

Influence of sample division on the quality of modeling and forecasting of real processes

Nina Kondrashova

International Research and Training Center of Information Technologies and Systems of the

National Academy of Sciences of Ukraine, Ukraine

nkondrashova@ukr.net

Abstract. *Sample division and a criterion for choice of the best division are important elements in GMDH algorithms. Sample divisions effective in the tasks of approximation, extrapolation and forecasting are considered in the paper. The main attention was placed to quasi-optimal sample division which enables to enhance the extrapolation and forecasting precision in combination with an adaptive prognosis. Some set of sample division methods allows to choose a proper technology for every task taking into account the object features.*

Keywords

GMDH, sample division, quasi-optimal, division by dispersion, approximation, extrapolation, forecast.

1 Introduction

Theoretical researches and numerical models experiments shows that the separation of optimum «nucleus» of sample is an important stage for the increase of exactness of modeling and forecasting [1].

The increase of exactness of models in GMDH algorithms is the purpose of this work due to the search of the best division of initial sample under the noise conditions. Sample W represented by a table has two basic parameters: number of variables M and of points N . Due to the nucleus optimization, the optimum subset of both variables (columns) s^* , m_s^* and points (rows) p^* , n_p^* is searched (where s^* is model complexity (structure) and p^* is composition (subset of rows) of subsample).

For taking into account features of approximation, extrapolation and forecasting tasks using GMDH we will differentiate sample division technologies for the construction:

- a) *approximation* models using input and output data of the whole sample $W = A \cup B$ without an examination subsample ($D = \emptyset$);
- б) *extrapolation* models using input and output information of subsamples A and B for learning and checking. Data of examination subsample D are used for extrapolation of the output variable ($W = A \cup B \cup D$, $D \neq \emptyset$, $A \cap B = \emptyset$, $A \cap D = \emptyset$, $B \cap D = \emptyset$);
- в) *forecasting* models which unlike to extrapolation models *do not use input data* of this subsample for the forecast of output variable on the examination subsample D .

2 Theoretical Part

2.1 Division types

Samples can be divided at will (*sequentially*) or in accordance with a criterion. It is possible to divide sample division criteria into two groups due to the purpose of modeling:

- for stationary objects or processes, (the hidden similarity of sample parts is assumed);
- for nonstationary objects or processes.

Types of division: 1. *«by dispersion»*,
2. ρ -*proportional* division [2].

Division *«by dispersion»* deals with the moment characteristics:

$$\sigma^2 = \left[M(\mathbf{x}_1 - M\mathbf{x}_1)^2, \dots, M(\mathbf{x}_N - M\mathbf{x}_N)^2 \right], \quad \sigma^2 = (\sigma_1^2, \dots, \sigma_N^2), \quad M\mathbf{x}_n = \frac{1}{M} \sum_{i=1, m} x_{n,i},$$

$$M(\mathbf{x}_n - M\mathbf{x}_n)^2 = \frac{1}{M-1} \sum_{i=1, m} (x_{n,i} - M\mathbf{x}_n)^2, \quad n = \overline{1, N},$$

where σ_n^2 , $n = \overline{1, N}$ is dispersion of arguments in n -th point; $M(\cdot)$ is mathematical expectation; \mathbf{x}_n is an n th row ($\dim \mathbf{x}_n = 1 \times M$) of the initial data matrix \mathbf{X} ($\dim \mathbf{X} = N \times M$).

ρ -*proportional* division deals with information matrices where $\chi_i = \mathbf{X}_i^T \mathbf{X}_i$, $i = \overline{1, 2}$ are matrices of complete rank, $\dim \mathbf{X}_i = n_i \times M$, $n_2 = (N - n_1)$ at the condition that $\max(n_1, n_2) \geq m_{\min} + 1$. m_{\min} is minimum number of variables in GMDH models.

ρ -*proportional «quasi-optimal»* division minimizes the norm of difference of information matrices (submatrices):

$$p_{\min}^*(N_v, n) = \arg \min_{\rho_\ell^2 \neq 0, \ell = \overline{1, L_v}} \left\| \rho_\ell^2 \chi_{1v_\ell} - \chi_{2v_\ell} \right\|, \quad v = \overline{1, V},$$

V is maximal number of samples being divided. Number of division variants equals to the number of simple combinations:

$$L_v = \sum_{n=n_{\min_v}}^{n_{\max_v}} C_{N_v}^n. \quad (1)$$

The value v determines the number of the subsample (sample) which is divided into two subsamples. We will write down for the result of matrix divisions:

$$\mathbf{X}_v^T = \left[\mathbf{X}_{1v}^T : \mathbf{X}_{2v}^T \right]^*, \quad \dim \mathbf{X}_{1v}^* = n^* \times M, \quad \dim \mathbf{X}_{2v}^* = (N_v - n^*) \times M.$$

Similar «by dispersion» (SD) division is determined as:

$$p_{\min \sigma}^*(N_v, n_i) = \min_{\ell = \overline{1, L_\sigma}} \left\| \sum_{i \in \Omega_{N_v}} \sigma_{i_\ell}^2 - \sum_{n \in \Omega_{N_v} \setminus \{i\}} \sigma_n^2 \right\|,$$

where $L_\sigma = C_{N_v}^{n_i}$ is the full quantity of division variants at the given $n_i : (N - n_i)$ relation of subsample lengths. The traditional similar *«by dispersion»* division applied in GMDH algorithms is one of variants of mismatch error minimization for such norm under the condition of monotonous change of σ_i^2 and σ_n^2 values.

Second group of division types for nonstationary objects maximizes of the moment characteristics mismatches.

Dissimilar «by dispersion» (DSD) division is determined as:

$$p_{\max \sigma}^*(N_v, n_i) = \max_{\ell = \overline{1, L_\sigma}} \left\| \sum_{i \in \Omega_{N_v}} \sigma_{i_{v_\ell}}^2 - \sum_{n \in \Omega_{N_v} \setminus \{i\}} \sigma_{n_{v_\ell}}^2 \right\|, \quad v = \overline{1, V}.$$

Traditional for GMDH algorithms *dissimilar «by dispersion»* division is found without exhaustive search by the simple division of the rows of initial matrix ranged in accordance with the $\hat{\sigma}_n^2$, $n = \overline{1, N}$ dispersions values in $n_i : (N - n_i)$ proportion:

$$p_{\max, \sigma} = \left(\sum_{n=1}^{n_i} \hat{\sigma}_n^2 - \sum_{n=n_i+1}^N \hat{\sigma}_n^2 \right)$$

Dissimilar ρ -proportional division is:

$$p_{\max}^*(N_v) = \arg \max_{\rho_\ell^2 \neq 0, \ell=1, L_v} \left\| \rho_\ell^2 \chi_{1v_\ell} - \chi_{2v_\ell} \right\|, \quad v = \overline{1, V}.$$

2.2 Comparison of traditional and quasi-optimal division methods

On the basis of numeric modeling it was established that the structure and parameters of GMDH model do not depend on the method of division in *the approximation task* if the informative matrices of input variables satisfy to the condition of ρ -proportionality subsamples and number of points of the proper subsamples for all divisions is equal. If the condition of ρ -proportionality subsamples is not fulfilled, the «quasi-optimal» ρ -proportional division which provides larger noise immunity has some advantage for the determination of a model structure.

Traditional methods of the sample division in combination with «quasi-optimal» division can enhance the accuracy of models essentially in *the extrapolation task*. Visualization of results helps to choose necessary models more operatively.

Figure 1 shows that the *forecast* by difference equations (by models) obtained from the data [3] by quasi-optimal division method is the most accurate.

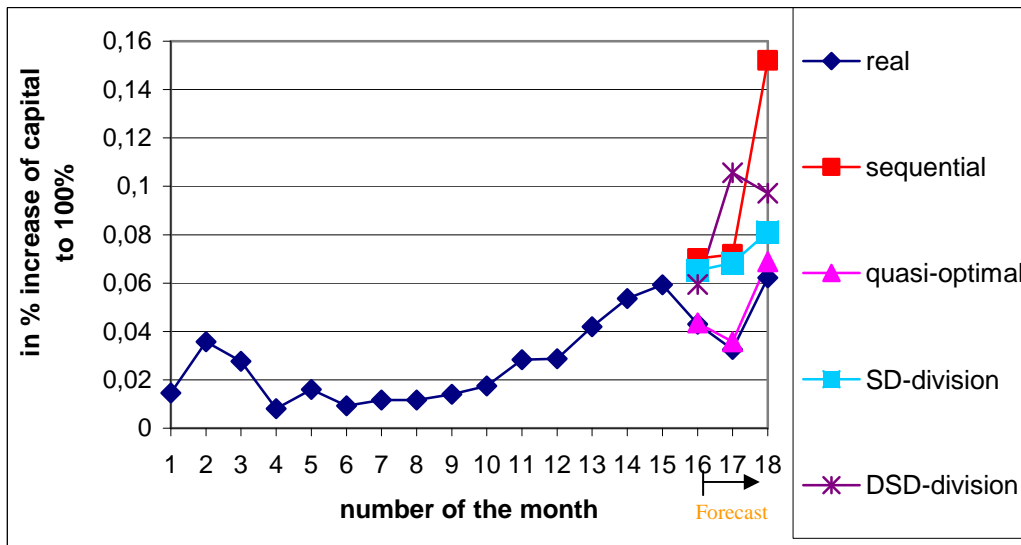


Fig. 1. Inflation index and adaptive forecast models as function of data division method

Points closed to the forecast point needed to use for building models to *forecast* the rapid change of the processes. The current forecasted values are added to the initial data for this purpose. From comparison of the graph values for models with adaptive forecast (points are connected by lines) and values in isolated points for models without adaptive forecast in a figure 2 it is visible that not only the «quasi-optimal» division but also the adaptive forecast is important for the considerable increase of the forecast accuracy [4]. The class of models and subsample volumes of different division methods was the same for the examples represented in these graphs. The type of difference models was the main distinction. The linear in parameters difference model of inflation index for $M=6$ initial variables given in $N_v=15$ points with an adaptive forecast looks as:

$$\hat{Y}_{k+L} = F(\boldsymbol{\theta}, x_{1,k}, \dots, x_{6,k}, x_{1,k-2}, x_{3,k-1}, x_{5,k-2}, Y_k, \hat{Y}_{k+1}, \dots, \hat{Y}_{k+L-1}), \quad L = 2, 3; \quad k = N_U$$

where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)$ is the vector of parameters.

The difference model of inflation index without an adaptive forecast has the form:

$$\hat{Y}_{k+L} = F(\boldsymbol{\theta}, x_{1,k}, \dots, x_{6,k}, x_{1,k-2}, x_{3,k-1}, x_{5,k-2}, Y_k), \quad L = \overline{1, n_D}, \quad k = N_U, \quad n_D = 3.$$

The forecasting error in the points of examination subsample is calculated as:

$$J_p = \sum_{i=1}^{n_D} (\hat{y}_{i_p} - y_i)^2 / \sum_{i=1}^{n_D} y_i^2.$$

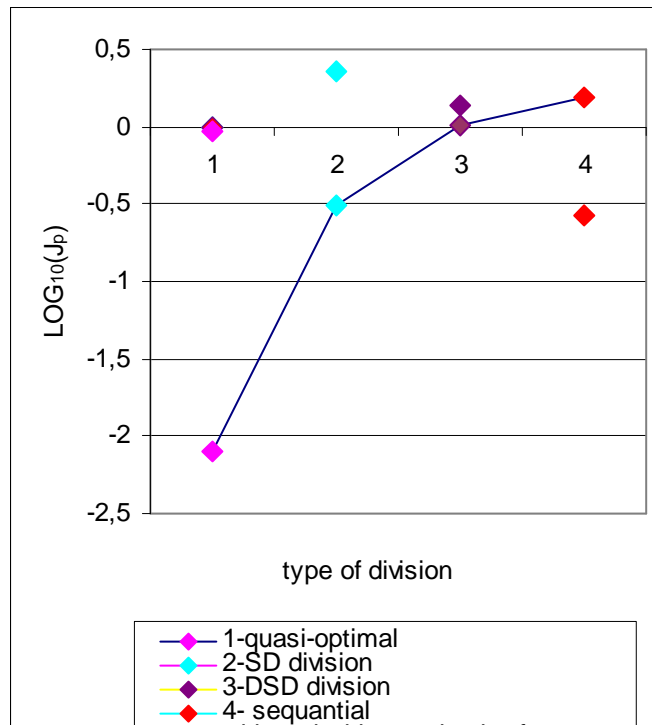


Fig. 2. Change of forecast error depending on division for models with and without an adaptive forecast.

3 Conclusion

Availability of a set of sample division methods allows to choose the proper division technology taking into account the features of an object being modeled and to compare it with alternative ones for any task.

1. The optimum division has advantage at the definition of relevance of model structure in the task of approximation.
2. A quasi-optimal division of the data gives the possibility to increase an accuracy of the extrapolation and the forecasting by GMDH models.
3. The combination of traditional and quasi-optimal methods of sample division can essentially increase accuracy of models for extrapolation of variables.
4. Using of adaptive forecast is important for the considerable increase of exactness in forecasting tasks.