The Fuzzy Group Method of Data Handling and Its Application to the Problems of the Macroeconomic Indexes Forecasting

Yurii P. Zaychenko
Institute for Applied System Analysis, National Technical University "KPI"
Av.Pobedy 37, Kiev 03056, Ukraine, E-mail: zaych@mmsa.ntu-kpi.kiev.ua
Alexey G. Kebkal
Institute for Applied System Analysis, National Technical University "KPI"
Av.Pobedy 37, Kiev 03056, Ukraine, E-mail: alexk@mmsa.ntu-kpi.kiev.ua
Valerii F. Krachkovckii
Institute for Applied System Analysis, National Technical University "KPI"
Av.Pobedy 37, Kiev 03056, Ukraine

Introduction

This paper deals with investigation in a fuzzy variant of a Group Method of Data Handling (GMDH) and its applications to the problems of the macroeconomic indexes forecasting. This method belongs to self-organising methods of complex systems, and allows revealing of the internal laws of the appropriate object area. The advantage of GMDH algorithms is the possibility of constructing optimal models with very small number of observations and unknown dynamics between these variables. Irrefutable advantage of such algorithms are objective analysis of model choice on the basis of the generation of partial descriptions and sequential selection of them by external criteria of accuracy that are the external supplement to the models. The goal is to build adequate model of the minimal complexity by using this algorithm.

Account must be taken of that in mathematical modelling, the natural system is influenced by large amount of environmental factors, and it is not possible to take into consideration all of them separately. One of the ways to take into account the aggregate of such factors is to use the fuzzy logic.

Problem description

There is a set of initial data $\Omega_n = \{Y, X_1, ..., X_n\}, X_n \in R^M$, where n are a number of variables, M are a number of observations. The task is to synthesise regression equation $y = f(x_1, ..., x_n)$ by using fuzzy GMDH that should be adequate to the initial data and that the resulting model should be of the minimal complexity.

As the experimental data the following macroeconomic indexes (from January 1995 to February 1997) were used: consumer prices index (CPI), wholesale prices' index for industrial production, industrial production, retailing, credit investments to economy, weighted mean dollar rate, money profit per capita, monetary aggregate M0, monetary aggregate M2, average salary, rate of the National Bank of Ukraine, credit debts between enterprises, volume changes of the gross national product.

This paper is aimed to construct an adequate model which reveals dependence of CPI index from the other macroeconomic indexes has the minimal complexity and is to be used for inflation process forecasting.

Fuzzy Group Method of Data Handling

Partial model construction

Linear interval regression model [4] for partial model construction has been put into use in as follows:

$$Y = B_1 z_1 + B_2 z_2 + ... + B_n z_n$$

where z_i are certain known variables, B_i are intervals, prescribed by the triangle fuzzy numbers (α_i, c_i) where α_i are the centres of the fuzzy number and c_i are its widths. Thus Y may be calculated in a form:

$$Y = \left(\sum \alpha_i z_i, \sum c_i |z_i|\right)$$

Inclusion relation for the two intervals B_i and B_j ($B_i \subset B_j$) may be described:

$$\alpha_j - c_j \le \alpha_i - c_i, \alpha_j + c_j \ge \alpha_i + c_i$$

In our case variables z_i accord with variables x_i and x_j for certain partial models as follows:

$$z_1 = 1, z_2 = x_i, z_3 = x_j, z_4 = x_1^2, ..., z_{n-1} = x_i \cdot x_j^{r-1}, z_n = x_j^r.$$

Let's consider the estimation method for the linear interval regression model. May we have M measurements of the n+1 variables, and besides n are standalone magnitudes, and n+1's depends upon them. And this dependence is unknown. Thus estimation of the linear interval model for partial fuzzy GMDH model may take form

$Y_i^* = A^*_{00} + A^*_{10} x_i + A^*_{11} x_j + A^*_{20} x_i^{\,2} + ... + A^*_{rr-1} x_i x_j^{\,r-1} + A^*_{rr} x_j^{\,r}$ and constructs as follows:

1. Initial data y_i are included to the estimation interval Y_j^*.
2. The estimation interval width should be minimised.

Such requirements may be reduced to the linear programming problem for model coefficient estimation as follows (for all nodes):

$$\min\left(c_0 \cdot M + c_1 \cdot \sum_{k=1}^{M} |X_{ki}| + ... + c_{C_{r+1}^2 - 1} \cdot \sum_{k=1}^{M} |X_{ki} \cdot X_{kj}^{r-1}| + c_{C_{r+1}^2} \cdot \sum_{k=1}^{M} |X_{kj}^r|\right)$$

$$\alpha_0 + \alpha_1 \cdot X_{ki} + ... + \alpha_{C_{r+1}^2} \cdot X_{kj}^r - (c_0 + c_1 \cdot |X_{ki}| + ... + c_{C_{r+1}^2} \cdot |X_{kj}^r|) \le Y_k$$

$$\alpha_0 + \alpha_1 \cdot X_{ki} + ... + \alpha_{C_{r+1}^2} \cdot X_{kj}^r + (c_0 + c_1 \cdot |X_{ki}| + ... + c_{C_{r+1}^2} \cdot |X_{kj}^r|) \ge Y_k$$

$$k = \overline{1, M}$$

$$c_p \ge 0, p = \overline{1, C_{r+1}^2}$$

The goal is to minimise the variable domain of changing of Y initial values at the cost of seeking such interval width values c_i and interval's centres $\alpha_i \;\; i = \overline{1, C_{r+1}^2}$ that guarantee the dispersion minimum of the Y if such measurements are in their interval. To solve this problem it is better to

turn to the following dual task:

$$\max\Big(\sum_{k=1}^{M} Y_k \cdot \delta_{k+M} - \sum_{k=1}^{M} Y_k \cdot \delta_k\Big)$$

$$\sum_{k=1}^{M} \delta_k - \sum_{k=1}^{M} \delta_{k+M} = 0$$

$$\sum_{k=1}^{M} X_{ki}\,\delta_k - \sum_{k=1}^{M} X_{ki} \cdot \delta_{k+M} = 0$$

....

$$\sum_{k=1}^{M} X^{r}_{kj}\,\delta_k - \sum_{k=1}^{M} X^{r}_{kj} \cdot \delta_{k+M} = 0$$

$$\sum_{k=1}^{M} \delta_k + \sum_{k=1}^{M} \delta_{k+M} \le M$$

$$\sum_{k=1}^{M} |X_{ki}| \cdot \delta_k + \sum_{k=1}^{M} |X_{ki}| \cdot \delta_{k+M} \le \sum_{k=1}^{M} |X_{ki}|$$

.....

$$\sum_{k=1}^{M} |X^{r}_{kj}| \cdot \delta_k + \sum_{k=1}^{M} |X^{r}_{kj}| \cdot \delta_{k+M} \le \sum_{k=1}^{M} |X^{r}_{kj}|$$

$$\delta_k \ge 0,\ i = \overline{1, 2 \cdot M}$$

Obviously such task always has the solution. And we are able to get this solution by any standard linear programming algorithm, and turn back to the straight problem to find the values of variables c_i and α_i, $i = \overline{1, C^2_{r+1}}$.

Sorting-out procedure description

One of the specific feature of the self-organising systems which differs them from the others is that the desired model is able to change its own structure by accumulating initial variables powers and adopting simultaneously coefficients of the model. Polynomial, which describes unknown model, depends on two variables, so it is possible to construct C_n^2 different models at the initial stage. After that it is essential to choose several (one) best models due to the external criteria and to try to forecast the unknown value with them. Such approach was named as single sorting–out procedure GMDH method. But after the first sorting-out procedure it is worth while to try to get much better models on a base of the resulting models. Such approach leads to the creation of the multiple sorting–out procedures GMDH method. And this method was realised in proposed work. Let's consider it properly.

In order to get the models of the second sorting–out procedure the reference function should be defined, the arguments of which are functions-models we've got at the previous sorting–out procedure.

In our case the reference function is the polynomial of the second power which depends on two variables - $Y_{ij}^{k+1} = f(Y_i^k, Y_j^k)$, where k is the number of sorting–out procedure.

The rule of selection termination is proximity of average models criterion of two adjoin sorting–out procedures:

$$-\varepsilon < N_{см}^{k+1} = \frac{1}{F} \cdot \sum_{i=1}^{F} n_{см}(Y_i^{k+1}) - N_{см}^{k} < \varepsilon$$

External criteria of accuracy

Let's consider regularisation and non-biased criteria applied to our investigation. A general feature for these criteria is their usage as external supplement.

Sampling N is split into training sampling N_A for model parameters estimation and verifying sampling N_B for determining of the model adequacy. Regularisation criterion determines mean square error (MSE) of the model for verifying sampling:

$$\Delta^2(B) = \sum_{t \in N_B} (y_t^M - y_t)^2 \Big/ \sum_{t \in N_B} y_t^2 \to \min$$

Using the assumption that good approximation quality in the past guaranties the good approximation in the immediate future, other factors being equal, it is reasonable to implement this criterion in short-term forecasting. The solution on the new nodes gives us only small deviation from the initial data.

However, there is possibility to loose significant variables in such selection process but their influence is included in other variables indirectly.

The next important quotient is a non-biased criterion. Analysts who possess good experience as a rule even don't imagine that they are working with conflicting system. The non-biased criterion is based on the fact that the resulting models should be close enough for different samplings of one object, other factors being equal.

Such criterion maybe represented as follows:

$$n_{см} = \frac{1}{R_1 + R_2} \cdot \sum_{r=1}^{R_1+R_2} (z_r^* - z_r^{**})^2$$

where R_1, R_2 are sizes of the first and second sub-samplings accordingly, z_r^*, z_r^{**} - forecast values of the first and second models accordingly. Non-biased criterion takes the value 0 on the adequate models. But in some cases this criterion takes the value 0 for false models. In such a situation it is worth while to split the sampling into three, four or more sub-samplings, till just but one non–biased model remains. Such model is to be optimal.

The following selection criterion have been applied to in our work, it is the convex combination of the regularisation and non-biased criteria:

$$K_{\Sigma} = \alpha \cdot \Delta^2(B) + (1-\alpha) \cdot n_{см}$$

where α is a weight coefficient, $0 \le \alpha \le 1$.

General algorithm description

1. A selection of a general model form, which will describe unknown dependence.
2. A selection of the external criteria of accuracy and of the freedom choice.
3. A selection of a general reference function form.
4. The counter of models number k and the counter of sorting–put procedures' number r are assigned the value of 0.
5. New partial model construction. Criteria calculation. k=k+1.
6. If k$\geq C_F^2$, then k=0, r=r+1.Estimation of the average models criterion $N_{см}^r$. If r=1 then go to the step 5, else go to the step 7.
7. If $\left|N_{см}^r - N_{см}^{r-1}\right| \leq \varepsilon$, then go to the step 8, else select F the best models according to an external criteria and go to the step 5.
8. Choose the best model from among F final models with external criterion. At last construct the analytic form of the best model with Goedel's indexing.

Work results and experimental data analysis

As result of the work program Designer was created, which allows to apply fuzzy GMDH method for any sampling of initial data. To set initial data one should form project-file in text-format where such parameters as complexity of reference function (maximum polynomial power), number of variables, number of interpolation nodes, matrix of initial data, freedom of choice and some auxiliary information should be specified. The program is to construct the model, which is adequate to initial sampling and depicts results in a table form, analytical form and figures. User has possibility in a run-time mode to change the freedom of choice, weight coefficient of external criterion and others parameters, to save results of work of the program in a text and graphical data formats.

On the basis of the problem definition, the solution is to find compromise due to the need of getting the high level of model constructing explicitly simultaneously receiving the minimal complexity. Thus, in principle there some such models exist, which satisfy the conditions, but a selection of the best one may be made by an expert. The program gives us the possibility not only to investigate the best model at the external criterion, but also to investigate all the models of last sorting- out procedure.

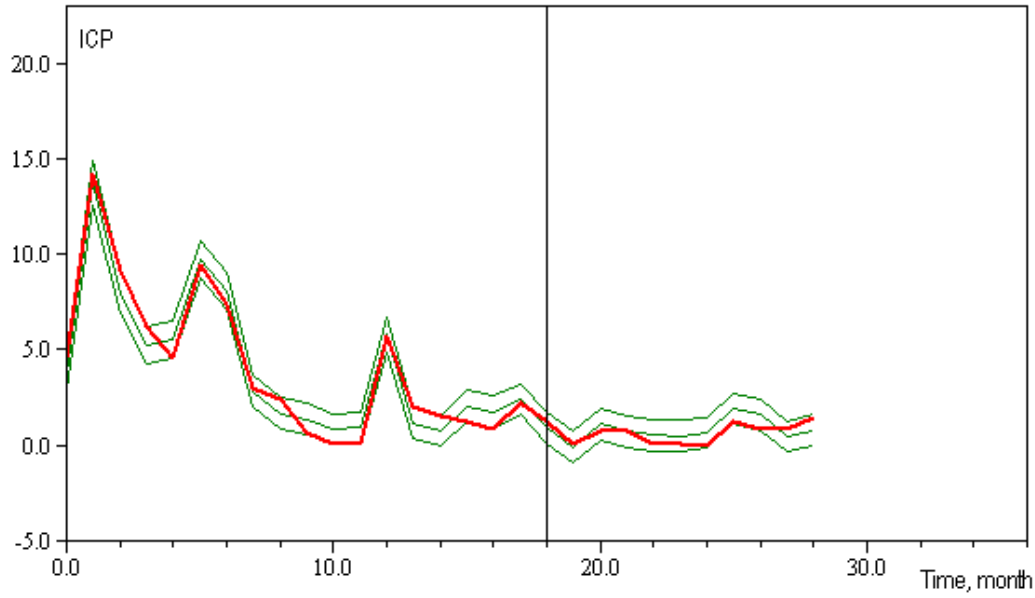

Figure1. Model, which has been constructed by fuzzy GMDH method. Vertical line splits the training and verifying samplings. The zero value fits of July 1995 and the twenty-eights fits of December 1997. MSE - 0.51.

Figure 1 demonstrates the results of model construction with a quadratic reference function. In accordance with the results we may ratify we have obtained an adequate model and it is close enough to real data both for training sampling and for verifying sampling.

During constructing of model with the quadratic reference function we faced the problem of selection of the weight coefficient of the external criterion which allowed us to build the adequate model. For different initial data (different training sampling size here) we have obtained different permissible domain for weight which allowed us to get adequate model. So we have researched the “effective domain” of weight coefficient which permitted us to get an adequate model. An example of such domain is demonstrated in the Figure 2. Effective domain is placed between two curves there.

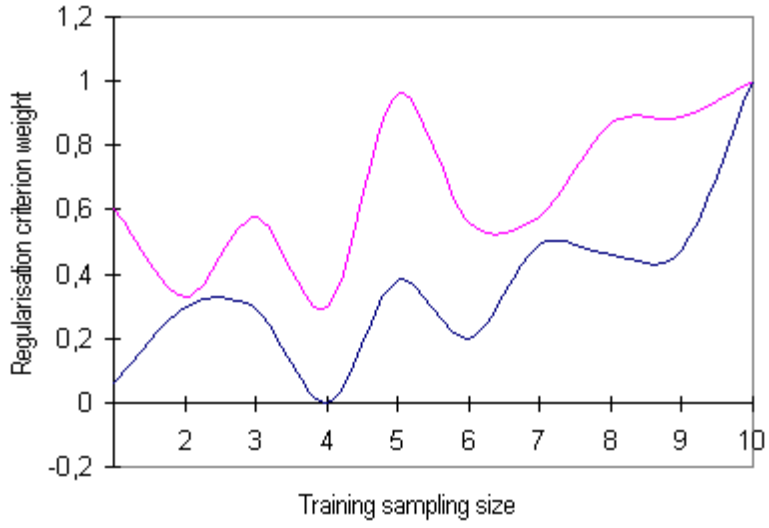

Figure.2. Effective domain of weight coefficient for external criterion with a quadratic reference function.

Table 1 demonstrates the MSE values for different training sampling sizes for models of different complexity. As evident from the table, in a case of linear reference function MSE is twice greater than in the case of quadratic reference function in effective domain. But the advantage of linear models is that they are faster at computing time, thus such models are good for robust estimation. But in a case of weight coefficient selection outside of the effective domain for quadratic reference function we get results which are nine times worse than the same ones in a linear case and eighteen times worse than the same ones in a case of quadratic models in effective domain. It should be pointed out that existence of such domain proves a necessity to use several external criteria at a time and that is essentially increasing quality of modelling.

Table 1. MSE for constructed linear and quadratic models using experimental data.

The size of training sampling	MSE		
	Linear models	Quadratic models in the effective domain	Quadratic models outside of the effective domain
1	0.821	0.386	0.506
2	0.463	0.410	19.427
3	1.880	0.483	4.922
4	0.922	0.562	1.632
5	0.842	0.526	6.862
6	1.073	0.901	1.550
7	1.039	0.703	3.044
8	0.984	0.661	2.847
9	1.053	0.634	3.738
10	1.212	0.511	52.523

Conclusions

The new variant of heuristic self-organisation method of complex system modelling – fuzzy GMDH method which uses the theory of fuzzy sets has been considered in this paper. The fuzzy GMDH method's algorithm and its results in experimental analysis have been represented. Usage of the method has been illustrated by an example of the problem of Ukraine macroeconomic indexes' forecasting. It should be particularly emphasised that in this work the combination of two external criteria was used for constructing of models that allowed us to increase the quality of the models. But such approach poses the following problem that is how to preestimate weight coefficients of external criteria.

Literature

1. Ивахненко А.Г., Мюллер И.А. Самоорганизаия прогнозирующих моделей. - Киев, Техника, 1985.-221с.
2. Ивахненко А.Г., Юрачковский Ю.П. Моделирование сложных систем по экспериментальным данным. - М.: Радио и связь, 1986. - 118с.
3. Ивахненко А.Г., Зайченко Ю.П., Димитров В.Д. Принятие решений на основе самоорганизации. М: Советское радио,- 1976.- 260с.
4. Прикладные нечеткие системы/ Т… Тэрано (ред.), Ю.Н.Ченышов (пер.).-М.:Мир, 1993.-386с.