https://doi.org/10.31891/csit-2026-2-5

Yurii TURBAL
Doctor of Engineering Sciences, Professor
Computer Science and Applied Mathematic Department
National University of Water and Environmental Egineering
https://orcid.org/0000-0002-5727-5334
e-mail: turbaly@gmail.com

Oleksandr KUBAI
Senior lecturer
Computer Science and Applied Mathematic Department
National University of Water and Environmental Egineering
https://orcid.org/0000-0002-2005-487X
e-mail: o.v.kubai@nuwm.edu.ua

Received: 13/04/2026
Accepted:12/05/2026
Published:31/05/2026

UDC 004.8:004.62

A FORECASTING METHOD BASED ON CLUSTERING THE POLYNOMIAL EXTRAPOLATION SEQUENCE

Time series forecasting is an important task in intelligent data analysis, especially under conditions of short samples, local non-stationarity, noise, and increased sensitivity to external disturbances. These properties are characteristic, in particular, of financial time series, where local trends, random fluctuations, and abrupt changes in dynamics may coexist even over small observation intervals. One of the promising approaches to short-term forecasting is polynomial extrapolation. However, the use of polynomials of different orders for the same segment of a series produces a set of alternative forecast values, which complicates the selection of the final forecast.

This paper proposes a short-term forecasting method based on clustering the values of the polynomial prediction sequence. For a local fragment of a time series, a sequence of polynomial forecasts p_m *is formed over a range of polynomial orders, after which cluster analysis is applied to this set of values. The densest interval method and the DBSCAN algorithm are used to identify the dominant forecast region, while the final forecast value is defined as the central characteristic of the detected cluster. The efficiency of the proposed approach is compared with the forecasting method based on averaging the polynomial extrapolation sequence.*

Experimental studies were carried out on deterministic functions, stochastic sequences, and real intraday stock data for Netflix using the Close parameter. It was found that the polynomial prediction sequence has an internal structure in the form of local extrema, concentration intervals, and distant values, which justifies the feasibility of its clustering. The scientific novelty of the study lies in refining the mechanism for selecting PPS elements by moving from index-based averaging to structural analysis of the spatial grouping of forecast values. The practical significance of the work lies in improving the robustness of short-term forecasting for financial time series.

Keywords: polynomial extrapolation; short-term forecasting; time series; polynomial prediction sequence; forecast clustering; DBSCAN; financial time series; stock market data.

Introduction

Time series forecasting is one of the fundamental tasks of intelligent data analysis, since forecast quality directly affects the effectiveness of decision-making in financial, economic, technical, and information domains. This task becomes particularly challenging when only a short local sample is available, the process is non-stationary, and the data contain noise components and random disturbances. Such conditions are typical of financial time series, especially intraday stock market dynamics, where local trends, short-term fluctuations, and abrupt directional changes may coexist even over small time intervals.

One of the promising directions in short-term forecasting under small-sample conditions is polynomial extrapolation. Its advantage lies in the possibility of constructing a forecast from a limited number of the most recent observations without requiring large training datasets or complex parameterized models. At the same time, applying polynomials of different orders to the same local fragment of a series produces not a single forecast value, but a set of alternative estimates. Such a set can reasonably be interpreted as a Polynomial Prediction Sequence (PPS).

It is precisely the mechanism for forming the final forecast from the PPS that represents one of the key problematic aspects of the polynomial approach. The use of a single fixed-order polynomial, or even averaging a certain part of the forecast sequence, does not always make it possible to account for the

internal structure of the set of p_m values. In practice, the PPS may contain both mutually close values and substantially distant estimates arising from the instability of individual polynomials, sensitivity to noise, or the specific local geometry of the series. Under such conditions, direct averaging may result in the loss of information about the actual organization of the forecast set.

In this regard, the development of a method capable of analyzing not only individual forecast values but also the pattern of their grouping is of clear relevance. If PPS values form a dominant region of concentration, then the center of such a region may provide a more stable and informative forecast estimate. In this paper, an approach is proposed in which a sequence of polynomial forecasts is generated for a local fragment of a time series over a predefined range of polynomial orders, after which cluster analysis is applied to this set of values. The densest interval method and the DBSCAN algorithm are used to identify the dominant forecast region. The effectiveness of the proposed approach is evaluated on deterministic, stochastic, and real financial data and is compared with the method based on averaging the polynomial extrapolation sequence.

Related Works

Over the past few years, research on time series forecasting has developed mainly along three directions: deep neural models for financial series, hybrid and ensemble forecast combination schemes, and clustering methods that make it possible to identify the hidden structure of either the data or the set of forecast estimates. A review of recent studies shows that the modern literature is increasingly shifting from isolated models to hybrid architectures; at the same time, however, the problems of overfitting, distribution shift, and insufficient interpretability of results remain unresolved. A limitation of most studies is that they systematize existing approaches rather than propose mechanisms for dealing with a small number of observations and the instability of local forecasts.

Study [1] considers financial forecasting from the perspective of comparing stand-alone neural network models and hybrid deep learning models. The authors emphasize that hybrid architectures, in particular combinations of convolutional neural networks, long short-term memory networks, and the attention mechanism (CNN – Convolutional Neural Network + LSTM – Long Short-Term Memory + Attention), often provide higher accuracy than individual models. The strength of this work lies in its clear representation of models as a sequence of stages: input data, feature extraction, and prediction. At the same time, an important limitation of such approaches is their reliance on complex architectures that require substantial amounts of data and computational resources. For short-term forecasting tasks on short financial series, this reduces their practical feasibility.

The systematic review [2], conducted in accordance with the PRISMA standard (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), covers 126 scientific papers across six thematic areas in finance and shows that machine learning and deep learning methods generally outperform classical models in many applied financial tasks. At the same time, the authors stress that direct comparison of results across different studies is complicated by the use of different input feature sets, different evaluation metrics, and non-uniform model validation procedures. This is an important critical conclusion, since the high quality of an individual model does not in itself indicate its universal reliability. Therefore, there remains a need for forecasting methods that ensure not only accuracy but also robustness to changes in the local structure of data. An additional limitation of this review is that it uses only materials from the ScienceDirect database, which means that some thematically relevant sources may have remained outside the scope of analysis.

A separate line of research is represented by studies on forecast combination. Review [3] shows that combining several forecasts has become one of the mainstream approaches in modern forecasting, and that combination schemes have evolved from simple averaging to nonlinear and time-varying weighting mechanisms. The strength of this work lies in its theoretical justification of why a set of forecasts is often more useful than the choice of a single “best” forecast. However, the review leaves open the question of how exactly to identify the most reliable subset of forecast values in a specific local situation when the set of forecasts contains outliers or several concentration zones. It is precisely here that clustering the sequence of polynomial forecasts may provide a more local and robust selection mechanism.

Recent financial time series studies are also characterized by the active use of transformer-type models. For example, study [4] proposes a Modality-aware Transformer that combines numerical time series with textual financial reports and uses multiple levels of the attention mechanism to align information from different source types. The advantage of this approach lies in its multimodal nature and the partial interpretability of results due to the attention mechanism. At the same time, its limitation is its orientation toward large volumes of heterogeneous data and a complex feature engineering system. Therefore, this approach does not solve the problem of constructing a robust forecast from a short local sample based solely on a sequence of numerical observations.

A similar idea is implemented in modern hybrid models. Thus, in [5], a combined model is proposed for forecasting the CSI 300 index that integrates Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN), the Informer architecture for modeling long temporal dependencies, and a Long Short-Term Memory recurrent neural network (LSTM). The effectiveness of the approach is evaluated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). Despite its high predictive power, this approach is characterized by substantial algorithmic complexity, a large number of parameters, and

dependence on the structural characteristics of a specific series. In contrast, the focus of our study is not on increasing model complexity but on making fuller use of the information contained in the set of polynomial forecasts.

Studies closest to the idea proposed in our research are those in which clustering is integrated directly into the forecasting procedure. Thus, in the TSEN model (Temporal-Spatial dependencies ENhanced deep learning model), time series first undergo screening and clustering stages, after which a Long Short-Term Memory recurrent neural network (LSTM) is applied to each identified cluster. Inter-cluster dependencies are then analyzed using the attention mechanism. The advantage of this approach lies in confirming that clustering can improve the structure of input data before the forecasting stage. At the same time, in this work clustering is applied to the time series themselves or to their groups, rather than to the set of alternative forecast values for a single time point [6]. Therefore, the problem of selecting the final forecast from several local estimates remains unresolved within this approach.

Study [6] develops an approach based on the combination of cluster analysis, a convolutional neural network (CNN), and a bidirectional long short-term memory network (BiLSTM, Bidirectional Long Short-Term Memory). The advantages of this approach are demonstrated on the M3 and M4 datasets, corresponding to the third and fourth M competitions for time series forecasting and widely used as benchmark test environments. An important strength of this work is its justification of the role of clustering as a means of preparing time series for the forecasting stage. At the same time, clustering is performed there at the level of transformed series rather than at the level of a set of alternative forecasts constructed for one-step-ahead prediction. Consequently, this work confirms the promise of clustering-based forecasting but does not solve the problem of selecting the final forecast from a sequence of alternative polynomial estimates.

In our own previous studies [7–9], it has already been shown that analyzing the entire sequence of polynomial forecasts is productive: first through an algorithm for automatically selecting the optimal polynomial degree, and later through averaging part of the forecast sequence. The strength of these works lies in the transition from a single polynomial to the analysis of a set of forecasts; their weakness is that averaging still smooths the structure of the set and does not allow one to explicitly identify the dominant region of concentration of values. It is precisely this circumstance that forms the scientific niche for a new method in which the final forecast is constructed on the basis of clustering the values of the polynomial prediction sequence rather than on the basis of a single degree or a global average.

Purpose

The aim of this study is to develop a short-term time series forecasting method based on clustering the values of the polynomial prediction sequence in order to refine the mechanisms for selecting PPS elements during averaging. To achieve this aim, the study is intended to: construct a sequence of polynomial forecasts for a local fragment of a time series; investigate its structural properties on deterministic, stochastic, and real financial data; develop procedures for identifying the dominant region of forecast values using the densest interval method and DBSCAN; compare the obtained results with the forecasting method based on averaging the polynomial extrapolation sequence; and evaluate the practical applicability of the proposed approach using intraday stock market data for Netflix shares based on the Close parameter.

Materials and Methods

The object of the study is the problem of short-term time series forecasting based on a local sample of the most recent observations. The subject of the study is a method for constructing a forecast value based on clustering the values of the polynomial prediction sequence. Unlike approaches in which the final forecast is determined either by a single polynomial of fixed degree or by averaging all constructed forecasts, this paper employs the idea of identifying the densest region (cluster) within the set of alternative polynomial predictive values (PPS). The general scheme of the forecasting method based on clustering the values of the polynomial prediction sequence is shown in Fig. 1.

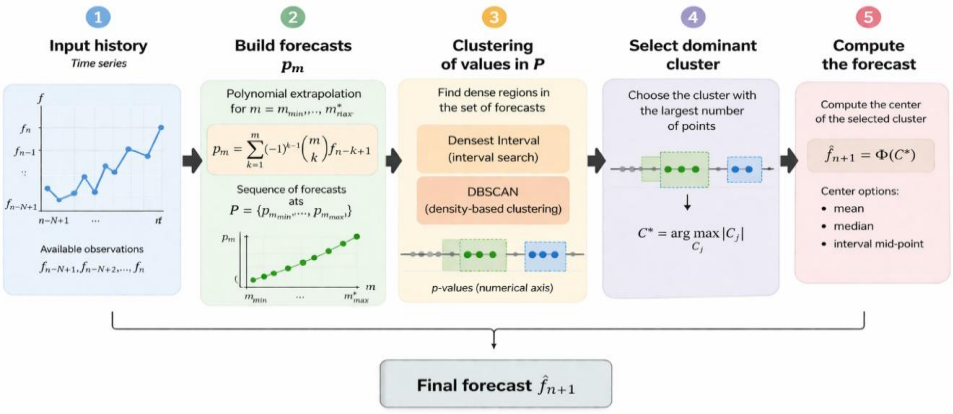

Fig. 1. A prediction algorithm scheme based on clustering

Let the time series $f_i = f(i\Delta),\ i = n, n-1, \ldots, n-N+1,$ be given, where Δ is the discretization step and, N is the number of the most recent available observations used for forecasting. It is required to construct a one-step-ahead forecast of the value $f_{n+1} = f((n+1)\Delta)$. To this end, for each $m = 1, 2, \ldots, k$ a polynomial forecast p_m is constructed, corresponding to the use of the last m values of the series. The value p_m is determined by the formula

$$p_m = f_{n+1}^m = \sum_{k=1}^{m} (-1)^{k-1} \binom{m}{k} f_{n-k+1}. \quad (1)$$

where $\binom{m}{k} = C_m^k$ are the binomial coefficients.

This formula corresponds to a fast method for computing the forecast based on an interpolation polynomial of degree $m-1$ nd is used both in the proposed approach and in the polynomial extrapolation sequence averaging method [8].

Thus, for the same forecasting instant, a PPS sequence is formed: $P = \{p_1, p_2, \ldots, p_m\}$. A distinctive feature of this approach is that each value p_m is an alternative estimate of the future value f_{n+1}, obtained on the basis of different history depths. Therefore, not only the individual value p_m is informative, but also the structure of the entire sequence P: the presence of concentrations, dispersion, local groups, and remote values. Fig. 2 shows an example of such a PPS for 9 points, computed from NFLX stock quotations and their placement on the numerical axis.

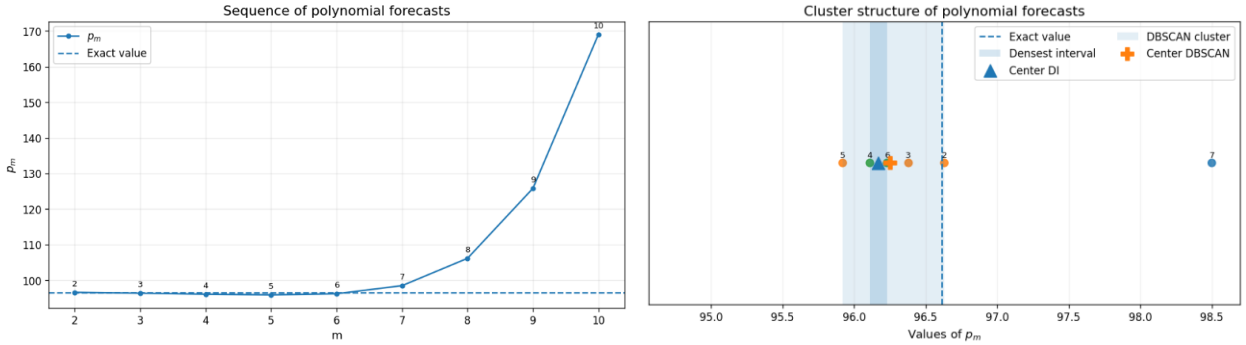

Fig. 2. Values of Polynomial Predictions Sequence and its cluster structure

Unlike approaches in which the final forecast is identified with either a single polynomial of fixed degree or the averaging of a part of the sequence P, this paper employs a clustering-based approach, in which the final value is determined on the basis of the region with the highest concentration of elements of the set P. To analyze the internal structure of the sequence p_m let us consider two adjacent elements of this sequence:

$$p_m = \sum_{k=1}^{m} (-1)^{k-1} \binom{m}{k} f_{n-k+1},$$

$$p_{m+1} = \sum_{k=1}^{m+1} (-1)^{k-1} \binom{m+1}{k} f_{n-k+1}.$$

Then

$$p_{m+1} - p_m = \sum_{k=1}^{m} (-1)^{k-1} \left(\binom{m+1}{k} - \binom{m}{k} \right) f_{n-k+1} + (-1)^m f_{n-m}.$$

Using Pascal's identity

$$\binom{m+1}{k} - \binom{m}{k} = \binom{m}{k-1}$$

we receive

$$p_{m+1} - p_m = \sum_{k=1}^{m} (-1)^{k-1} \binom{m}{k-1} f_{n-k+1} + (-1)^m f_{n-m}. \quad (2)$$

Let's use substitution $j = k-1$:

$$p_{m+1} - p_m = \sum_{j=0}^{m} (-1)^j \binom{m}{j} f_{n-j}. \quad (3)$$

Thus, expression (3) is m-th inverse finite difference at the point, f_n, i.e.

$$p_{m+1} - p_m = \nabla^m f_n.$$

We obtained a recurrence formula for a sequence of polynomial predictions:

$$p_{m+1} = p_m + \nabla^m f_n. \quad (4)$$

Relation (4) shows that each subsequent term of the sequence of polynomial predictions differs from the previous one by the corresponding inverse finite difference. Therefore, the sequence P is an informative object of study, and not just a set of isolated forecast values.

For comparative analysis, a method is used in which the final forecast value is constructed as the arithmetic mean of a certain number of the first elements of a sequence of polynomial forecasts. In the relevant work, it was proposed to search for a forecast in the form of averaging a subsequence $\{p_1, p_2, \ldots, p_m\}$, and the optimal value m determined by minimizing the internal deviation between the forecast at the point $n+1$ and the estimate constructed for the point $n+2$ [8].

According to this formulation, the averaged forecast has the form

$$\hat{f}_{n+1}^{(avg,m)} = \frac{1}{m} \sum_{i=1}^{m} f_{n+1}^{(i)}.$$

In the cited work, an equivalent analytical representation of this averaging procedure in terms of time-series values is also provided, and the internal characteristic

$$\Delta_m = \frac{1}{m} \left| \hat{f}_{n+2}^{(m)} - f_{n+1}^{(m)} \right|$$

is used to assess the quality of the choice of m.

After that, the optimal value of m is determined as

$$m^* = \arg\min_{m \in \{1,2,\ldots,n\}} \Delta_m.$$

Then, the final forecast of the averaging method is defined as

$$\hat{f}_{n+1}^{(avg)} = \hat{f}_{n+1}^{(avg,m^*)}.$$

It is precisely this approach that is further used as the baseline comparison method for evaluating the effectiveness of cluster-based forecast formation. Its key idea consists in selecting an initial fragment of the polynomial prediction sequence. In the approach proposed in this study, the selection is performed not according to the index position of the elements, but according to their spatial arrangement on the numerical axis. Regions with the highest concentration of p_m values are identified, and the final forecast is constructed on the basis of the central characteristic of this region.

Let C be a subset of the values of P corresponding to the identified cluster or the densest interval. Then the final forecast is defined as

$$f_{n+1} = \Phi(C),$$

where $\Phi(C)$ is the rule for computing the cluster center. The software implementation provides three options:

1) arithmetic mean:

$$\Phi(C) = \frac{1}{|C|} \sum_{x \in C} x;$$

2) median:

$$\Phi(C) = \text{median}(C);$$

3) interval midpoint:

$$\Phi(C) = \frac{L_C + R_C}{2},$$

where L_C and R_C are the left and right boundaries of the identified cluster or interval.

Two clustering approaches are implemented in the study: the Densest Interval (DI) method and DBSCAN (Density-Based Spatial Clustering of Applications with Noise).

The first clustering approach, DI, is based on searching for the densest interval among the values of the sequence P. For this purpose, the elements of the set are ordered in ascending order:

$$v_1 \le v_2 \le \ldots \le v_s.$$

Then all possible contiguous subsequences

$$[v_l, v_s], \quad 1 \le l < r \le s,$$

containing at least q points are considered. For each such interval, its width is computed as

$$w_{l,r} = v_r - v_l$$

as well as the classical density measure

$$\rho_{l,r} = \frac{r - l + 1}{\max(v_r - v_l, \varepsilon_0)},$$

where $\varepsilon_0 > 0$ is a small number introduced to avoid division by zero.

The optimal interval is defined as

$$(l^*, r^*) = \arg\max_{l,r} \rho_{l,r}.$$

After that, the cluster is formed as

$$C_{DI} = \{v_{l^*}, v_{l^*+1}, \ldots, v_{r^*}\},$$

and the forecast value is determined as

$$\hat{f}_{n+1}^{(DI)} = \Phi(C_{DI}).$$

The second clustering approach is based on the DBSCAN algorithm. The set P is considered as a one-dimensional sample

$$P = \{p_1, p_2, \ldots, p_m\}.$$

The algorithm uses two parameters:

- ε – the radius of the local neighborhood;
- $MinPts$ – the minimum number of points in the ε-neighborhood.

Points that have a sufficient number of neighbors within radius ε form dense regions, which are interpreted as clusters. Points that do not belong to any such region are treated as noise. As a result, a set of clusters is formed:

$$C_{DB} = \{C_1, C_2, \ldots, C_K\}.$$

As the final cluster, the one containing the largest number of elements is selected:

$$C_{DB}^* = \arg\max_{C_j \in C_{DB}} |C_j|.$$

Then the final forecast is determined as

$$\hat{f}_{n+1}^{(DB)} = \Phi(C_{DB}^*).$$

To investigate the properties of the proposed approach, three types of data are used. In controlled experiments, the values of the following functions are analyzed:

$$y = \cos(hx),$$

$$y = \cos(h \ln x),$$

$$y = e^{hx},$$

where h is a function parameter.

To study the behavior of the algorithms on noisy sequences, random data generated from a normal distribution are used:

$$y_i \sim \mathrm{N}(0, \sigma^2),$$

where σ determines the noise level.

To assess the practical applicability of the method, real financial time series loaded from Excel files are used. As experimental data, intraday quotations of Netflix (NFLX) stock over one trading day (08.12.2025) with a discretization step of $\Delta = 30$ minutes (fields: Timestamp, Close) are considered. In addition, the effectiveness of the method was tested on daily quotations of Apple (AAPL) stock over the period 2019–2025. From the selected Close column, a sequence of values is formed in which the last observation is used as the reference value, while the preceding observations are used as the history for forecast construction.

For each approach (DI, DBSCAN), the absolute error is computed as

$$\Delta_{abs} = \left| f_{n+1} - \hat{f}_{n+1} \right|$$

and the relative error as

$$\Delta_{rel} = \frac{\left| f_{n+1} - \hat{f}_{n+1} \right|}{\left| f_{n+1} \right|} \cdot 100\%.$$

In addition to numerical characteristics, the software implementation employs visual analysis tools, including a plot of the original time series with the true and forecast values, a plot of the p_m values with point numbering, a plot of the cluster analysis on the numerical axis with cluster boundaries highlighted, and a histogram of the distribution of the p_m values. It is also recorded separately which particular p_m values are included in the identified clusters. This makes it possible to analyze not only the accuracy of the final forecast, but also the internal structure of the set of alternative polynomial forecasts.

Experiments

Experimental studies were conducted to evaluate the effectiveness of the proposed forecasting method based on clustering the values of the polynomial prediction sequence and to compare it with the forecasting method based on averaging the polynomial extrapolation sequence [8]. In all experiments, a unified scheme was used for constructing the polynomial prediction sequence $P = \{p_1, p_2, \ldots, p_m\}$ (1) over a given range of m values. After that, two cluster analysis approaches were applied to this set: the Densest Interval (DI) method and the DBSCAN method.

The study was carried out for three classes of data:

- deterministic functions;
- stochastic (random) sequences;
- real financial time series.

For each type of data, the prediction sequence p_m was constructed for varying parameters $m \in [m_{min}, m_{max}]$, after which its structure, distribution density, and behavior were analyzed.

The following functions were used as test functions:

$$y = \cos(hx), \quad y = \cos(h \ln x), \quad y = e^{hx}.$$

Fig. 3 shows a fragment of the graph of the function $y = \cos(hx)$ and the graph of the polynomial prediction sequence for the following given parameters: $x_0 = 1$ (initial value), $N = 30$ (number of historical points), $h = 0,3$ (parameter), $\Delta = 1$ (step), $m_{\min} = 2, m_{\max} = 20$ (minimum and maximum values of the number of points for the PPS sequence), $\varepsilon = 0,5$ (radius of the local neighborhood for DBSCAN).

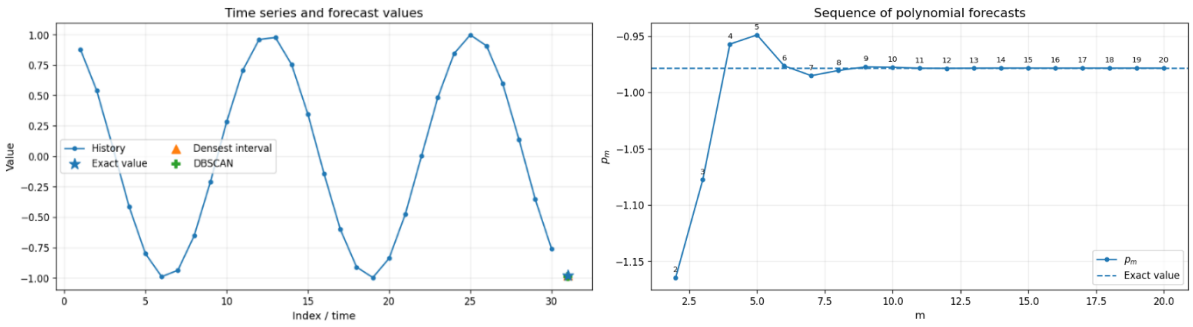

Fig. 3. Plot of function $y = \cos(hx)$ and its Polynomial Predictions Sequence (PPS).

For deterministic functions, as can be seen from Fig. 3, the sequence of polynomial predictions p_m demonstrates structured and predictable behavior, namely:

- local extrema are observed;
- random noise is absent;
- the values are concentrated within a limited interval.

This is explained by the fact that the input data have a regular structure, while polynomial extrapolation effectively reconstructs the local smoothness of the function.

The distribution of PPS values on the numerical axis, the cluster structure, the histogram, and the KDE (Kernel Density Estimation) density estimate are shown in Fig. 4. For better visualization of point placement, a vertical displacement parameter (jitter) was used.

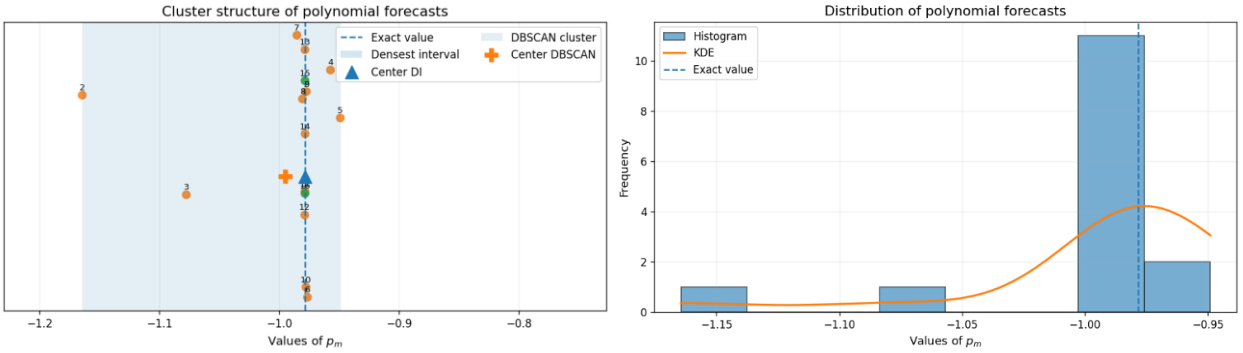

Fig. 4. Cluster structure and distribution of PPS for $y = \cos(hx)$.

The results of the cluster analysis obtained using the DI and DBSCAN methods are presented in Table 1.

Table 1

Results of cluster analysis using DI and DBSCAN methods

Method	Exact Value	Predicted Value	Absolute Error	Relative Error (%)	Left Bound	Right Bound	Cluster Count	Cluster Width	Cluster Mode
Densest Interval (DI)	-0,97845	-0,97845	0,00001	0,00065	-0,97846	-0,97843	2	0,00003	mean
DBSCAN	-0,97845	-0,99442	0,01597	1,63220	-1,16445	-0,94898	15	0,21547	mean

The stochastic data used in the experiments were generated with a pseudorandom number generator as a sample from a normal (Gaussian) distribution with zero mean and a prescribed standard deviation σ, which determined the noise level in the series. The generation of values is described by the relation

$$y_i \sim N(0, \sigma^2)$$

To ensure reproducibility of the results, a fixed initial state of the generator (seed) was used. In each experiment, a sequence of $N+1$ values was generated, where the first N elements were used as the history

$$\{f_0, f_1, \ldots, f_{N-1}\},$$

and the last value f_N was treated as the reference (true) value for evaluating forecast accuracy. The input data were generated without additional assumptions regarding trend or seasonality, which made it possible to investigate the behavior of the polynomial prediction sequence p_m under the absence of deterministic structure and to assess the ability of the clustering method to detect internal regularities even in random data.

Fig. 5 shows the graph of the stochastic data and the graph of the polynomial prediction sequence for the following specified parameters: $N = 30$ (number of historical points), $Seed = 42$ (reproducibility parameter of the pseudorandom distribution), $\sigma = 0,5$ (noise level), $m_{\min} = 2, m_{\max} = 20$ (minimum and maximum number of points for the PPS sequence), $\varepsilon = 0,5$ (radius of the local neighborhood for DBSCAN).

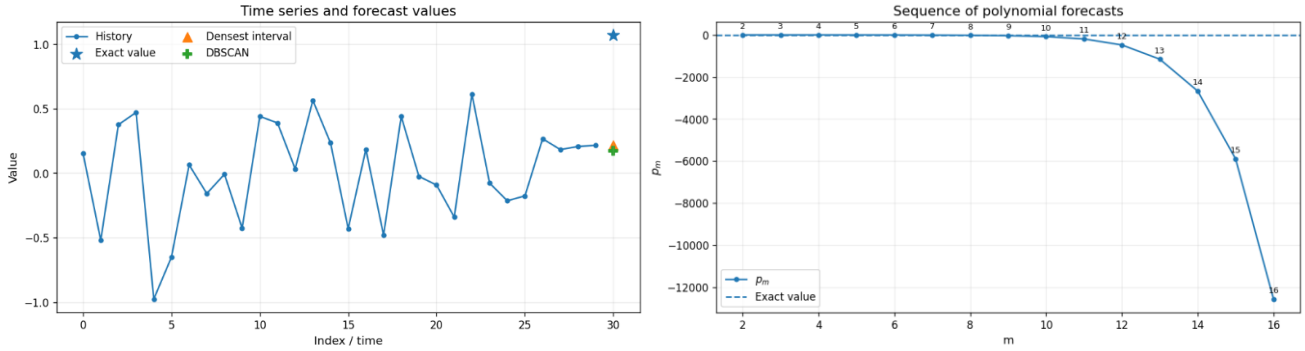

Fig. 5. Graph of stochastic data and its Polynomial Predictions Sequence (PPS).

Despite the random nature of the input series, the sequence p_m demonstrates features of deterministic behavior. In particular:

- a chaotic structure is absent;
- intervals of monotonicity are preserved;
- local extrema arise;
- clustering of values is observed.

The distribution of PPS values on the numerical axis, the cluster structure, the histogram, and the KDE (Kernel Density Estimation) density estimate are shown in Fig. 6. For better visualization of the arrangement of points, a vertical displacement parameter (jitter) was used.

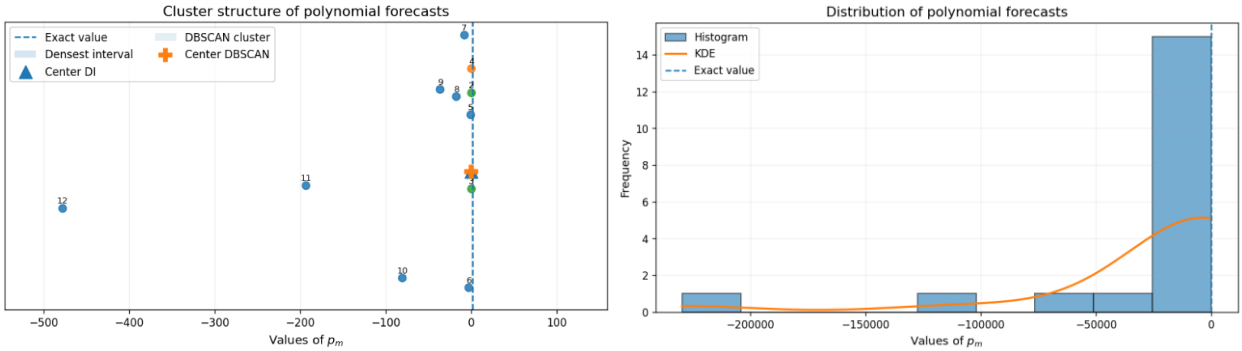

Fig. 6. Cluster structure and distribution of PPS for stochastic data

The results of the cluster analysis of the polynomial prediction sequence for stochastic data obtained using the DI and DBSCAN methods are presented in Table 2.

Table 2

Results of cluster analysis using DI and DBSCAN methods for stochastic data.

Method	Exact Value	Predicted Value	Absolute Error	Relative Error (%)	Left Bound	Right Bound	Cluster Count	Cluster Width	Cluster Mode
Densest Interval (DI)	1,070824	0,217155	0,853669	79,72079	0,209855	0,224455	2	0,01460	mean
DBSCAN	1,070824	0,174162	0,896661	83,73567	0,088178	0,224455	3	0,13628	mean

For experiments on real financial data, an intraday dataset of Netflix (NFLX) stock quotations was used, provided in Excel format with a 30-minute discretization interval. The main analyzed indicator was the Close parameter, i.e., the closing price of the corresponding 30-minute interval. This variable directly reflects the final market state within each observation. Such a dataset is representative for short-term forecasting tasks, since it combines local fluctuations, microtrends, and random variations characteristic of intraday stock market dynamics. The use of 30-minute intervals makes it possible to analyze the series over a sufficiently short time horizon, where polynomial extrapolation and subsequent PPS clustering can reveal local regularities.

Fig. 7 shows the graph of NFLX stock quotations and the graph of the polynomial prediction sequence for the following specified parameters: $N = 31$ (number of historical points), $m_{\min} = 2, m_{\max} = 12$ (minimum and maximum numbers of points for the PPS sequence), $\varepsilon = 0,5$ (radius of the local neighborhood for DBSCAN).

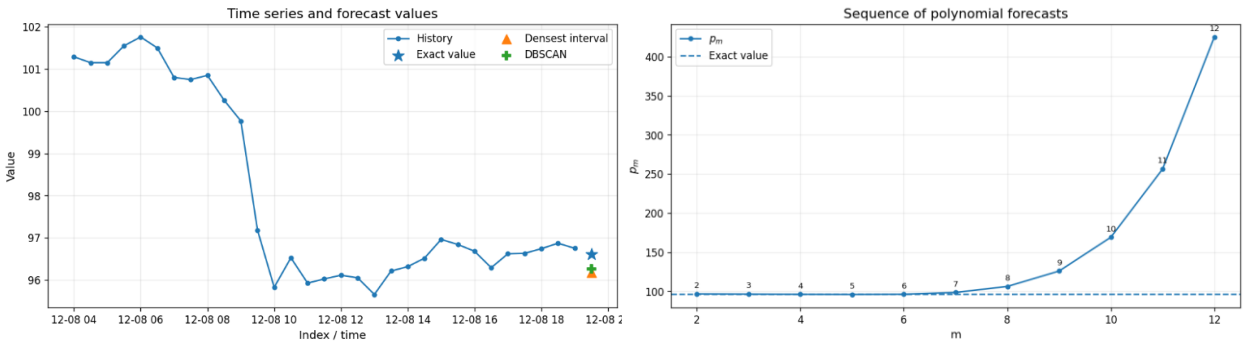

Fig. 7. Graph of stock data (NFLX, 30 min) and its Polynomial Predictions Sequence (PPS).

For stock market quotations, the PPS sequence has its own specific features, namely:

- it has a mixed nature, being partly smooth and partly noisy;
- dominant clusters are clearly distinguished;

DBSCAN consistently identifies the most representative group of values (Fig. 8).

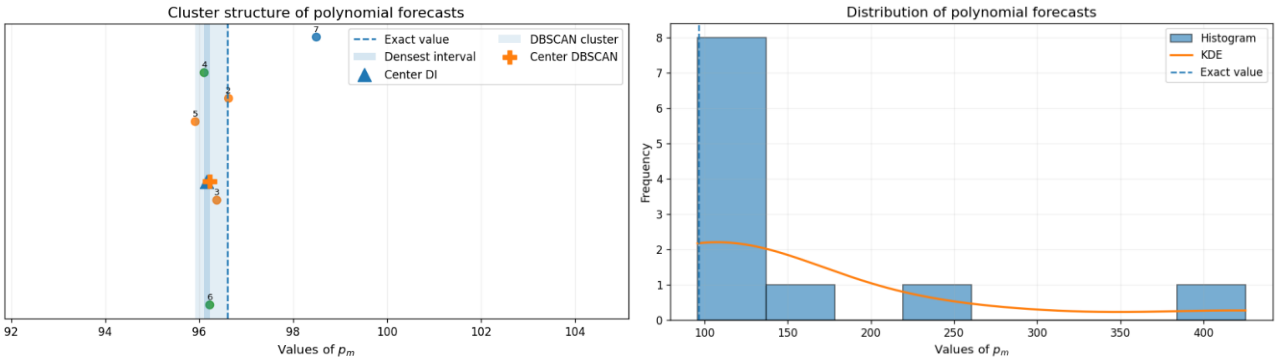

Fig. 8. Cluster structure and distribution of PPS for stock data (ticker NFLX, 30 min)

The results of the cluster analysis of the polynomial prediction sequence for stochastic data obtained using the DI and DBSCAN methods are presented in Table 3.

Table 3

Results of cluster analysis using DI and DBSCAN methods for stock data (ticker NFLX, 30 min).

Method	Exact Value	Predicted Value	Absolute Error	Relative Error (%)	Left Bound	Right Bound	Cluster Count	Cluster Width	Cluster Mode
Densest Interval (DI)	96,6101	96,1695	0,4406	0,45606	96,1096	96,2294	2	0,1198	mean
DBSCAN	96,6101	96,2294	0,3807	0,39406	95,9195	96,6298	5	0,7103	mean

The obtained results show that, regardless of the type of input data, the sequence of polynomial predictions p_m possesses an internal structure manifested in the form of local extrema and dense regions of values. This substantiates the expediency of applying clustering methods to determine the final forecast. The most stable results were obtained using the DBSCAN algorithm, which makes it possible to identify dominant groups of values even in the presence of noise.

Results and Discussion

The conducted experiments showed that the sequence of polynomial predictions $P = \{p_1, p_2, \ldots, p_m\}$, is characterized, in a certain sense, by deterministic behavior even for stochastic data. In particular, for deterministic functions, convergence is observed in most cases. In the case of divergence, it is strictly monotonic.

For stochastic data, specifically for samples drawn from a normally distributed population, the sequence P exhibits a clearly expressed deterministic behavior characteristic of periodic oscillations with increasing amplitude; local concentrations, intervals of relative monotonicity, and distinct dominant regions on the numerical axis are observed. A similar behavior was exhibited by the sequence for real stock market data of Netflix shares (for the Close parameter). In this case, it contained compact groups of close forecast values. This created the prerequisites for the application of cluster analysis. In numerical terms, the values of errors and metrics are presented in Tables 1–3. The proposed approach was compared with the baseline method of averaging the polynomial extrapolation sequence [7].

A one-step-ahead forecast was performed for the last Close value in the file with 30-minute NFLX quotations. The first 31 Close values were taken as the historical data, while the last value at timestamp 2025-12-08 19:30:00 was

assumed to be unknown and was used only for validation; its actual value is 96.6101. For the averaging method, the algorithm described in [7] was used: for each m the polynomial forecast f_{n+1}^{m} was computed, then the estimate $\hat{f}_{n+2}^{m}$, and then the deviation

$$\Delta_m = \frac{1}{m}\left|\hat{f}_{n+2}^{m} - f_{n+1}^{m}\right|,$$

was selected

$$m^* = \underset{m\in\{1,2,..,n\}}{\operatorname{argmin}} \Delta_m .$$

For comparison with the clustering methods, the following parameters were used: $N = 31$ (number of historical points), $m_{\min} = 2, m_{\max} = 15$ (minimum and maximum numbers of points for the PPS sequence), $\varepsilon = 3$ (radius of the local neighborhood for DBSCAN). The following results were obtained (Table 4):

Table 4

Comparison results of cluster methods with the algorithm for averaging PPS.

Method	Exact Value	Predicted Value	Absolute Error	Relative Error (%)	Left Bound	Right Bound	Cluster Count	Cluster Width	Cluster Mode
Densest Interval (DI)	96,61010	96,16950	0,44060	0,45606	96,10960	96,2294	2	0,1198	mean
DBSCAN	96,61010	96,62788	0,01778	0,01841	95,91950	98,4993	6	2,5798	mean
Averaging PPS	96,61010	96,62788	0,01778	0,01841	–	–	–	–	–

Thus, for the reference point 96,6101, the best results were obtained by the averaging method and DBSCAN, whereas the densest interval method produced the largest error. This comparison refers specifically to the last Close value and to exactly these parameter settings of the software implementation. For another range of m or another value of ε the DBSCAN result may change. This is clearly seen from the comparison with Table 3: for $m_{\min} = 2, m_{\max} = 12$ and $\varepsilon = 0,5$ the relative error was larger and amounted to 0,39406%.

Although the densest interval algorithm is specifically designed to identify the region of maximum local concentration of forecast values, in practical experiments it may yield inferior results compared with DBSCAN. This can be explained by the fact that the density criterion $\rho = \frac{count}{width}$ favors very narrow local groups, which do not necessarily represent the dominant trend of the entire set of forecast values. Instead, DBSCAN better captures the structure of the principal concentration of points and therefore, in some cases, provides a more accurate final forecast.

The obtained results can be explained by the fact that the individual elements of the sequence are not independent forecasts, but are formed on the basis of a common local history and are interconnected through the structure of finite differences. That is why, even when the polynomial degree changes, the forecasts do not disperse chaotically, but tend to form distinct regions of concentration. If such a region corresponds to a locally stable part of the prediction sequence, then its center turns out to be a more representative estimate of the future value than either an individual forecast of fixed degree or the averaging of elements selected only according to their position in the sequence. In essence, clustering reveals not merely the "average" value of the forecasts, but the dominant geometric structure of the set P.

The advantages of the proposed solution are ensured by several of its features:

1) he method is not limited to selecting a single "best" polynomial of fixed degree, but uses the entire sequence of alternative forecasts obtained from the same local fragment of the series;

2) within the method, the spatial arrangement of the values p_m on the numerical axis is taken into account, that is, the actual concentration of forecasts rather than merely their index position in the sequence;

3) clustering makes it possible to reduce the influence of individual anomalous or unstable values that may arise for higher polynomial degrees;

4) the method is sufficiently flexible, since it allows different ways of forming the final forecast – through the densest interval or through DBSCAN – as well as different ways of determining the cluster center.

Compared with the known method of averaging the polynomial extrapolation sequence [7], the advantage of the proposed approach lies in the fact that it refines the mechanism for selecting PPS elements: not according to the rule "take the first m elements," but according to the rule "take those elements that form the dominant region of forecast values." This constitutes the main scientific novelty and practical value of the study.

At the same time, the results of the study also have a number of limitations that should be taken into account in practical applications and in further theoretical developments. First of all, the proposed method is focused on short-term one-step forecasting; its behavior in multi-step forecasting requires separate investigation. A second important

limitation is the parameterization of the clustering procedures: for DBSCAN, the choice of ε is crucial, while for the densest interval method, the choice of the minimum number of points is essential. It should also be taken into account that, for real financial data, intraday series may depend substantially on volatility, the news background, and microstructural market effects, which are not directly incorporated into the polynomial model itself.

Among the shortcomings of the study is the fact that, at the current stage, the experimental validation has been carried out on a limited set of scenarios and does not cover the full range of financial instruments, market regimes, and time horizons. Another drawback is the absence, at the present stage, of strict theoretical criteria that would unambiguously establish under which conditions the clustering-based approach is guaranteed to outperform the baseline averaging method.

Further development of the research appears advisable in several directions:

1) transition from individual test examples to batch analysis on a large number of sequential windows of real financial series. Such an approach would make it possible to provide a statistically substantiated answer regarding the advantages of one or another aggregation method.

2) adaptive selection of clustering parameters, primarily ε for DBSCAN, based on the internal characteristics of the set P.

3) further theoretical investigation of the form and properties of the polynomial prediction sequence, in particular the conditions for the emergence of clusters, local extrema, monotonicity intervals, and convergence effects.

4) extension of the research to multi-step forecasting.

These directions constitute a natural continuation of the study, since they make it possible to improve the practical robustness of the method and deepen its mathematical justification.

Overall, the results of the study indicate that clustering the values of the polynomial prediction sequence is a justified and promising way to refine the mechanism for selecting PPS elements in forming the final forecast. The proposed approach is a further development of the forecasting method based on averaging the sequence of polynomial predictions.

Conclusions

This paper develops a method for short-term time series forecasting based on clustering the values of the polynomial prediction sequence. Unlike approaches in which the final forecast value is determined either by a single polynomial of fixed degree or by averaging the first elements of the PPS sequence, the proposed method is based on analyzing the spatial structure of the set of alternative forecasts and identifying its dominant region. This made it possible to refine the mechanism for selecting PPS elements in the formation of the final forecast.

In the course of the study, it was established that the sequence of polynomial predictions is characterized by the presence of concentration intervals, which confirms the expediency of applying cluster analysis.

The experimental results showed that the use of the densest interval method and the DBSCAN algorithm makes it possible to generate more accurate forecast estimates.

The obtained results confirmed the practical applicability of the approach to short-term forecasting of intraday stock market series, in particular Netflix stock data for the Close parameter.

Further research should be directed toward adaptive selection of clustering parameters, batch analysis on a large number of financial series, and theoretical investigation of the conditions under which the cluster structure of the polynomial prediction sequence is formed.

ADDITIONAL INFORMATION

AUTHOR CONTRIBUTIONS

Conceptualization, Y.T. and O.K.; methodology, Y.T.; software, O.K.; validation, Y.T. and O.K.; formalanalysis, Y.T.; investigation, Y.T and O.K.; data curation, O.K.; writing–original draft preparation, O.K.; writing–review and editing, Y.T.; visualization, O.K.; supervision, Y.T.; project administration, Y.T.; funding acquisition, O.K. All authors have read and agreed to the published version of the manuscript.

DECLARATION ON THE USE OF GENERATIVE ARTIFICIAL INTELLIGENCE TOOLS

The authors confirm that they did not use artificial intelligence technologies in creating the submitted work.

1. Chen, W., Hussain, W., Cauteruccio, F., Zhang, X. Deep Learning for Financial Time Series Prediction: A State-of-the-Art Review of Standalone and Hybrid Models // CMES – Computer Modeling in Engineering & Sciences. 2023. Vol. 139, no. 1. P. 187–224. https://doi.org/10.32604/cmes.2023.031388.

2. Nazareth, N., Ramana Reddy, Y. V. Financial applications of machine learning: A literature review // Expert Systems with Applications. 2023. Vol. 219. Art. 119640. https://doi.org/10.1016/j.eswa.2023.119640.

3. Wang, X., Hyndman, R. J., Li, F., Kang, Y. Forecast combinations: An over 50-year review //

International Journal of Forecasting. 2023. Vol. 39, no. 4. P. 1518–1547. https://doi.org/10.1016/j.ijforecast.2022.11.005.

4. Emami, H., Dang, X.-H., Shah, Y., Zerfos, P. Modality-aware Transformer for Financial Time series Forecasting // arXiv. 2024. Art. arXiv:2310.01232. https://doi.org/10.48550/arXiv.2310.01232.

5. Li, J.-C., Sun, L.-P., Wu, X., Tao, C. Enhancing financial time series forecasting with hybrid Deep Learning: CEEMDAN-Informer-LSTM model // Applied Soft Computing. 2025. Vol. 177. Art. 113241. https://doi.org/10.1016/j.asoc.2025.113241.

6. Yang, H., Chen, Y., Chen, K., Wang, H. Temporal-spatial dependencies enhanced deep learning model for time series forecast // International Review of Financial Analysis. 2024. Vol. 94. Art. 103261. https://doi.org/10.1016/j.irfa.2024.103261.

7. Turbal, Y., Shlikhta, G., Turbal, M., Turbal, B. The polynomial forecasts improvement based on the algorithm of optimal polynomial degree selecting // Eastern-European Journal of Enterprise Technologies. 2023. Vol. 5, no. 4 (125). P. 34–42. https://doi.org/10.15587/1729-4061.2023.289292.

8. Turbal, Y., Turbal, M., Kubai, O., Smirnov, D., Melnychuk, M. Forecasting method based on averaging a polynomial extrapolation sequence. Modeling, Control and Information Technologies: Proceedings of International Scientific and Practical Conference. No. 8. P. 326–329. https://doi.org/10.31713/MCIT.2025.102.

9. Turbal, Y., Bomba, A., Turbal, M., Turbal, B., Smirnov, D. Forecasting Algorithm Based on Intellectual Analysis of Polynomial Extrapolation. Lecture Notes on Data Engineering and Communications Technologies, 244, 98–113. DOI: 10.1007/978-3-031-88483-2_5.

Юрій ТУРБАЛ, Олександр КУБАЙ
Національний університет водного господарства та природокористування

МЕТОД ПРОГНОЗУВАННЯ НА ОСНОВІ КЛАСТЕРИЗАЦІЇ ПОЛІНОМІАЛЬНОЇ ЕКСТРАПОЛЯЦІЙНОЇ ПОСЛІДОВНОСТІ

Прогнозування часових рядів є важливою задачею інтелектуального аналізу даних, особливо в умовах коротких вибірок, локальної нестаціонарності, шуму та підвищеної чутливості до зовнішніх збурень. Такі властивості характерні, зокрема, для фінансових часових рядів, де на малих інтервалах спостереження поєднуються локальні тренди, випадкові коливання та різкі зміни динаміки. Одним із перспективних підходів до короткострокового прогнозування є поліноміальна екстраполяція. Проте використання поліномів різного порядку для одного й того самого фрагмента ряду формує множину альтернативних прогнозних значень, що ускладнює вибір підсумкового прогнозу.

У роботі запропоновано метод короткострокового прогнозування на основі кластеризації значень поліноміальної екстраполяційної послідовності. Для фрагмента часового ряду формується послідовність поліноміальних прогнозів p_m у заданому діапазоні порядків, після чого до множини цих значень застосовується кластерний аналіз. Для виділення домінантної області прогнозів використано метод найщільнішого інтервалу та алгоритм DBSCAN, а фінальне прогнозне значення визначається як центральна характеристика знайденого кластера. Ефективність запропонованого підходу порівнюється з методом прогнозування на основі усереднення поліноміальної екстраполяційної послідовності.

Експериментальні дослідження проведено на детермінованих функціях, стохастичних послідовностях та реальних внутрішньоденних біржових даних акцій Netflix за параметром Close. Встановлено, що послідовність поліноміальних прогнозів має внутрішню структуру у вигляді локальних екстремумів, інтервалів концентрації та віддалених значень, що обґрунтовує доцільність її кластеризації. Наукова новизна роботи полягає в уточненні механізму вибору елементів PPS шляхом переходу від індексного усереднення до структурного аналізу просторового групування прогнозних значень. Практична значущість полягає у підвищенні стійкості короткострокового прогнозування фінансових часових рядів.

Ключові слова: поліноміальна екстраполяція; короткострокове прогнозування; часові ряди; послідовність поліноміальних прогнозів; кластеризація прогнозів; DBSCAN; фінансові часові ряди; біржові дані.