# CLUSTERING OF MULTIVARIATE HETEROGENEOUS TIME SERIES

**Antonova I.V., Chikina N.A.**
***National Technical University «Kharkiv Polytechnic Institute», Kharkiv***

The task of clustering objects presented in the form of multivariate heterogeneous time series is one of the actual problems of mathematical modeling. Heterogeneous series are the series where the description of the state of the studied object $S$ at each time moment $t$ is carried out by quantitative, ordinal or categorical variables. Such problems arise when building object models in poorly structured research areas, for example, in medicine, when determining a risk factors set, forming diseases development risk groups, etc. In this case, the identification of possible system states can be carried out both for each of these indicators, and for their arbitrary combination, depending on the goal of research.

The widely known clustering algorithms use the concept of «distance» between objects (groups of objects). Determining the distance or measure of the difference between the studied time series has additional difficulties. So, for example, the series can be of different lengths, have a large dimension, consist of heterogeneous components. In this case, the task is complicated by the diversity of characteristics of the studied object $S$. Equally important is the presence of dependencies between the characteristics. It is known that it is impossible to introduce a metric in the case of heterogeneous characteristics of the studied object.

To solve the problem of automatic classification in the 70s of the twentieth century G. Lbov (Sobolev Institute of Mathematics, Siberian Branch of the Russian Academy of Sciences) proposed a method for detecting logical regularity for constructing logical decision rules (decision tree). The use of such a class of decision functions has several advantages: visibility and simple interpretability of the resulting rules.

Let be given $N$ time series $Y^i$, $n_i \left(i = \overline{1,N}\right)$ is the length of the series $Y^i$. Let us denote by $X^i_{tк}$ the set of values of the heterogeneous characteristics of the studied object $S$ in the series $Y^i$, measured at the time $t_{iк}$ $\left(1 \le k \le n_i\right)$, $y^i_{tk}$ is the corresponding value of the objective function. Thus, the series can be represented as $Y^i = \left\{X^i_{t1}, y^i_{t1}; ...; X^i_{t_{ni}}, y^i_{t_{ni}}\right\} \left(i = \overline{1,N}\right)$.

If $\mu_{ij}$ is the root-mean-square error of the decision tree, then the smaller value of $\mu_{ij}$, the more likely that the time series $Y^i$ and $Y^j$ belong to the same cluster. In this case, it is natural to define the difference measure $\rho(i, j)$ between the series as the average value of $\mu_{ij}$ and $\mu_{ji}$.