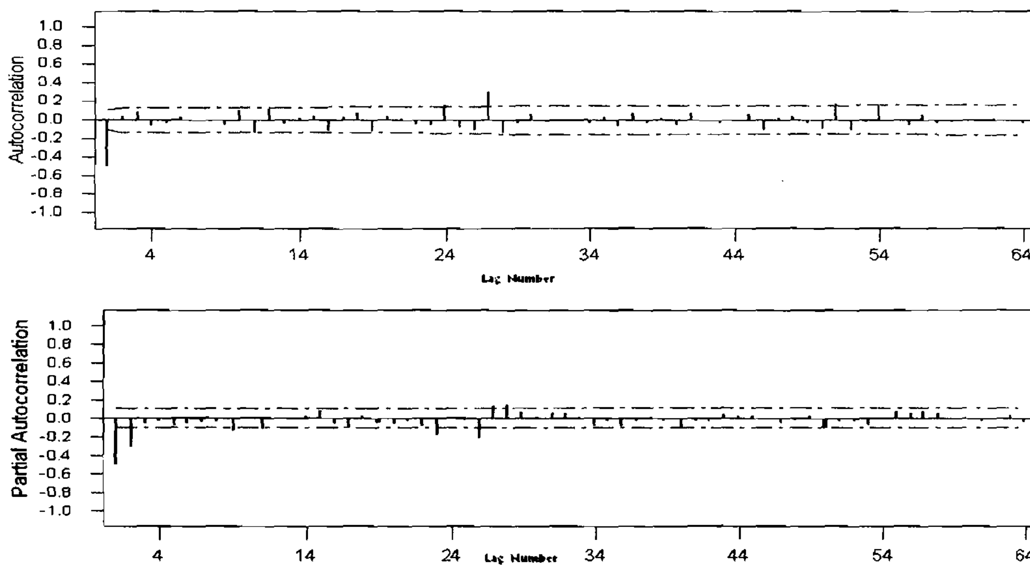

Figure 3: The ACF and PACF for the data study

Table 2 Performance Variation of a Three-Layer ANN during training with the number of neurons in the hidden layer for ANN

Criterion	Number of neurons in the hidden layer											
	3	9	15	21	27	33	39	45	51	57	63	70
RMSE	16093	15733	15043	14263	14197	13794	13768	12863	13199	13334	**12590**	12791
MAE	0.118	0.129	0.115	0.114	0.114	0.113	0.106	0.099	0.102	0.103	**0.097**	0.101

It is observed that the performance of ANN is improved as the number of hidden neurons increases. However, too many neurons in the hidden layer may cause over-fitting problem, which results in the network can learn and memorize the data very well, but lacks the ability to generalize. If the number of neurons in hidden layer is not enough then the network may not be able to learn. So, an ANN with 63 neurons in the hidden layer seems to be appropriate.

6 FITTING GMDH MODELS TO THE DATA

In designing the GMDH model, one must determine the following variables: the number of input nodes, the number of hidden layers and the number of output. The selection of the number of input corresponds to the number of variables play important roles for many successful applications of GMDH. The issue of determining the optimal number of input nodes is a crucial yet complicated one. There is no theory that can used to guide the selection the number of input.

In this study, the ACF and PACF are used also to select the number of input nodes. Based on these analyses, the maximum number of lags, 27, was identified suitable to use as inputs for the proposed GMDHM. The input pattern was assigned as $x(t-1), x(t-2),..., x(t-27)$ and thus the output pattern is:

$$y(t) = f(x(t-1), x(t-2),..., x(t-27))$$

The values predicted by the GMDH were compared with the ANN model. Table 1 shows the comparison of modeling/forecasting precision among the two approaches based on two statistical measures. In Table 1, the lowest RMSE and MAE are found with the ANN in the modeling while the GMDH in forecasting. The result, demonstrating that the GMDH provides a better approach for forecasting rice yields data.

Table 1 Comparison of modeling and forecasting precision among the four algorithms

Algorithms		ANN	GMDHZ
Modelling	RMSE	12540.19	13431.18
	MAE	0.0989	0.1038
Forecasting	RMSE	16245.206	**7549.243**
	MAE	0.1071	0.0512

7 CONCLUSION

This study investigated the applicability and capability of the GMDH model in rice yields forecasting. The performances of the GMDH model and observations were compared and evaluated based on their performance in the training and testing sets. The ANN models was also investigated for the same set of data and the results are reported. Based on the performance of two models, it can be concluded that GMDH is an effective method to forecast rice yields while the ANN method is superior to the GMDH method in the modeling of time-series. The results show that the combine the ANN and GMDH can be applied successfully to establish time-series forecasting models, which could provide accurate forecasting and modeling of time-series.

ACKNOWLEDGEMENTS

Special thanks to Ministry of Science, Technology and Innovation (MOSTI) for funding this research and MADA for their contribution of data.

REFERENCES

[1] Chang, F. J. and Hwang, Y.Y. A self-organization algorithm for real-time flood forecast, *Hydrological processes,* 13, 123-138.

[2] Coulibaly, P., Anctil, F., Aravena, R. and Bobee, B., (2001), Artificial Neural Networks modelling of water table depth fluctuations, *Water Resources Research,* 37(4), 885-896.

[3] Hsu, K.L., Gupta, H.V. and Sorooshian, S., (1995), Artificial Neural Networks modelling of rainfall-runoff process, *Water Resources Research,* 31(10), 2517- 2530.

[4] Imrie, C.E., Dcan, S. and Korre, A., (2000), River flow using Artificial Neural Networksuru: Generalization beyond the calibration range, *Journal of Hydrology,*233, 138-153.

[5] Islam, S. and Kothart, R, (2000), Artificial Neural Networks in Remote Sensing of Hydrologic Processes, *Journal of Hydrologic Engineering*, 5(2), 138-144.

[6] Ivakheneko A G. and Ivakheneko, G.A. (1995), The review of problems solvable by algorithms of the GMDH, *Pattern Recognition and Image Analysis*, 5(4): 527- 535.

[7] Khush, G.S., "Harnessing Science and Technology For Sustainable Rice Based Production Systems", Conf. *on Food and Agricultural Organization Rice Conference*, 2004.

[8] Madala, H.R. and Ivakhnenko, A.G. (1974). *Inductive Learning Algorithms for Complex System Modeling*. Boca Raton: CRC.

[9] Maier, H.R. and Dandy, G.C., (2000), Neural Networks for the production and forecasting of water resource Environmental modelling and software variables: a review and modelling issues and application, , 15, 101-124.

[10] Mark S. Voss and James C. Howland III. (2003). Financial Modelling using Social Programming. FEA 2003: *Financial Engineering and Applications*

[11] Mohammadi K, Eslami HR, Kahawita R. 2006. Parameter estimation of an ARMA model for river flow forecasting using goal programming. Journal of Hydrology 331: 293–299.

[12] Moradkhani, H, Hsu, K.L. Gupta, H.V. and Sorooshian, S, (2004). Improved streamflow forecasting using self-organizing radial basis function Artificial Neural Networks, *Journal of Hydrology*, 295, 246-262.

[13] Tamura, H. and Kondo, T. (1980). Heuristics Free Group Method of Data Handling Algorithm of Generating Optimal Partial Polynomials with Application to Air Pollution Prediction. *Int. J. System Sci.* 11:1095-1111.

[14] Tamura, H., and Kondo, T. (1980), Heuristics free group method of data handling algorithm of generating optimal partial polynomials with application to air pollution prediction, *Int. J. Syst. Sci.*, 11, 1095-1111.

[15] Thirumalaiah, K. and Deo, M.C., (2000), Hydrological Forecasting Using Neural Networks, *Journal of Hydrologic Engineering*, 5(2), 181-189.

[16] Tokar, A.S. and Markus, M., (2000), Precipitation-Runoff Modelling Using Artificial Neural Networks and Conceptual Models, *Journal of Hydrologic Engineering*, 5(2), 156-161.

[17] Tong, X.H., Kuang, J.C., Wang X.Y. and Qi, T.X., (1996), Setting up prediction model of gas well prediction rate by various methods. *Natural Gas Industry*, 16(6), 49-53.

[18] Wu Jr. S., Han J, Annambhotla S. and Bryant, S., (2005), Artificial Neural Networks for Forecasting Watershed Runoff and Stream Flows, *Journal of Hydrologic Engineering*, 5(2), 216-222.

[19] Yim, J. and Mitchell,H. Comparison of country risk models: hybrid neural networks, logit models, discriminant analysis and cluster techniques, *Expert Systems with Applications* **28** (2005) (1), pp. 137–148.

[20] Zealand, C.M., Burn, D.H. and Simonovic, S.P., (1999), Short term stream flow forecasting using Artificial Neural Networks, *Journal of Hydrology*, 214, 32-48.

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

[22] 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.