# Fuzzy GMDH-type Model and Its Application to Financial Demand Forecasting for the Educational Expenses

**Heung Suk Hwang[1*], Miyoung Seo[2],**
[1*] Professor, College of Port & Logistics, Tongmyong University,
505, Yondang-Dong, Nam-Gu, Busan, Korea
Tel: 051-620-3486
e-mail: hshwang@tu.ac.kr

[2] Professor, Department of Applied Foreign languages,
No. 1 Kainan Rd., Lu-jhu, Taoyuan, 338, Taiwan
Tel: +886-3-3412500 ext. 6096,
e-mail: seomy@mail.knu.edu.tw

## Abstract

In this paper, we developed the fuzzy group method data handling-type (GMDH) Model and applied it to demand forecasting of educational expenses. At present, GMDH family of modeling algorithms discovers the structure of empirical models and it gives only the way to get the most accurate identification and demand forecasts in case of noised and short input sampling. In distinction to fuzzy system, the results are explicit mathematical models, obtained in a relative short time. In this paper, an adaptive learning network is proposed as a kind of fuzzy GMDH. The proposed method can be reinterpreted as a multi-stage fuzzy decision rule which is called as the fuzzy GMDH. The fuzzy GMDH-type networks have several advantages compared with conventional multi-layered GMDH models. Therefore, many types of nonlinear systems can be automatically modeled by using the fuzzy GMDH. A computer program is developed and successful applications are shown in the field of demand forecasting problem of educational expenses with the number of factors considered.

**Keywords:** GMDH, Fuzzy GMDH, Adaptive Network Analysis

## 1. Introduction

Recently, the demand forecasting problem of educational expenses with its unreliable factors is one of the special research issues. In this study, we focused on the based on the data of house incomes and number children to support educational. Generally, educational expenses involve a considerable capital investment and a very complex control problem has to be solved to ensure the proper support for educational expenses needed. Mathematical models, in which many input variables are involved, require a wide range of input and output data since the number of parameters increases with the input variables. We use group method of data handling method, GMDH by Ivakhnenko (1971), to obtain the optimal combination of input variables of non-linear models. GMDH is a method of identifying nonlinear systems with many input variables and its mathematical form is represented by a hierarchical network of partial descriptions. For the purpose of identification of a mathematical model that has many input variables and limited data needs, we used a fuzzy GMDH model which is a kind of adaptive learning network in the hierarchical structure. Also this method has the ability of self-organizing a number of layers and the ability of self-selecting useful input variables. The useless input variables are eliminated and useful input variables are selected automatically. Because of this feature, it is very easy to apply this method to the identification problems of practical complex systems. Also, we developed a computer program for fuzzy GMDH algorithm, and applied it to an example demand forecasting problem of educational expenses. It is shown that the fuzzy GMDH model could be applied easily and that it is known to be a useful method for the complicated problems.

## 2. Heuristic Self-organization Method, GMDH

The architectures of the fuzzy GMDH are organized automatically by using the heuristic self-organization method (Farlow, 1988) which is used in the GMDH algorithm (Ivakhnenko, , 1968, 1971, 1995). The GMDH algorithm was developed first by Ivakhnenko (1988) for the sorting methods for modeling and clustering and we can well understand GMDH algorithm by a survey of GMDH papers Ivakhnenko (1988, 1994). GMDH algorithm is implemented according the following steps:

**Step 1**: Data collection and dividing data into two sets, training set and testing set. The training data are used for the estimation of the weights of neural networks and the testing data are used for organizing the network architectures.

**Step 2**: Construction of new variables:

In this step we take all the independent variables one or two at a time and construct the partial descriptions. The output of selected partial descriptions is treated as the input in the next layer. These steps are repeated until a termination criterion is satisfied.

**Step 3**: Rating the results of estimated dependent variable, $Z_{ij}$, by a rate criterion using only checking data.

$$r_j^2 = \frac{\sum_{i=nt+1}^{n} (y_i - z_{ij})^2}{\sum_{i=nt+1} y_i^2}, j = 1, 2, \ldots, \binom{n}{2} \qquad (1)$$

Where $y_i$ is input data in i$^{th}$ data set and $r_j$ is the rating index of estimated j$^{th}$ variable using number of testing data n. It $r_j$ < R, the new variables are passed to the next level of algorithm, where R is a predetermined value.

**Step 4**: Testing for optimality:.

If RMIN, the smallest of $r_j$, of the layer of which analysis value is greater than that of the previous design, then optimal Ivankhnenko polynomial is obtained by the following equation.

$$y_0 = a + \sum_{i=1}^{m} b_i x_i + \sum_{i=1}^{m}\sum_{j=1}^{m} c_{ij} x_i x_j +, \ldots, \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} a_{ij} x_i x_j x_k \qquad (2)$$

Where, $x_j$ is input variable and $y_0$ is estimation equation.

The heuristic self-organization method plays very important roles for organization of a neuro-fuzzy GMDH model.

## 3. Fuzzy-GMDH

There are a number of ways of fuzzy logic which can be used with neural networks. One of the simple ways is to use a fusilier function to pre-process or post-process data for a neural network. So far we have considered how fuzzy logic plays a role in neural networks (Zurada, 1992). The converse relationship, neural networks in fuzzy systems, is also an active area of the researches. In this paper, we used fuzzy networks to improve the conventional GMDH algorithms and fuzzy GMDH with improving in terms of identification accuracy.

### 3.1 Fuzzy Membership Function

In this paper, we used a simplified fuzzy reasoning rule (Takashi, *et al.*, 1998) which is given by:

If $x_1$ is given by $F_{k1}$ and $x_2$ is given by $F_{k2}$, then, output $y$ is given by $w_k$. We used Gaussian membership function, $F_{kj}$ of $k^{th}$ fuzzy rule in the domain of the $j^{th}$ input values $x_j$ is given by

$$F_{kj}(x_j) = \exp(-(x_j - a_{kj})^2 / b_{kj})$$

the parameters $a_{kj}$ and $b_{kj}$ are given for each rule,

the output $y$ is given by $y = \sum u_k w_k$,

where $u_k = \prod_j F_{kj}(x_j)$ and $w_k$ is a real number of the concluding part of the $k^{th}$ rule.

This simplified fuzzy reasoning model is used as the partial description of GMDH type adaptive learning network which is called neuro-fuzzy GMDH (Moody, & Darken, 1989).

### 3.2 Neuro-fuzzy GMDH-type Network Model

The neuro-fuzzy GMDH is an adaptive learning network (network-type of GMDH) in the hierarchical structure (Takashi, *et al.*, 1998). Figure 1 shows an example structure of neuro-fuzzy GMDH, and shows the outputs from each partial description in each layer becomes the input variables in the next layer, respectively, such as the M$^{th}$ model in the P$^{th}$ layer are the output variables of the (M-1)$^{th}$ and the M$^{th}$ models in the (P-1)$^{th}$ layer.

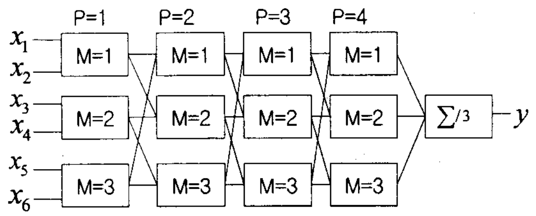


Figure 1. Sample the structure of Neuro-fuzzy GMDH

The fuzzy GMDH model can be constructed by the following five procedures (Poggio, & Girosi, 1990):, (Kondo (1997, 1998).

1) Normalizing the data: Normalizing the input and output data into intervals [0, 1]
2) Separating the original data into training and testing data
   The original data are separated into training data and testing data by "leave one out" cross validation method.
3) Generating the optimal partial descriptions: According to the following procedure, each description is generated from the 1st layer upward, and corresponding output values are obtained. The $M^{th}$ model in the $P^{th}$ layer is the input values of the $(M-1)^{th}$

$$y^{PM} = f(y^{P-1,M-1}, y^{P-1,M}) = \sum_k \mu_k^{PM} w_k^{PM} \qquad (3)$$

Where, $y^{PM}$ is the output variable of the $M^{th}$ model in the $P^{th}$ layer,

$\mu_k^{PM}$ and $w_k^{PM}$ are the degree of compatibility of the premise part with real number of conclusion part respectively which was defined in section 3.1.

Let $y^*$ be the observed value, then the performance index of the error of the models

$$Error = (y^* - y)^2 / y \tag{4}$$

where $y^*$ is observed value, and $y$ is estimated value.

4) Criteria of accuracy: Using the testing data the mean square error between observed value $y^*$ and the estimates $y$ is determined for the layer $P$ by,

$$\Delta_P{}^2 = \sum_{d=1}^{nch} (y_d{}^* - y_d)^2 / nch \tag{5}$$

where, $P$ is the layer number,

d is subscript denoting the testing data number,

$nch$ is the number of testing data, and

$y_d$ is the mean of output from the partial descriptions in the top layer.

5) Stopping rule:

When the errors of the testing data in each layer stop decreasing the iterative computation is terminated,

If $\Delta^2{}_P \geq \Delta^2{}_{P+1}$, then the iteration continue, and if

$\Delta^2{}_P < \Delta^2{}_{P+1}$, then, the iteration terminates and the models up to that layer are adopted

We developed a computer program for this model, and applied it in demand forecasting problem of educational expenses. The flow chart of this model is shown in Figure 2

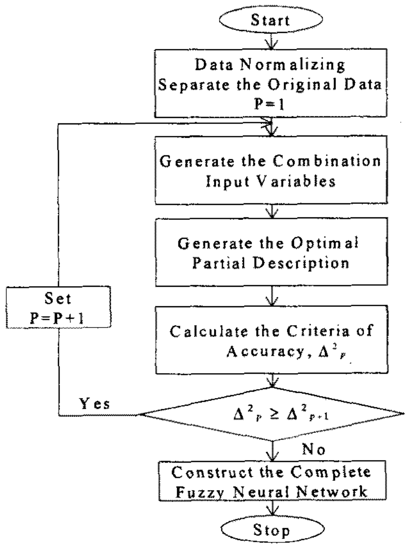


Figure 2. Flow chart of fuzzy GMDH Model

## 4. An Application to Demand Forecasting Problem of Educational Expenses

For the purpose of illustration of the fuzzy GMDH model, we apply it to the forecasting of the amount of educational expenses in Korea, and we compared the sample results of the fuzzy GMDH model with that of conventional GMDH model. Table 1 shows training data and checking data sets, among 1986 – 2004.

In Table 1, y is observed value giving the amount of educational expenses (unit $10^6$),

$x_j$, j = 1, 2, 3, 4 are independent variables:

$x_1$: amount of population (@1,000 ),

$x_2$: number of house holds (@1,000 ),

$x_3$: amount of average expenditure per house hold

$x_4$: number of student

Table 1. Training data and checking data sets
(where, TRD: training data, CHD: checking data)

| No | Year | Obs.Value ,y@$10^6$ | $x_1$ @$10^3$ | $x_2$ @$10^3$ | $x_3$ @$10^3$ | $x_4$ @$10^3$ |
|---|---|---|---|---|---|---|
| 1 | 1986 | 596,190 | 41,021 | 8,204 | 965 | 8,517 |
| 2 | 1987 | 791,775 | 41,627 | 8,325 | 1,006 | 10,557 |
| 3 | 1988 | 806,520 | 42,135 | 9,784 | 1,071 | 10,340 |
| 4 | 1989 | 846,814 | 42,721 | 9,795 | 1,093 | 10,323 |
| 5 | 1990 | 874,905 | 43,411 | 10,097 | 1,126 | 10,293 |
| 6 | 1991 | 913,140 | 43,613 | 10,903 | 1,173 | 10,146 |
| 7 | 1992 | 928,720 | 43,917 | 10,979 | 1,223 | 9,935 |
| 8 | 1993 | 958,048 | 44,232 | 11,058 | 1,353 | 9,776 |
| 9 | 1994 | 1,060,005 | 44,510 | 11,713 | 1,353 | 9,581 |
| 10 | 1995 | 1,040,820 | 44,609 | 11,760 | 1,425 | 9,462 |
| 11 | 1996 | 1,078,815 | 44,828 | 12,452 | 1,611 | 9,381 |
| 12 | 1997 | 1,129,560 | 45,991 | 12,775 | 1,692 | 9,413 |
| 13 | 1998 | 1,156,625 | 46,430 | 12,548 | 1,531 | 9,253 |
| 14 | 1999 | 1,233,225 | 46,858 | 13,016 | 1,720 | 9,135 |
| 15 | 2000 | 1,342,320 | 47,275 | 14,391 | 1,883 | 9,588 |
| 16 | 2001 | 1,429,950 | 47,692 | 14,452 | 1,906 | 9,533 |
| 17 | 2002 | 1,515,680 | 47,639 | 14,887 | 1,925 | 9,473 |
| 18 | 2003 | 1,714,440 | 47,275 | 15,225 | 2,123 | 9,508 |
| 19 | 2004 | 1,998,570 | 47,554 | 15,833 | 2,435 | 9,517 |

Source: Ministry of Education and Korea National Statistical Office

The input and output data are divided into 10 for training and 9 for testing data. It is assumed that the network is consisted of 4 layers and 3 models for each layer. Generally the estimation equation can be given by a polynomial equation, $f(x) = f(x_1, x_2, \cdots, x_k)$ where $k$ is the number of initial input variable (in this example, $k = 3$). We used the estimation equation (2).

In the first layer, we use variable $x_j$, $j = 1, 2, 3, 4$ and from the second layer the output of layer is used as input variable of next layer. Thus, the output equation can be given by equation (2).

We developed a computer program for both conventional GMDH and fuzzy GMDH method according the flow diagram as Figure 2. We used the data set for both methods and summarized the results as follows. In this example, the result of the minimum mean square error is obtained by 88,380.66 in the third layer by fuzzy GMDH using the checking data and 118,616.66 for conventional GMDH method. By this example, the minimum mean square error by fuzzy GMDH method is a little bit less than that by conventional GMDH method. The estimated values using 9 checking data by both the fuzzy GMDH and conventional GMDH method are summarized in Table 2. These values are estimated by the result of best estimated model (Ivakhnenko equation).

Figure 3 shows the mean square of errors between the estimates and the desired outputs for testing sets by both conventional GMDH and fuzzy GMDH model. For the training data set, these errors are similar with the case of checking data set but a little bit less than the case of checking data set. There are a little difference in the performance between fuzzy GMDH and conventional GMDH method, but in fuzzy GMDH model, the result gives a little better identification and forecasting accuracies, and also faster convergence than conventional GMDH.

Table 2. Observed and estimated values by both, fuzzy ($y_{fG}$) and conventional GMDH ($y_{cG}$) using checking data

| Fuzzy GMDH | | Conventional GMDH | | | | |
|---|---|---|---|---|---|---|
| No | Year | Observed Value, y | Estimated, value, $y_{fG}$ | Diff, $y - y_{fG}$ | Estimated value, $y_{cG}$ | Diff, $y - y_{cG}$ |
| 2 | 1987 | 791,775 | 791,335 | -440 | 792,234 | +459 |
| 4 | 1989 | 846,814 | 846,429 | -385 | 847,229 | +415 |
| 6 | 1991 | 913,140 | 912,908 | -232 | 913,462 | +322 |
| 8 | 1993 | 958,048 | 957,856 | -192 | 958,347 | +299 |
| 10 | 1995 | 1,040,820 | 1,040,670 | -150 | 1,040,979 | +159 |
| 12 | 1997 | 1,129,560 | 1,129,444 | -116 | 1,129,742 | +182 |
| 14 | 1999 | 1,233,225 | 1,233,013 | -212 | 1,233,502 | +277 |
| 16 | 2001 | 1,429,950 | 1,429,603 | -347 | 1,430,345 | +395 |
| 18 | 2003 | 1,714,440 | 1,714,038 | -40/.2 | 1,714,890 | +450 |

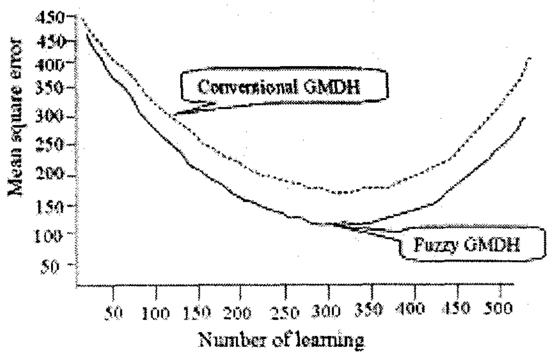


Figure 3. Comparison of accuracy of fuzzy GMDH and conventional GMDH model

Figure 4 shows the observed values and estimated values of sample problem using both conventional GMDH and fuzzy GMDH model using the data of Table 2.

By both fuzzy-GMDH and conventional GMDH method, we compute the estimates of the amount of educational expenses (in 1,000,000 units) for year 2008 and 2010, $y_{2008}$ and $y_{2010}$ as:

1) by fuzzy GMDH model: $y_{2008}$=1,971,210, and $y_{2010}$=2,034,740,
2) by conventional GMDH: $y_{2008}$=2,203,250, and $y_{2010}$ = 2,372,430.

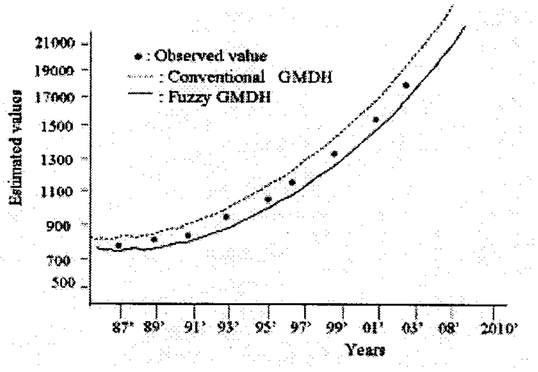


Figure 4. Observed and estimated values using checking data by both fuzzy and conventional GMDH method

The GMDH algorithm is generally implemented in a multi-layered network structured problem and enables us to construct a model containing a large number of parameters as polynomial function. In this study we proposed a fuzzy GMDH model as an improved model from conventional GMDH models. In the further study, we can compare with the other estimation models with detail data such as; Winter, ARIMAX or the other regression models.

## 5. Conclusions

Conventional regression analysis methods are constructed based on regression functions and with many assumptions. Also the conventional GMDH model is based on linear regression analysis.

In this paper, we proposed a fuzzy GMDH model which is formulated from the view point of possible model GMDH and which can automatically organize the optimal network architectures using the heuristic self-organization method. In this research, we proposed a fuzzy GMDH model and its algorithm when the simplified fuzzy reasoning is applied to practical descriptions in GMDH hierarchical structure. We applied it to demand forecasting problem of educational expenses in Korea. We used input data within a possible extends as; the amount of portion of population, amount of house holds and the amount of average expenditure per house holds. For the comparative purpose, when we compared the result of this example problem with that of conventional GMDH method, the proposed fuzzy GMDH method was known to be excellent for the complicated forecasting problems. We developed the computer program for both conventional GMDH and fuzzy GMDH model. For the further research, following researches can be performed; 1) to improve the self organizing method, we can develop a more effective self-learning engine for this model, 2) to develop a graphical user-interface programming for the users.

## References

[1] Farlow, S.J. (1984). *Self-organizing Methods in Modeling, GMDH-type Algorithms*, Marcel Dekker, Inc., New York.

[2] Ivakhnenko A.G. (1968). The group method of data handling; a rival of the method of stochastic approximation, *Soviet Automat, Control*, 13( 3), pp. 43-55.

[3] Ivakhnenko A.G. (1971). Polynomial theory complex systems, IEEE Trans. *Systems Man. Comp.*, 14(1), pp. 364-378.

[4] Ivakhnenko A.G. (1988). Sorting methods for modeling and clustering (survey of GMDH papers for the years 1983-1988) the present stage of GMDH development, *Soviet Journal of Automation and information Sciences*, 4, 1-13.

[5] Ivakhnenko A.G. (1988). The Group Method Data Handling, *a Rival to Stochastic Approximation, Soviet Automatic Control*, 13, pp. 43-55.

[6] Ivakhnenko, A.G. (1994). Heuristic self-organization in problems of engineering cybernetics, *Automatica*, 6(2), pp. 207-219.

[7 Ivakhnenko, A.G., Ivakhnenko, G.A. and Muller, J.A. ( 1994). Self-organization of the neural networks with active neurons, *Pattern Recognition and Image Analysis*, 4(2), pp. 177-188.

[8] Ivakhnenko, G.A. (1995). Self-organization on neuro-net with active neurons for effects of nuclear test explosions forecasting. *System Analysis Modelling Simulation (SAMS),* 20, pp.107-116.

[9] Kondo, T. (1998). GMDH neural network algorithm using the heuristic self-organization method and its application to the pattern identification problem, *Proc. Of 37th SICE American Conference, International Session Paper*, pp.1143-1148.

[10] Kondo, T. (1997). Medical image recognition of the lungs by using the neural network, *Bull. Sch. Med. Sci. Univ. Tokushima*, 7(2), pp.19-26.

[11] Moody, J., and Darken, C. J. (1989). Fast learning in networks of locally-tuned processing unit, *Neural Computation*, 1(1), pp.281-294.

[12] Poggio, T., and Girosi, F. (1990). Regulation algorithms for learning that are equivalent to multilayer networks, *Sciences,* 247, pp.978-982.

[13] Takashi O., Hidetomo I., Tetsuya M., and Kazunori N. (1998). Orthogonal and successive projection methods for the learning of neuro-fuzzy GMDH, *Introduction Sciences*, 110, pp. 5-24.

[14] Zurada, J.M. (1992). *Introduction to Artificial Neural Systems*, PWS, Boston, Wesley, 320.