2011 International Nuclear Atlantic Conference - INAC 2011
Belo Horizonte,MG, Brazil, October 24-28, 2011
ASSOCIAÇÃO BRASILEIRA DE ENERGIA NUCLEAR - ABEN
ISBN: 978-85-99141-04-5

VARIABLE IDENTIFICATION IN GROUP METHOD OF DATA HANDLING METHODOLOGY

Iraci Martinez Pereira[1] and Elaine Inácio Bueno[2]

[1] Instituto de Pesquisas Energéticas e Nucleares (IPEN / CNEN - SP)
Av. Professor Lineu Prestes 2242
05508-000 São Paulo, SP
martinez@ipen.br

[2] Instituto Federal de Educação, Ciência e Tecnologia – Campus Guarulhos
Av. Salgado Filho, 3501
07115-000 Guarulhos, SP
elainebueno@gmail.com

ABSTRACT

The Group Method of Data Handling - GMDH is a combinatorial multi-layer algorithm in which a network of layers and nodes is generated using a number of inputs from the data stream being evaluated. The GMDH network topology has been traditionally determined using a layer by layer pruning process based on a pre-selected criterion of what constitutes the best nodes at each level. The traditional GMDH method is based on an underlying assumption that the data can be modeled by using an approximation of the Volterra Series or Kolmorgorov-Gabor polynomial. A Monitoring and Diagnosis System was developed based on GMDH and Artificial Neural Network – ANN methodologies, and applied to the Ipen research Reactor IEA-R1. The GMDH was used to study the best set of variables to be used to train an ANN, resulting in a best monitoring variable estimative. The system performs the monitoring by comparing these estimative calculated values with measured ones. The Ipen Reactor Data Acquisition System is composed of 58 variables (process and nuclear variables). As the GMDH is a self-organizing methodology, the input variables choice is made automatically, and the real input variables used in the Monitoring and Diagnosis System were not showed in the final result. This work presents a study of variable identification of GMDH methodology by means of an algorithm that works in parallel with the GMDH algorithm and traces the initial variables paths, resulting in an identification of the variables that composes the best Monitoring and Diagnosis Model.

1. INTRODUCTION

The Group Method of Data Handling - GMDH method is composed by an algorithm proposed by Ivakhnenko [1]. The methodology can be considered as a self-organizing algorithm of inductive propagation applied at the solution of many complex practical problems. Moreover, it is possible to get a mathematical model of the process from observation of data samples, which will be used in identification and pattern recognition or even though to describe the process itself.

In applications using ANN, there are a lot of concerns due to the appropriate variable input selection. In a Nuclear Power Plant control room, there are hundreds of monitored variables which indicate the plant status operation. Thus, the correct variable selection is important to choose the minor possible variable numbers that contain the necessary information to the plant monitoring using ANN. Sometimes, it is necessary to use specialist knowledge to do the

appropriate variables input selection, or perform so many tests with different combinations of previously variables until an excellent result will be reached. Because of this, it will be very interesting to have an input automatic selection method which will be used in ANN without using the specialist knowledge. The results obtained will be ANN with a minor number of input variables, a faster training time and to discard the use of specialist knowledge to do this work.

A Monitoring and Diagnosis System was developed based on GMDH and ANN methodologies, and applied to the Ipen research Reactor IEA-R1 [2]. The GMDH was used to study the best set of variables to be used to train an ANN, resulting in a best monitoring variable estimative. The system performs the monitoring by comparing these estimative calculated values with measured ones. The Ipen Reactor Data Acquisition System is composed of 58 variables (process and nuclear variables). As the GMDH is a self-organizing methodology, the input variables choice is made automatically, and the real input variables used in the Monitoring and Diagnosis System were not showed in the final result.

This work presents a study of variable identification of GMDH methodology by means of an algorithm that woks in parallel with the GMDH algorithm and traces the initial variables paths, resulting in an identification of the variables that composes the best Monitoring and Diagnosis Model

2. GROUP METHOD OF DATA HANDLING - GMDH

The network constructed using the GMDH algorithm is an adaptive, supervised learning model. The architecture of a polynomial network is formed during the training process. The node activation function is based on elementary polynomials of arbitrary order. This kind of networks is shown in Figure 1.

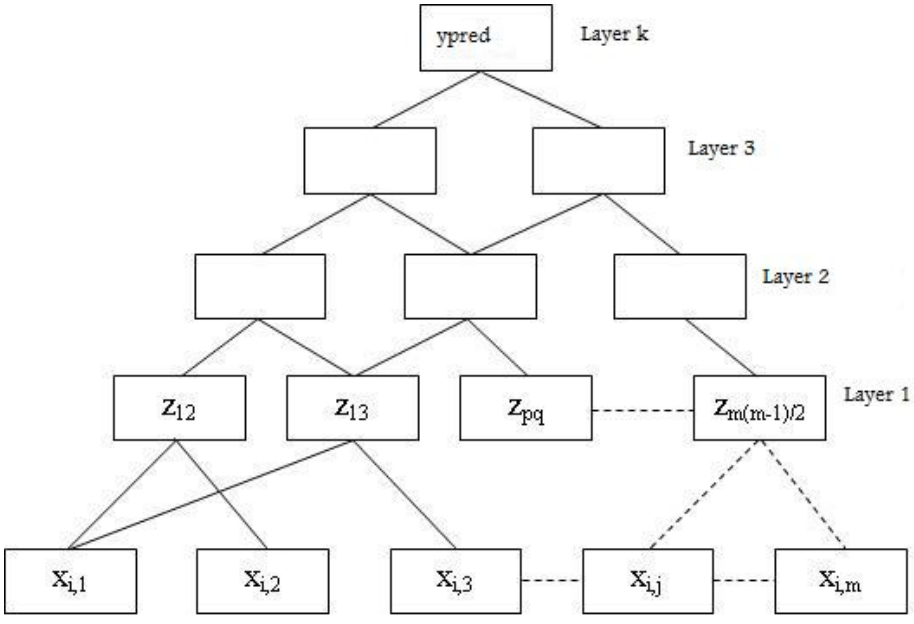

Figure 1. Self-organizing GMDH structure with m inputs and k layers.

This method solves the multidimensional problem of model improvement by the choice procedure and selection of models chosen from a set of candidate models in accordance with a supplied criterion. The majority GMDH algorithms use reference polynomial functions. A generic connection between inputs and outputs can be expressed by the series functions of Volterra which is the discrete analogous of the polynomial of Kolmogorov-Gabor [5], equation (1):

$$y = a + \sum_{i=1}^{m} b_i x_i + \sum_{i=1}^{m}\sum_{j=1}^{m} c_{ij} x_i x_j + \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} d_{ijk} x_i x_j x_k + \cdots \quad (1)$$

Where:

{x_1, x_2, x_3 ...}: inputs
{a, b, c...}: polynomials coefficients
y: the node output

The components of input matrix can be changeable independent, functional forms or terms of finite differences, moreover, can be used other nonlinear reference functions. The methods still allow, simultaneously finding the model structure and the output system dependence as a function of the most important inputs system values.

The following procedure is used for a given set of ***n*** observations of the ***m*** independent variables $\{x_1, x_2,\ldots, x_m\}$ and their associated matrix of dependent values $\{y_1, y_2, \ldots y_n\}$ [3] .

- Subdivide the data into two subsets: one for training and other for testing;
- Compute the regression polynomial using the equation (2), for each pair of input variables x_i and x_j and the associated output y of the training set which best fits the dependent observations y in the training set. From the observations, m(m-1)/2 regression polynomials will be computed from the observations;

  - $y = A + Bx_i + Cx_j + Dx_i^2 + Ex_j^2 + Fx_i x_j$ (2)

- Evaluate the polynomial for all ***n*** observations for each regression. Store these ***n*** new observations into a new matrix Z. The other columns of Z are computed in a similar manner. The Z matrix can be interpreted as new improved variables that have better predictability than those of the original generation $x_1, x_2,\ldots, x_m$;
- Screening out the last effective variables. The algorithm computes the root mean-square value (regularity criterion – r_j) over the test data set for each column of Z matrix. The regularity criterion is given by the equation (3);

$$r_j^2 = \frac{\sum_{i=1}^{nt} \left(y_i - z_{ij}\right)^2}{\sum_{i=1}^{nt} y_i^2} \quad (3)$$

- Order the columns of Z according to increasing r_j, and then pick those columns of Z satisfying $r_j < R$ (R is some prescribed value chosen by the user) to replace the original columns of X;
- The above process is repeated and new generations are obtained until the method starts overfitting the data set. One can plot the smallest of the r_j's computed in each generation and compare it with the smallest r_j's of the most recent generation start to have an increasing trend.

3. ARTIFICIAL NEURAL NETWORKS

An Artificial Neural Network - ANN is a massively parallel distributed processor made up of simple processing units, which has a natural propensity for storing experiential knowledge and making it available for use. The knowledge is acquired by the networks from its environment through a learning process which is basically responsible to adapt the synaptic weights to the stimulus received by the environment. The fundamental element of a neural network is a neuron, which has multiple inputs and a single output, Figure 2. It is possible to identify three basic elements in a neuron: a set of synapses, where a signal x_j at the input of synapse j connected to the neuron k is multiplied by the synaptic weight w_{kj}, an adder for summing the input signals, weighted by the respective synapses of the neuron; and an activation function for limiting the amplitude of the output of a neuron. The neuron also includes an externally applied *bias*, denoted by b_k, which has the effect of increasing or lowering the net input of the activation function, depending on whether it is positive or negative, respectively [4].

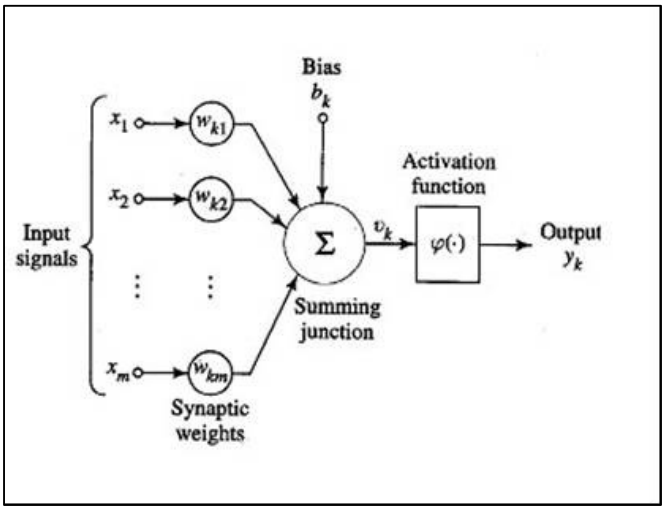

Figure 2. Neuron Model.

In this work, it was used the MLP (Multilayer Perceptron Neural Network). In this kind of architecture, all neural signals propagate in the forward direction through each network layer from the input to the output layer. Every neuron in a layer receives its inputs from the neurons in its precedent layer and sends its output to the neurons in its subsequent layer. The training is performed using an error backpropagation algorithm, which involves a set of connecting weights, which are modified on the basis of a Gradient Descent Method [4] to

minimize the difference between the desired output values and the output signals produced by the network, as show the equation (4):

$$E = \frac{1}{2}\sum_{m=1}^{m}\left(y_{dj}(n) - y_j(n)\right)^2 \qquad (4)$$

Where:

E: mean squared error
m: number of neurons in the output layer
y_{dj}: target output
y_j: actual output
n: number of interactions

4. VARIABLE IDENTIFICATION IN GMDH METHODOLOGY

The variable identification in GMDH Methodology will be described considering a set of n observations of four input variables x1, x2, x3 e x4 (matrix X), and the corresponding set of output variables Y.

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & x_{n3} & x_{n4} \end{bmatrix} \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} \qquad (5)$$

Layer 1:

As described previously, the first step consists of compute the regression polynomial using the equation (2), for each pair of input variables x_i and x_j and the associated output y of the training set which best fits the dependent observations y in the training set.

$$y = A + Bx_i + Cx_j + Dx_i^2 + Ex_j^2 + Fx_i x_j \qquad (2)$$

From the observations, m(m-1)/2 regression polynomials will be computed; for 4 input variables the possible number of combinations is 6.

$$C = \frac{4(4-1)}{2} \Rightarrow C = 6 \qquad (6)$$

And then we will have the following polynomials that will be evaluate for all ***n*** observations for each regression.

$$y_1 = a_{12} + b_{12}x_1 + c_{12}x_2 + d_{12}x_1^2 + e_{12}x_2^2 + f_{12}x_1x_2 \quad (7)$$

$$y_2 = a_{13} + b_{13}x_1 + c_{13}x_3 + d_{13}x_1^2 + e_{13}x_3^2 + f_{13}x_1x_3 \quad (8)$$

$$y_3 = a_{14} + b_{14}x_1 + c_{14}x_4 + d_{14}x_1^2 + e_{14}x_4^2 + f_{14}x_1x_4 \quad (9)$$

$$y_4 = a_{23} + b_{23}x_2 + c_{23}x_3 + d_{23}x_2^2 + e_{23}x_3^2 + f_{23}x_2x_3 \quad (10)$$

$$y_5 = a_{24} + b_{24}x_2 + c_{24}x_4 + d_{24}x_2^2 + e_{24}x_4^2 + f_{24}x_2x_4 \quad (11)$$

$$y_6 = a_{34} + b_{34}x_3 + c_{34}x_4 + d_{34}x_3^2 + e_{34}x_4^2 + f_{34}x_3x_4 \quad (12)$$

The calculated coefficients are then used to calculate the z values. These ***n*** new observations are stored into a new matrix Z.

$$z_{i,pq} = a_{pq} + b_{pq}x_{ip} + c_{pq}x_{iq} + d_{pq}x_{ip}^2 + e_{pq}x_{iq}^2 + f_{pq}x_{ip}x_{iq} \quad (13)$$

$$X = \begin{bmatrix} z_{11} & z_{12} & z_{13} & z_{14} & z_{15} & z_{16} \\ x_{21} & x_{22} & x_{23} & x_{24} & z_{25} & z_{26} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ z_{n1} & z_{n2} & z_{n3} & z_{n4} & z_{n5} & z_{n6} \end{bmatrix} \quad (14)$$

After that, the next step consists in screening out for the last effective variables. The algorithm computes the root mean-square value (regularity criterion – r_j) over the test data set for each column of Z matrix. The regularity criterion is given by the equation (15);

$$r_j^2 = \frac{\sum_{i=1}^{nt} \left(y_i - z_{ij}\right)^2}{\sum_{i=1}^{nt} y_i^2} \quad (15)$$

According to increasing r_j, the columns of Z are ordered and then those columns of Z satisfying $r_j < R$ (R is some prescribed value chosen by the user) are considered to replace the original columns of X. For this explanation, we will consider that the columns of matrix z are already ordered.

In order to proceed to the next layer, we have to consider the number of columns of matrix z that will not to cause a too big possible number of combinations of the input variables. For this explanation, we will consider the 4 first columns of matrix z to continue to the next layer.

Layer 2:

The same procedure will be repeated for the second layer, but it has to be noted that the matrix z columns represent the variables x_1x_2, x_1x_3, x_1x_4, x_2x_3. The polynomials of the second layer will be represented in a shot form, to avoid a too long polynomial representation:

$$y_1 = f(x_1x_2x_1x_3) \Rightarrow y_1 = f(x_1^2x_2x_3) \quad (16)$$

$$y_2 = f(x_1x_2x_1x_4) \Rightarrow y_1 = f(x_1^2x_2x_4) \quad (17)$$

$$y_3 = f(x_1x_2x_2x_3) \Rightarrow y_1 = f(x_1x_2^2x_3) \quad (18)$$

$$y_4 = f(x_1x_3x_1x_4) \Rightarrow y_1 = f(x_1^2x_3x_4) \quad (19)$$

$$y_5 = f(x_1x_3x_2x_3) \Rightarrow y_1 = f(x_1x_2x_3^2) \quad (20)$$

$$y_6 = f(x_1x_4x_2x_3) \Rightarrow y_6 = f(x_1x_2x_3x_4) \quad (21)$$

The new matrix z is calculated, the columns are ordered according to r_j, and again the 4 first columns are considered to the next layer.

Layer 3:

The same procedure is repeated to the layer 3, where the matrix columns are now composed by the variables x_1x_2 x_1x_3, x_1x_2 x_1x_4, x_1x_2 x_2x_3, $x_1x_3x_1x_4$ and we have the following polynomials:

$$y_1 = f(x_1x_2x_1x_3x_1x_2x_1x_4) \Rightarrow y_1 = f(x_1^4x_2^2x_3x_4) \quad (22)$$

$$y_2 = f(x_1x_2x_1x_3x_1x_2x_2x_3) \Rightarrow y_2 = f(x_1^3x_2^3x_3^2) \quad (23)$$

$$y_3 = f(x_1x_2x_1x_3x_1x_3x_1x_4) \Rightarrow y_3 = f(x_1^4x_2x_3^2x_4) \quad (24)$$

$$y_4 = f(x_1x_2x_1x_4x_1x_2x_2x_3) \Rightarrow y_4 = f(x_1^3x_2^3x_3x_4) \quad (25)$$

$$y_5 = f(x_1x_2x_1x_4x_1x_3x_1x_4) \Rightarrow y_5 = f(x_1^4x_2x_3x_4^2) \quad (26)$$

$$y_6 = f(x_1x_2x_2x_3x_1x_3x_1x_4) \Rightarrow y_6 = f(x_1^3x_2^2x_3^2x_4) \quad (27)$$

Layer 4:

The same procedure is repeated to the layer 4, where the matrix columns are now composed by the variables $x_1x_2x_1x_3x_1x_2x_1x_4$, $x_1x_2x_1x_3x_1x_2x_2x_3$, $x_1x_2x_1x_3x_1x_3x_1x_4$, $x_1x_2x_1x_4x_1x_2x_2x_3$ and we have the following polynomials:

$$y_1 = f(x_1x_2x_1x_3x_1x_2x_1x_4x_1x_2x_1x_3x_1x_2x_2x_3) \Rightarrow y_1 = f(x_1^7x_2^5x_3^3x_4) \quad (28)$$

$$y_2 = f(x_1x_2x_1x_3x_1x_2x_1x_4x_1x_2x_1x_3x_1x_3x_1x_4) \Rightarrow y_2 = f(x_1^8x_2^3x_3^3x_4^2) \quad (29)$$

$$y_3 = f(x_1x_2x_1x_3x_1x_2x_1x_4x_1x_2x_1x_4x_1x_2x_2x_3) \Rightarrow y_3 = f(x_1^7x_2^5x_3^2x_4^2) \quad (30)$$

$$y_4 = f\left(x_1x_2x_1x_3x_1x_2x_2x_3x_1x_2x_1x_3x_1x_3x_1x_4\right) \Rightarrow y_4 = f\left(x_1^7 x_2^4 x_3^4\right) \quad (31)$$

$$y_5 = f\left(x_1x_2x_1x_3x_1x_2x_2x_3x_1x_2x_1x_4x_1x_2x_2x_3\right) \Rightarrow y_5 = f\left(x_1^6 x_2^6 x_3^2 x_4\right) \quad (32)$$

$$y_6 = f\left(x_1x_2x_1x_3x_1x_3x_1x_4x_1x_2x_1x_4x_1x_2x_2x_3\right) \Rightarrow y_6 = f\left(x_1^7 x_2^4 x_3^3 x_4^2\right) \quad (33)$$

Assuming that the best result was reached for this layer, and the polynomials are already ordered, we now can evaluate which input variables are predominant in the resulting polynomial by means of the variable exponent. In this example, the best result is y_1 and the corresponding input variables and respective exponents are: x_1^7 x_2^5 x_3^3 and x_4. We can then consider that the variables x_1 and x_2 are the most relevant variables for the GMDH algorithm. These variables will be used as input variables in a ANN algorithm to perform the Monitoring and Diagnosis System.

5. RESULTS

The GMDH variables identification was applied in a Monitoring and Diagnosis System for the Ipen research Reactor IEA-R1. The Ipen nuclear research reactor IEA-R1 is a pool type reactor using water for the cooling and moderation functions and graphite and beryllium as reflector. Its first criticality was in September 16th, 1957. Since then, its nominal operation power is 2 MW. In 1997 a modernization process was performed to increase the power to 5 MW, in a full cycle operation time of 120 hours, in order to improve its radioisotope production capacity.

The IEA-R1 reactor Data Acquisition System DAS monitors 58 operational variables including temperature, mass flow rate, pressure, nuclear radiation, nuclear power and rod position. From these 58 variables 30 were chosen to be monitored by the Monitoring and Diagnosis System [2].

The methodology developed was applied to a GMDH model for this set of input variables. For each monitored variable, there is a set of input variables chosen by the GMDH algorithm as described in the variable identification topic. Table 1 shows the variables description and the predominant variables used in the GMDH model.

Table 1. Monitoring and Diagnosis System Variables

Monitored Variable	Variable description	Variables identified by GMDH
Z1	Control rod position	Z2 R2M3 R8M3 R11M3
Z2	Safe Rod position	Z1 Z3
Z3	Safe Rod position	Z1 Z4
Z4	Safe Rod position	Z1 Z3
N2	% power	Z1 N7
N3	% power	N2 R8M3 T1
N4	% power	Z3 N7
N6	% power	Z1 N7
N7	% demand	N6 R8M3
N8	N16 power	Z4 N4
C1	Water conductivity	R8M3 T1
C2	Water conductivity	R8M3 T1
R1M3	Nuclear dose rate (left side bridge)	Z1 R2M3 R7M3 R8M3 R11M3
R2M3	Nuclear dose rate (right side bridge)	Z1 R1M3 R8M3 R11M3 T4
R3M3	Nuclear dose rate	R1M3 R2M3 R8M3 R11M3 T7
R7M3	Nuclear dose rate	Z2 N4 N7 N8 R8M3 T8
R8M3	Nuclear dose rate	Z1 Z2 R1M3 R2M3 R11M3
R9M3	Nuclear dose rate	N3 N7 T6
R10M3	Nuclear dose rate	R11M3 R12M3
R11M3	Nuclear dose rate	Z1 Z4 R1M3 R2M3 R8M3 T1 T6
R12M3	Nuclear dose rate	Z1 R8M3 R11M3 T1 T6
R14M3	Nuclear dose rate	R8M3 R11M3 T1
T1	Pool surface temperature	N3 T2 T6
T2	Pool temperature (half high)	T1 T4
T3	Pool temperature (over the nuclear core)	T1 T6
T4	Inlet decay tank temperature	T1 T6
T6	Outlet decay tank temperature	T4 T7
T7	Outlet primary loop temperature	T6 T8
T8	Inlet secondary loop temperature	Z1 R7M3 R11M3 T6 T9
T9	Outlet secondary loop temperature	T6 T8

The results obtained were compared with a Monitoring and Diagnosis System developed using all the variables as input variables to the ANN, that means without using GMDH as input variable selection. The monitoring is performed by comparing the variable value estimated by the System with those measured by the Data Acquisition System (actual value). This comparison is computed chosen calculating the residual, as shown the equation (34) where y_{dj} are the desired output and y_j the actual output. Results are shown in Table 2 and Figure 3.

$$residual = \frac{\left| y_{dj} - y_j \right|}{y_j} \cdot 100 \qquad (34)$$

Table 2. Results for Monitoring and Diagnosis Systems

Monitored Variable	Residual(%)	
	ANN+GMDH	ANN
Z1	0.5279	0.3267
Z2	0.1053	0.2803
Z3	0.0019	0.3142
Z4	0.0036	0.2955
C1	2.0564	2.0083
C2	0.1030	0.0955
N2	0.1413	0.3044
N3	0.1642	0.2726
N4	0.1511	0.3041
N6	0.1342	0.3046
N7	0.0468	0.3041
N8	0.4293	0.6101
R1M3	2.9378	3.2171
R2M3	2.6268	2.9562
R3M3	2.8834	4.0342
R7M3	2.8065	2.8201
R8M3	1.9831	1.7784
R9M3	1.6833	1.7832
R10M3	3.1599	3.2419
R11M3	2.3390	2.3411
R12M3	5.0551	5.0919
R14M3	6.0936	6.1500
T1	0.1137	0.5015
T2	0.1049	0.5172
T3	0.1180	0.5741
T4	0.0023	0.5271
T6	0.1041	0.5243
T7	0.1436	0.5166
T8	0.2004	1.1617
T9	0.1229	0.5810

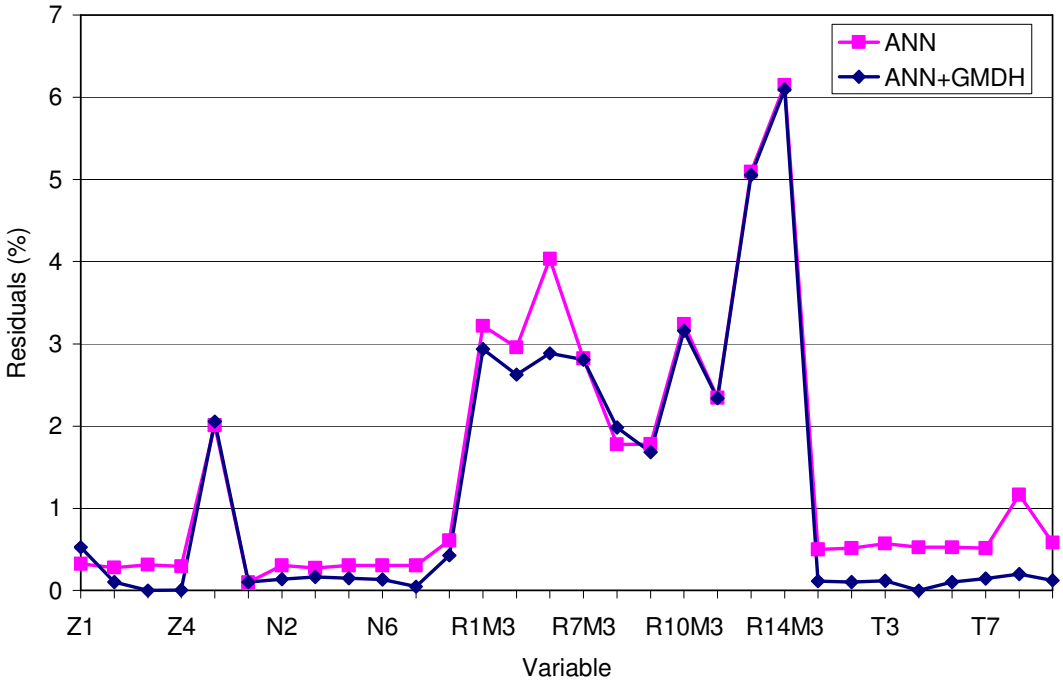

Figure 3. Results of Monitoring and Diagnosis Systems using ANN with and without GMDH input variable selection.

The residuals are smaller using GMDH for ANN input variable selection than using just ANN for 26 variables monitored and for the 4 remainder variables the residuals are of the same magnitude.

6. CONCLUSION

This work presents a study of variable identification of GMDH methodology by means of an algorithm that woks in parallel with the GMDH algorithm and traces the initial variables paths, resulting in an identification of the variables that composes the best Monitoring and Diagnosis Model.

The GMDH variables identification was applied in a Monitoring and Diagnosis System for the Ipen research Reactor IEA-R1. The IEA-R1 reactor Data Acquisition System DAS monitors 58 operational variables including temperature, mass flow rate, pressure, nuclear radiation, nuclear power and rod position. From these 58 variables 30 were chosen to be monitored by the Monitoring and Diagnosis System. The methodology developed was applied to a GMDH model for this set of input variables. For each monitored variable, there is a set of input variables chosen by the GMDH algorithm as described in the variable identification topic.

The results obtained were compared with a Monitoring and Diagnosis System developed using all the variables as input variables to the ANN, that means without using GMDH as input variable selection. The monitoring is performed by comparing the variable value estimated by the System with those measured by the Data Acquisition System (actual value). From the 30 monitored variables, for 26 variables the residuals are smaller using GMDH for ANN input variable selection than using just ANN and for the 4 remainder variables the residuals are of the same magnitude.

REFERENCES

1. Ivakhnenko, A. G. "The group method of data handling - A rival of the method of stochastic approximation". *Avtomatika,* No. 3, 1968.
2. Bueno, E. I. "Group Method of Data Handling e Redes Neurais Na Monitoração e Detecção de Falhas em Reatores Nucleares". Tese (Doutorado), Universidade de São Paulo - IPEN, 2011.
3. Farlow, S. J. *Self-organizing methods in modeling: GMDH-type algorithms*. New Your: M. Dekker, 1984.
4. Haykin, S. *Neural Networks - A Comprehensive Foundation*. USA: Prentice Hall, 1999.
5. Nelles, O. *Nonlinear system identification: from classical approaches to neural networks and fuzzy models.* Springer, 2001.