COMPARATIVE ANALYSIS OF SHORT TIME SERIES
Antonova I.V., Chikina N.O.
National Technical University «Kharkiv Polytechnic Institute», Kharkiv

Statistical data on most social, demographic, and environmental indicators are collected most often once a year. Such series are available only for small amounts of data of a few years, i.e., short time series. Most existing clustering methods are not suitable for analysing short time series. Therefore, it becomes an actual problem how to present information about numerical series so that it is sufficient for the purposes of clustering tasks for short time series.

In recent scientific literature, there have been publications recommending the inclusion of additional information derived from statistical indicators of time series data, along with time-invariant indicators that characterise their dynamics and variability.

The task of choosing measures of series proximity remains relevant. Existing studies on the comparative analysis of proximity measures tend to show that the Euclidean distance is the best quality in time series clustering tasks.

In this paper, the authors studied time series reflecting the number of various skin disease cases in the regions of Ukraine over a certain period of time. The results of clustering such time series make it possible to compare the data and identify groups of regions of Ukraine with the «same» situation on this issue.

Further studies will identify the causes of the respective morbidity rates, such as water, air, and environmental pollution. To solve this problem, cluster analysis procedures for short time series with three proximity measures were performed: Euclidean distance, a measure based on the series' statistical characteristics, and a measure characterising the dynamics of the numerical series.

As demonstrated by the results of the performed procedures, proximity measures based on the computation of statistical characteristics and dynamic properties of time series have confirmed their reliability, enabling their use in clustering tasks involving short time series. This also supports the confident application of variability-based features, such as sequences of second-order differences, as valid proximity measures for time series data.

The clustering results of the analysed time series of morbidity rates for the selected three regions of Ukraine will be further examined to identify the underlying factors driving the observed grouping of oblasts into clusters, as well as the causes of anomalous morbidity levels for certain skin diseases in specific cities within the regions. These anomalies are generally associated with environmental indicators such as air and water quality deviations. The exact nature and magnitude of the morbidity risks will require further investigation in dedicated studies.