*2013 International Nuclear Atlantic Conference - INAC 2013*
*Recife, PE, Brazil, November 24-29, 2013*
*ASSOCIAÇÃO BRASILEIRA DE ENERGIA NUCLEAR - ABEN*
***ISBN:*** **978-85-99141-05-2**

# STUDY OF INPUT VARIABLES 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 ANN methodologies, and applied to the Ipen research Reactor IEA-1. The system performs the monitoring by comparing the GMDH and ANN calculated values with measured ones. As the GMDH is a self-organizing methodology, the input variables choice is made automatically. On the other hand, the results of ANN methodology are strongly dependent on which variables are used as neural network input.

## 1. INTRODUCTION

Group Method of Data Handling was applied in a great variety of areas for data mining and knowledge discovery, forecasting and systems modeling, optimization and pattern recognition. Inductive GMDH algorithms give possibility to find automatically interrelations in data, to select optimal structure of model or network and to increase the accuracy of existing algorithms [3].

This original self-organizing approach is substantially different from deductive methods used commonly for modeling. It has inductive nature − it finds the best solution by sorting-out of possible variants. By sorting of different solutions, GMDH algorithms aims to minimize the influence of the author on the results of modeling. Computer itself finds the structure of the model and the laws which act in the system.

In mathematical statistics it is need to have *a priori* information about the structure of the mathematical model. In neural networks the user estimates this structure by choosing the number of layers and the number and transfer functions of nodes of a neural network. This requires not only knowledge about the theory of neural networks, but also knowledge of the

object nature and time. Besides this the knowledge from systems theory about the systems modeled is not applicable without transformation in neural network world. But the rules of translation are unknown. These problems can be overcome by GMDH that can pick out knowledge about object directly from data sampling. The Group Method of Data Handling is the inductive sorting-out method, which has advantages in the cases of rather complex objects, having no definite theory, particularly for the objects with fuzzy characteristics.

This work presents a study of input variable of GMDH methodology. Results obtained by statistical learning networks and especially GMDH algorithms are comparable with results obtained by neural networks. The well-known problems of an optimal (subjective) choice of the neural network architecture are solved in the GMDH algorithms by means of an adaptive synthesis (objective choice) of the architecture. Such algorithms combining the best features of neural nets and statistical techniques in a powerful way discover the entire model structure directly from data sample − in the form of a network of polynomial functions, difference equations or another structure type. Models are selected automatically based on their ability to solve the task (approximation, identification, forecasting, and classification).

## 2. GROUP METHOD OF DATA HANDLING − GMDH

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.

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.

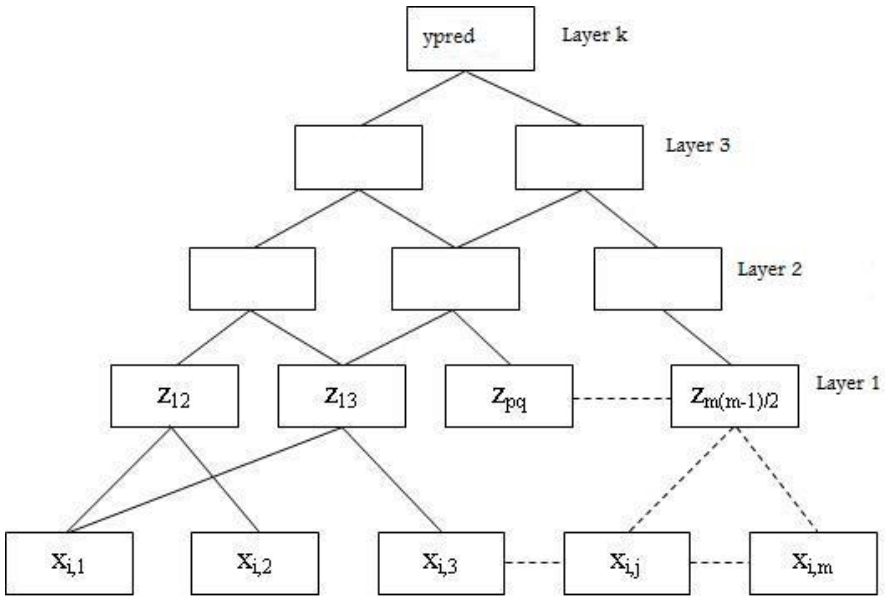


**Figure 1: Self-organizing GMDH structure with m inputs and k layers.**

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;

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

- 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 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, as we can see in Figure 2. It is possible to identify three basic elements in a neuron: a set of synapses, where a signal xj at the input of synapse j connected to the neuron k is multiplied by the synaptic weight wkj, 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 bk, 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].

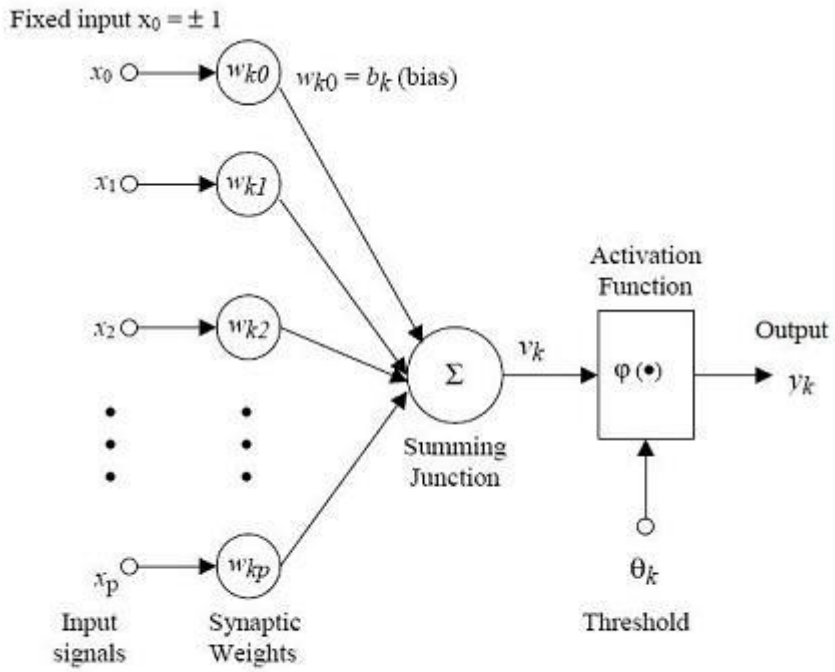


**Figure 2: Neuron Model.**

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 to minimize the difference between the desired output values and the output signals produced by the network.

## 4. IEA-R1 RESEARCH REACTOR

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 was 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. Figure 1 shows a flowchart diagram of the Ipen nuclear research reactor IEA-R1.

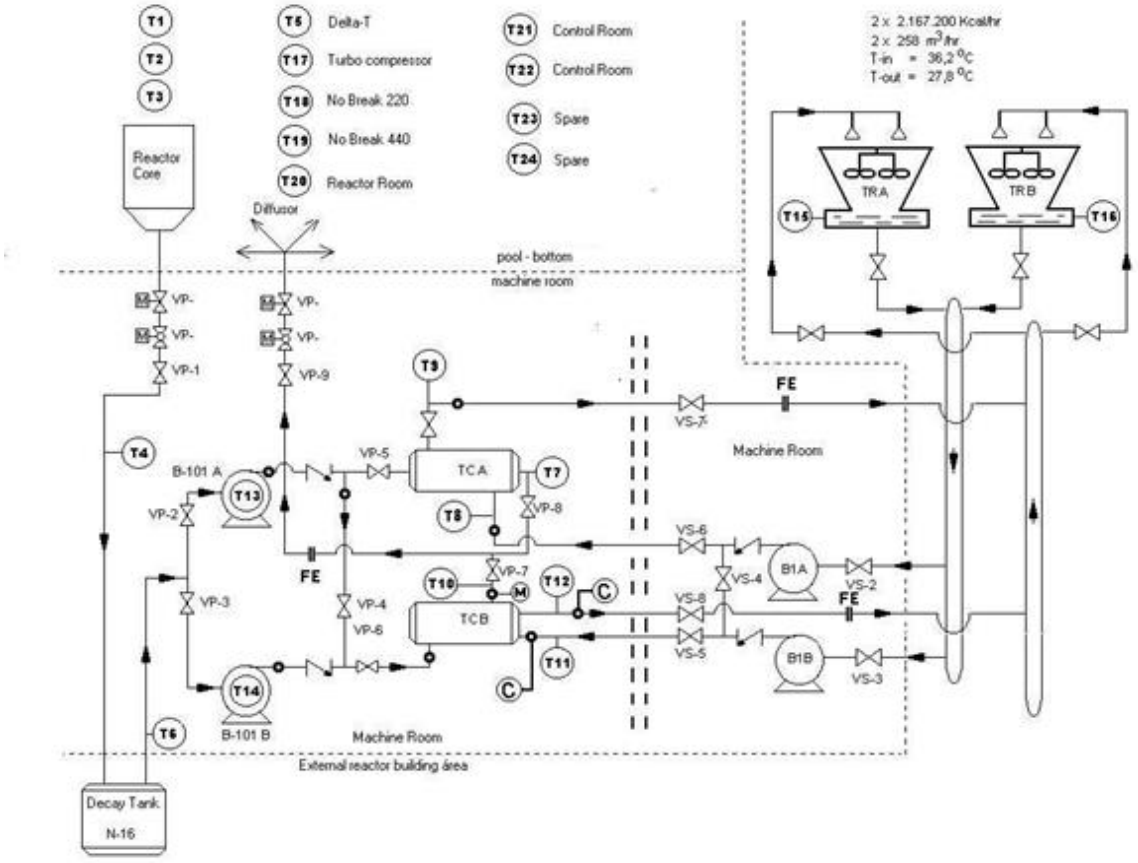


**Figure 3: IEA-R1 experimental reactor schematic diagram.**

### 4.1. IEA-R1 Data Acquisition System (DAS)

The Ipen reactor Data Acquisition System monitors 58 operational variables, including temperature, flow, level, pressure, nuclear radiation, nuclear power and rod position (Table 1). The DAS performs the storage the temporal history of all process variables monitored and does not interfere with the reactor control.

**Table 1. IEA-R1 DAS variables.**

| | |
|---|---|
| Z1 | Control rod position [0 a 1000 mm] |
| Z2-Z4 | Safety rod position 1, 2 and 3[0 a 999 mm] |
| N2-N4 | % power (safety channel 1, 2 and 3) [%] |
| N5 | Logarithm Power (log channel) [%] |
| N6-N8 | % power [%] |
| F1M3 | Primary loop flowrate [gpm] |
| F2M3 | Secondary loop flowrate [gpm] |
| C1-C2 | Pool water conductivity [µmho] |
| L1 | Pool water level [%] |
| R1M3-R14M3 | Nuclear dose rate [mR/h] |
| T1-T3 | Pool water temperature [º C] |
| T4 and T6 | Decay tank inlet and outlet temperature [º C] |
| T5 | (T4-T3) [º C] |
| T7 | Primary loop outlet temperature (heat exchanger A) [º C] |
| T8-T9 | Secondary loop inlet and outlet temperature (heat exchanger A) [º C] |
| T10 | Primary loop outlet temperature (heat exchanger B) [º C] |
| T11-T12 | Secondary loop inlet and outlet temperature (heat exchanger B) [º C] |
| T13-T14 | Housing pump B101-A and B102-A temperature [º C] |
| T15-T16 | Cooling tower A and B temperature [º C] |
| T17 | Housing turbo compressor temperature [º C] |
| T18-T19 | NO-BREAK temperature –220V and 440V [º C] |
| T20-T24 | Room temperature [º C] |

## 5. RESULTS

### 5.1. Monitoring System using Neural Networks

A Monitoring and Diagnosis System was developed using Artificial Neural Networks [2]. The neural network was trained to monitor the temperature, nuclear power and dose rate variables. It was used IEA-R1 reactor data from a typical operation week. To prevent overfitting, the method of Early Stopping was used, which database division in three subsets: training (50%), validation (25%) and testing (25%).The ANN architecture used has a Multilayer Perceptron Network with three layers: one input layer, one hidden layer and on output layer. The input layer is composed by three neurons and its activation function is linear. In the hidden layer, 10 cases was studied and tested with different number of neurons varying from 1 to 10, in order to find the best number of neurons. The activation function is the hyperbolic tangents. The output layer is composed by a neuron that represents the output of the network.

The ANN developed was used to estimate the IEA-R1 variables. The values obtained were compared with the actual measured values according to Equation 4 where: *res* is the residual, $y_{ANN}$ is the variable estimated by the neural network and $y$ is the variable value measured.

$$res = ((y_{ANN} - y)/y) * 100 \quad (4)$$

Figure 4 shows the residual % values for T3 and N2 IEA-R1 variables with number of neurons of the hidden layer varying from 1 to 10. Note that there are different values, and we do not know *a priori* which one will be the best result. The results are similar for the other IEA-R1 variables. Figure 5 illustrates that the Monitoring System result using Artificial Neural Networks strongly depends on the input variables.

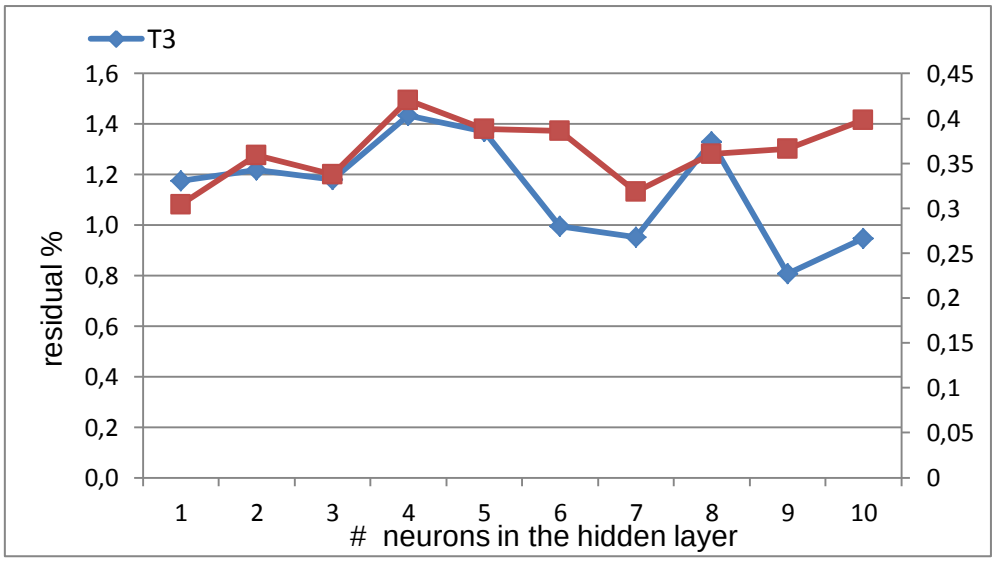


**Figure 4: Monitoring System Residual for different ANN architectures.**

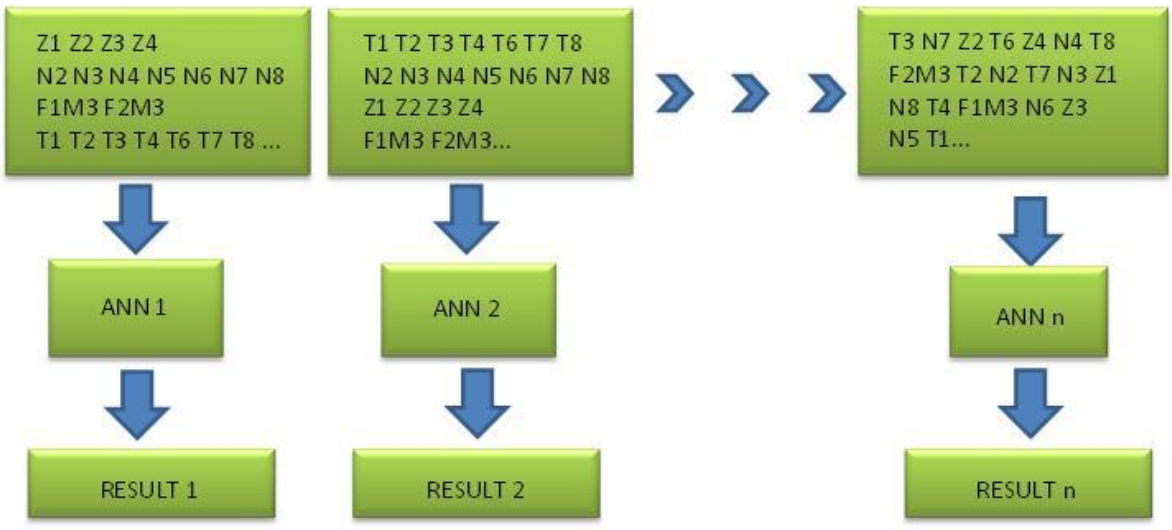


**Figure 5: Different results for different ANN architectures.**

## 5.2. GMDH Monitoring System

A Monitoring and Diagnosis System was developed for Ipen research reactor IEA-R1operation data based on the GMDH (Group Method of Data Handling) methodology [6]. The system performed the monitoring, comparing GMDH model calculated values with measured values.

For each one of the 58 SAD variables, 57 different GMDH models were developed with different architectures, varying the set of input variables. GMDH models were tested and all gave the same result. Figures 6 show the results for GMDH Monitoring System for variables T3 and N2, and Figure 7 illustrates the independence on the input variables architecture on GMDH results.

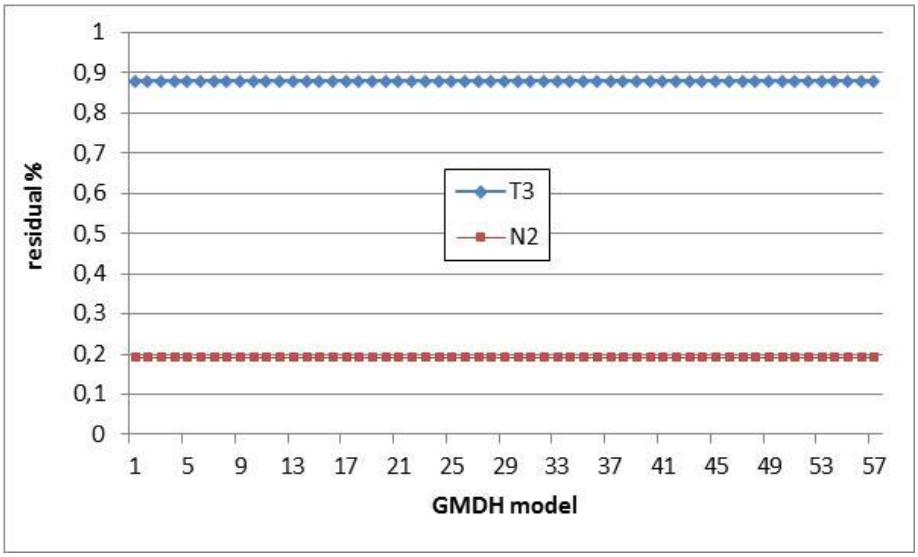


**Figure 6: Monitoring System Residual for different GMDH architectures.**

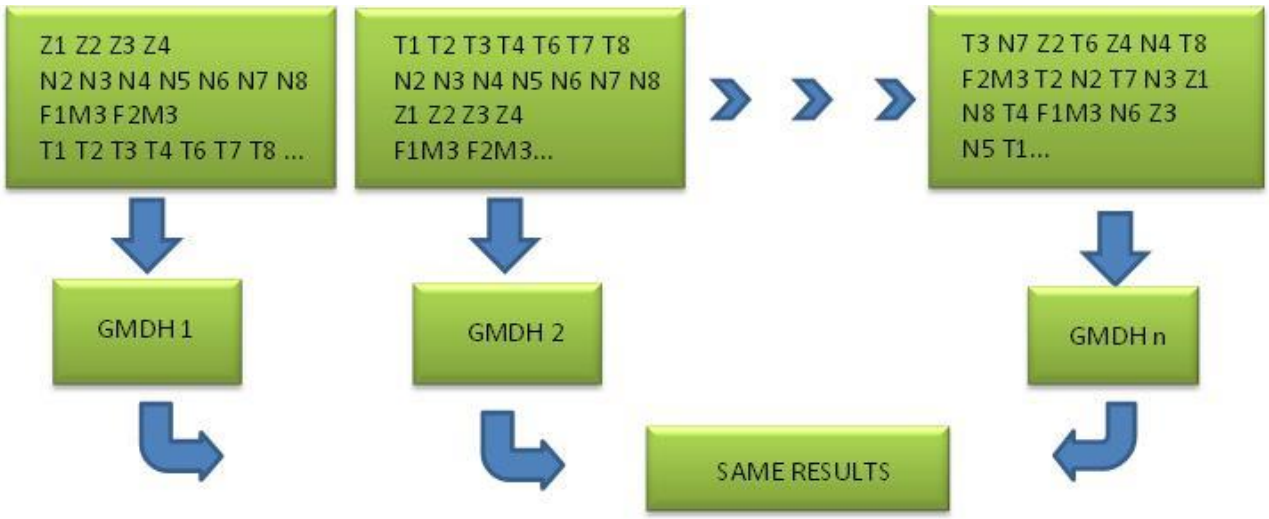


**Figure 7: Same results for different GMDH input variable architectures.**

## 6. CONCLUSIONS

A Monitoring and Diagnosis System was developed based on GMDH and ANN methodologies, and applied to the Ipen research Reactor IEA-1. The system performs the monitoring by comparing the GMDH and ANN calculated values with measured ones. The results of ANN methodology are strongly dependent on which variables are used as neural network input. On the other hand, as the GMDH is a self-organizing methodology, the input variables choice is made automatically. This work presented a study of input variable of GMDH methodology. For each one of the 58 SAD variables, 57 different GMDH models were developed with different architectures, varying the set of input variables. GMDH models were tested and all gave the same result. By sorting of different solutions, GMDH algorithms aims to minimize the influence of the author on the results of modeling. Computer itself finds the structure of the model and the best solution.

## ACKNOWLEDGMENTS

The authors would like to express their thanks to Mr. Walter Ricci and the IEA-R1 operators for reactor data.

## 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.
6. Gonçalves, I. M. P. “Monitoração e diagnóstico para detecção de falhas de sensores utilizando a metodologia GMDH”. Tese (Doutorado), Universidade de São Paulo - IPEN, 2006.