ESTIMATION OF NONLINEAR RELATIONSHIPS IN DYNAMICS OF PUBLIC COMPANY MARKET CAPITALIZATION IN THE TASK OF DETERMINATION OF MARKET SYNTHETIC INDEX

A PREPRINT

Anastasia Nasykhova **Elizaveta Kovtun** **Semen Budennyy**

2023

ABSTRACT

Studying the dynamics of the market capitalization of public companies is of great importance in finance and economics. This allows you to analyze and predict the behavior of financial markets and investments, as well as identify trends and patterns in the stock market. In this scientific paper, a study was conducted on the possibility of using hierarchical clustering based on DTW (Dynamic Time Warping) to improve the predictions of the TFT (Temporal Fusion Transformer) model in financial time series. To assess the quality of predictions, the MAPE metric (Mean Absolute Percentage Error) was used. Experimental results have shown that using time series from a single cluster when training a model allows achieving more accurate predictions in comparison with a model trained on only one time series. The research makes a valuable contribution to the field of time series analysis and predictive modeling, introducing a new method to improve the accuracy of predictive models.

Keywords Time series similarity · Stock price prediction · Time series clustering · Synthetic index

Contents

1 Introduction

Identifying relationships between companies is important for understanding how changes in one company can affect other companies in the industry or in the market as a whole. Understanding these relationships can help make more informed investment decisions, as investors can better predict how markets will behave in the future. In this regard, there is increasing interest in an in-depth study of the relationships between companies, in particular, to identify nonlinear relationships that are characterized by complex patterns and require the use of special analysis methods. This article presents the results of a study of nonlinear relationships between companies and discusses their possible practical application.

Contribution. Within the framework of this research the following work was done:

- Conducted a thorough review of the literature on clustering algorithms for time series analysis and methods for their prediction;
- Created a synthetic index based on the evaluation of clustering results;
- Implemented TFT model for predicting financial time series data;
- Applied a synthetic index to improve the quality of prediction.

2 Literature Review

When working with time series, many standard machine learning difficulties arise: high dimensionality of input data, correlation in data, missing values. However, there are a couple of additional difficulties in the task of clustering sequences. Firstly, there may be a different number of counts in the rows. Secondly, when working with sequential data, we have more freedom in determining the similarity of one object to another, because the way the data changes over time also contributes.

2.1 Distance between time series

Time series similarity measures and time series indexing [2] This article discusses ways to calculate distances between time series. Most of them have a key drawback: the metric does not take into account the shift in the data. This means that if in one time series there is a sequence from another series that is similar in structure, but it is located in a different time interval, the metric will count these series as different. And since it is important to find a common pattern of the behavior of a series for the task of determining a synthetic index, such metrics cannot be used. Fortunately, the article also describes a metric that solves this problem – DTW (Dynamic Time Warping).

Recent Techniques of Clustering of Time Series Data: A Survey [5] In a later article, the authors explore the latest developments in the field of time series clustering. The most important thing is that researchers pay special attention to the correlation of distance metrics with clustering algorithms. The paper presents a variety of approaches to working with time series. Most often, for financial data, a simple Euclidean distance is chosen as a metric, and modified K–Means is used as an algorithm. By it, the writers mean a basic algorithm adapted to work with sequential data.

2.2 Clustering

Clustering Time Series Data — A Survey [3] The article discusses the main approaches to clustering time series. In general, time series clustering can be divided into 2 types:

• Feature based approach in which objects are clustered based on statistical data collected from a time series;

• Raw data based approach in which clustering is applied to time series vectors without any transformations.

Both methods have proven themselves well when working with financial data, but the first approach is rarely used, because it does not take into account the hidden connections between time series.

Time-series clustering - A decade review [1] The authors provide a comprehensive review of time series clustering research over the past decade. They focus on clustering methods for data from different areas. They confirm the studies cited earlier about the fact that standard algorithms are well suited for financial time series. An interesting feature of this article is that in addition to successful solutions, researchers cite a number of unsuccessful experiments, as well as talk about possible causes of failures. The most common reasons are the features of sequential data: high dimensionality, high noise level in the data, strong correlations between values.

2.3 Stock price prediction

Many forecasting problems are related to time analysis. In our case, time series data can be defined as a chronological sequence of observations for the stock price. Most of the ways to solve this problem can be divided into two classes: linear and nonlinear models.

Stock price prediction using LSTM, RNN and CNN-sliding window model [6] In 2017, researchers from India applied the most popular linear algorithm and several different neural network architectures to predict stock prices of some Indian companies. After conducting a series of experiments, they showed that deep learning algorithms are able to identify hidden patterns and underlying dynamics in data through the process of self-learning, which leads to better prediction of time series.

Recently, in the field of temporal data processing, there has been a growing interest in the use of recurrent neural network (RNN) and transformer models in combination with fusion methods for modeling temporal dependencies. As part of this trend, a new model has been proposed, called the "Temporal Fusion Transformer Model" (TFTM), in which information from several sources is efficiently combined to predict time series.

Temporal Fusion Transformers for interpretable multi-horizon time series forecasting [4] The authors of the article present a TFTM model for multidimensional time series forecasting. TFTM is a hybrid model that combines the functionality of transformers and unifying methods such as gated-fusion and decomposition of time series into trend, seasonality and noise. To test the model, experiments were conducted on several data sets, including energy consumption, stock prices and the number of passengers in air transport. The results showed that TFTM is superior to standard time series forecasting models such as ARIMA, LSTM and GRU.

3 Methodology

This section describes the general methodology of the study, including clustering algorithms and a prediction model.

3.1 Clustering algorithm

Clustering is the process of grouping similar time series together to identify underlying patterns and structures, this study uses a hierarchical clustering algorithm, K-Means and Feature based approach.

K-Means clustering. K-Means is a clustering algorithm used to split a set of objects into a given number of clusters. It is based on minimizing the sum of squared distances between objects and cluster centers. To apply the algorithm to time series, you need to choose a metric by which the distances between the series will be estimated. Usually Euclidean distances or DTW are used.

After initializing the initial centroids of k clusters, the algorithm goes through the following steps:

1. Each time series refers to the nearest cluster by the distance to its centroid.
2. For each cluster, a new centroid is calculated as the average of all rows in this cluster.
3. Repeat steps 1 and 2 until the centroids stop changing or the maximum number of iterations is reached.

One of the main parameters of the algorithm is the choice of initial centroids. It is a good practice to use random points from a common dataset, as well as to conduct several runs with different initial centroids in order to more accurately determine the optimal number of clusters and avoid falling into the local optimum.

The limitations of K-Means include the need to specify a prefixed number of clusters, as well as sensitivity to outliers and heterogeneity of clusters in size. However, with the right choice of parameters and understanding of the features of the data, K-Means can be an effective tool for clustering time series.

Hierarchical clustering. Hierarchical clustering is one of the most common time series clustering algorithms. This algorithm treats all time series as a set of points in n-dimensional space and builds a hierarchical cluster structure where each cluster consists of more similar data series.

The clustering algorithm can be divided into two stages:

1. Creating a distance matrix. The first stage is to create a matrix of distances between time series. The distance between time series can be determined in various ways.
2. Clustering. Each time series forms its own cluster. Then the algorithm finds the smallest distance between all possible pairs of clusters and combines them within a new cluster. This process continues until all time series are combined into a single cluster or until the number of clusters specified by the user is reached.

Feature based approach. The feature based approach to time series clustering is to use a set of features describing each time series. In the future, the statistics obtained are clustered by the usual K-means algorithm, as simple tabular data.

3.2 Forecasting model

After clustering the time series, the Temporal Fusion Transformer (TFT) model from the darts package was used to predict time series.

TFT Model. The Temporal Fusion Transformer (TFT) is an attention-based neural network architecture that can incorporate multiple sources of information, such as temporal features and calendar events, into the time series forecasting process. The basic idea is that TFT predicts the values of time series at different points in time simultaneously, taking into account the context, using the transformer method as the basis for the model architecture.

A feature of TFT is that it uses a hybrid neural network architecture that combines convolutional, recurrent and transformer blocks. A time series is received at the input of the model, which is then converted into a multidimensional context vector containing information about the sequence of time series signals.

Next, multidimensional decoding is carried out, which is carried out using transformer blocks, and each block takes as input not only the output of the previous block, but also the context vector. This allows the model to take into account the context when generating forecasts for each point in time in the future and use metrics for estimating forecasts at all time steps for optimization.

TFT uses additional mechanisms to improve the quality of forecasts, including autoregressive and multitasking training. Autoregressive learning is used to improve the generation of forecasts for different time intervals, and multitasking training allows the model to cope with the simultaneous prediction of several time series using general information about the data.

Thus, the TFT model is an effective, flexible and powerful tool for predicting time series. It has shown high quality predictions on several real data sets, and can also be applied in various fields, including economics, finance, medicine and transport.

3.3 Problem Statement

There is a hypothesis that the behavior of time series of stock prices may be similar for different companies at different time intervals. If the hypothesis is correct, using information about another company can improve the prediction of the prices of the company in question. The key goal of the work is to test this hypothesis.

The goal will be achieved using a cluster approach that will effectively group similar financial time series data and use them to create synthetic indexes that are expected to reflect the unique characteristics of each group.

To evaluate the effectiveness of the proposed method, experiments will be conducted with real data sets of financial time series, comparing the accuracy of forecasting using the obtained index and without it. The prediction results will be evaluated using metrics such as average absolute percentage error (MAPE), average absolute error (MAE) and symmetric mean absolute percentage error (SMAPE).

Clustering methods have shown promising results in reflecting the basic structure of financial time series, but their effectiveness in improving the accuracy of forecasting has not been thoroughly investigated. In general, this work demonstrates the potential of using clustering methods to improve the accuracy of forecasting financial time series and analyzing the dynamics of market capitalization of public companies.

3.4 Proposed approach

In this work, three global stages can be identified.

1. 3 types of clustering will be investigated: Hierarchical, K-Means, Feature based. An indicator of the similarity of time series is the Euclidean distance and DTW.
2. Construction of a basic model for predicting stock prices. At this stage, it is important to select hyperparameters for the TFT model from the darts package.
3. Use of the obtained clusters to improve the prediction of stock prices. Training a TFT model on a list of time series compiled from companies in the same cluster.

4 Experiments

4.1 Datasets

The dataset used in this work was collected by Yahoo Finance over a 7-year period. The data set consists of time series data of closing prices of companies included in the NASDAQ-100 index. In addition to time series, the dataset also includes the attributes listed in Table 1.

Attribute	Description	Type
Ticker	Unique identifier used to identify a company on the exchange	str
Short name	Short name of the company	str
Country	The country in which the company is registered	str
Market	The stock exchange on which the company is listed	str
Sector	The sector to which the company belongs	str
Industry	Provides a more specific classification within the sector	str
Time series	Stock closing price for every day for the last 7 years (2014-2023)	float64

Table 1: Time series dataset properties description

The dataset provides a comprehensive view of the stock market by combining time series data with characteristics of specific companies.

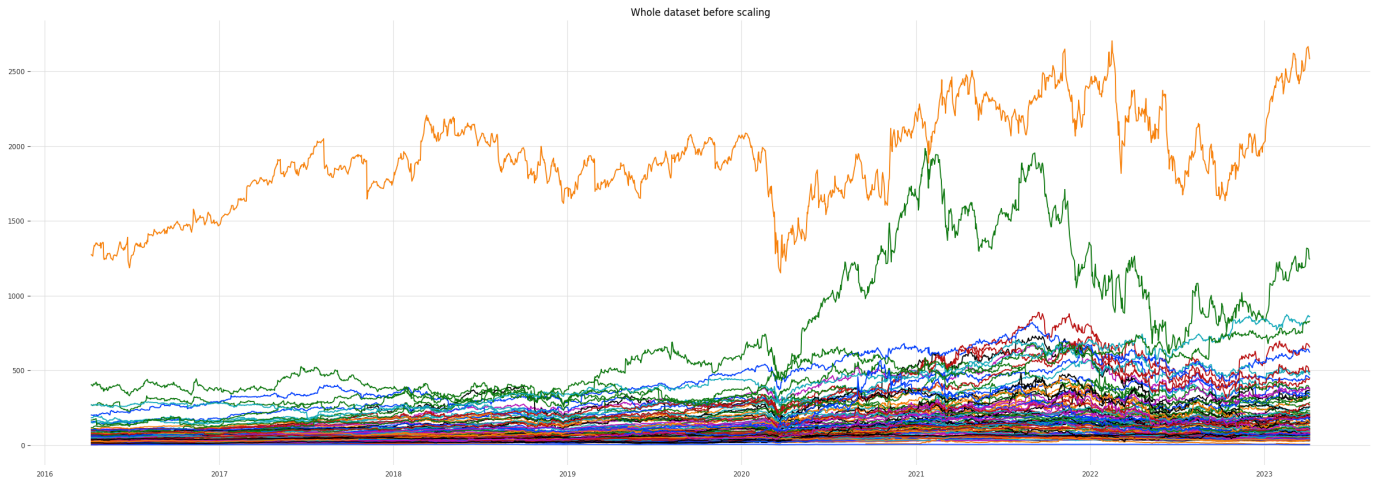

Figure 1: Dataset

Before the direct analysis, all data was preprocessed to remove any missing values, outliers, and noise. The data was also normalized (each row was divided by its maximum value) to ensure that each time series had the same scale.

4.2 Metrics

4.2.1 Time series similarity

To cluster time series data, it is necessary to use a similarity measure that can effectively capture the temporal relationships between data points. In this study, DTW and euclidean distance were used as similarity indicator to compare different time series.

Euclidean distance. The distance, which is defined as the square root of the sum of the squares of the differences between the corresponding elements of the series. The formal record of this distance looks like this:

$$Euclidean = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2},$$

where x and y are two time series, and n is the number of elements in the series.

Euclidean distance has several advantages, such as simplicity and intuitiveness, as well as a good ability to compare time series of different lengths. However, it also has some disadvantages, such as sensitivity to outliers and bias, which can reduce the quality of clustering.

DTW. Dynamic Time Warping (DTW) is a time series alignment algorithm that allows measure the similarity between two sequences, considering the possible offset (time shift) between them.

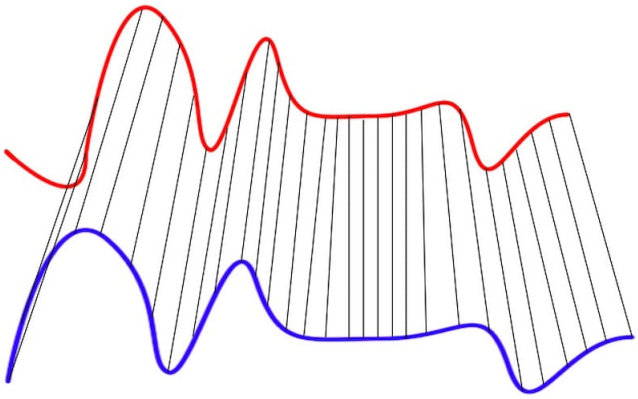

Figure 2: Dynamic time warping between two time series

The essence of the method is that two sequences are compared in pairs, and then aligned relative to each other. In the process of alignment, each point of one sequence correlates with a point of another sequence that is located at the closest distance in time, even if it differs significantly from their original position.

The DTW method works as follows:

1. First, the preliminary distance between two time series is calculated.
2. Then a distance matrix is constructed, where each element corresponds to the distance between a pair of points in the time series.
3. The optimal path with the minimum sufficient distance is calculated, points in two time series are compared.
4. The final distance is calculated as the sum of the distances between the points on the optimal path.

4.2.2 Quality of time series prediction

To assess the quality of time series prediction, the metrics MAE, MAE, SMAPE were used.

MAPE. Mean Absolute Percentage Error (MAPE) is a metric that is used to evaluate the accuracy of time series forecasting models. It is calculated as the average absolute value of the percentage error at all forecast points. The mathematical formula of MAP looks like this:

$$MAPE = \frac{1}{n} * \sum_{i=1}^{n} |\frac{actual_i - predicted_i}{actual_i}| * 100\%,$$

where actual is the actual value in the time series, predicted is the value predicted by the model and n is the number of points in the time series.

The main advantage of using MAPE are the fact that MAPE allows you to estimate the accuracy of the prediction model as a percentage, which is clear and easy to interpret.

MAE. Mean Absolute Error (MAE) measures the average absolute difference between the predicted values and the actual values. MAE is calculated using the formula:

$$MAE = \frac{\sum_{i=1}^{n} |actual_i - predicted_i|}{n},$$

where actual is the actual value in the time series, predicted is the value predicted by the model and n is the number of points in the time series.

The advantages of the MAE metric are that it is easily interpreted and is not sensitive to outliers in the data, since the absolute value is used in the calculation. Also, it is additive and scalable (i.e. it can be averaged over different ranges, for example, for each sample or for each class in a multiclass classification).

Compared to other techniques, such as MSE (root-mean-square error), MAE is better suited for estimating small for predicted values, since it does not increase the error in the case of large predictions.

SMAPE. Symmetric Mean Absolute Percentage Error (SMAPE) is the percentage deviation of the forecast from the actual value. Formula for calculating SMAPE:

$$SMAPE = \frac{1}{n} * \sum_{i=1}^{n} \frac{2 * |predicted_i - actual_i|}{|actual_i| + |predicted_i|} * 100\%,$$

where actual is the actual value in the time series, predicted is the value predicted by the model and n is the number of points in the time series.

SMAPE is best used to assess the quality of predictions when it is important to achieve a balance between small and large values, as well as to reduce the impact of outliers. SMAPE is also resistant to outliers because it uses the number of absolute changes in errors rather than their area. Moreover, it is a more interpretable metric than RMSE (Root Mean Squared Error), which does not provide a visual interpretation such as SMAPE.

4.3 Results

4.3.1 Time series clustering

For this study, we used data on the stock prices of companies that make up the NASDAQ-100 index during the period from April 6, 2020 to April 5, 2023. Only companies whose stocks are traded on the stock exchange for the entire time period were taken. A time series was formed for each company, representing the daily value of the stock price at the close of trading.

Hierarchical clustering. Hierarchical clustering was performed based on DTW distances between pairs of time series. The number of clusters was not determined in advance, but was calculated based on the minimum threshold of the distance between rows to get into one cluster.

The best clustering result was at a DTW distance of 2.5, 4 clusters were obtained. Figure 3 illustrates the result of the distribution of companies into clusters.

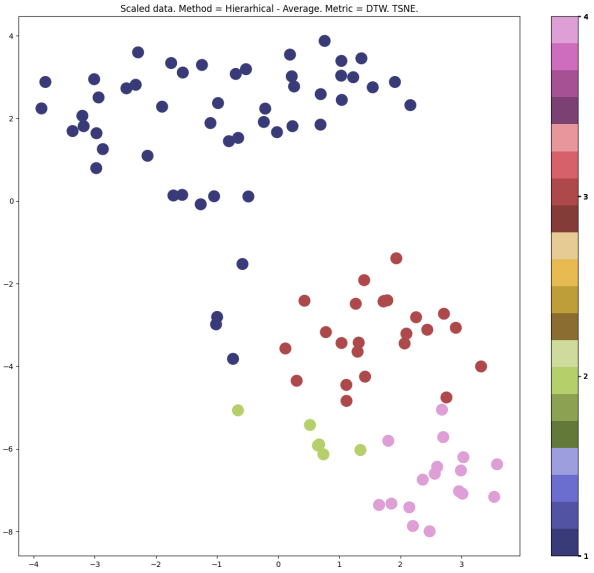

Figure 3: Visualization of clusters as a result of hierarchical clustering using tsne

As a result of this experiment, it became clear that the optimal number of clusters is 4 - 5, it will be used in further experiments.

K-Means. Clustering based on K-Means was performed based on Euclidean distance and DTW. Figures 4 and 5 show the results.

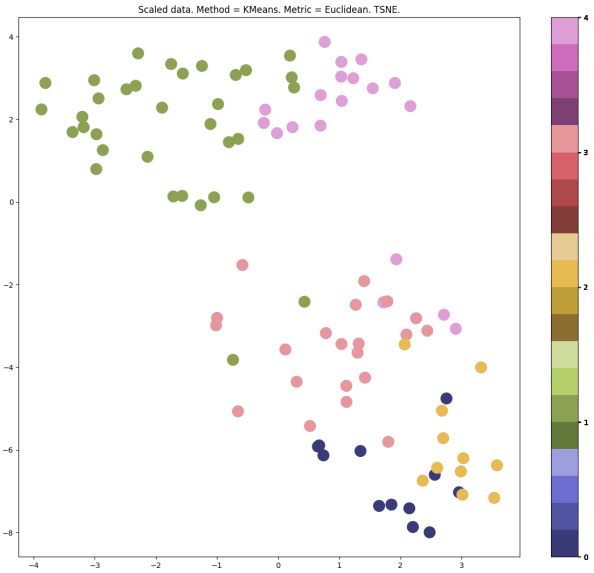

Figure 4: Visualization of clusters as a result of K-Means + Euclidean using tsne

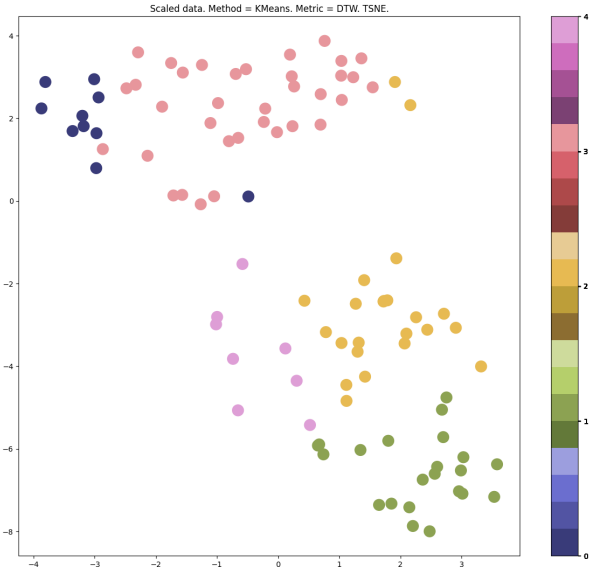

Figure 5: Visualization of clusters as a result of K-Means + DTW using tsne

It is noticeable that the Euclidean distance copes worse, because it does not take into account the shift in time series. With DTW, the result is better, but due to the sensitivity of the algorithm to outliers, some companies fall into different clusters, although there is a small distance between them.

In general, both approaches (hierarchical clustering and K-Means) give good results, but we will assume that the results of the first experiment are better, because it did not need to know the number of clusters. If we need to apply the developed approach to another task, hierarchical clustering will be more adaptive.

Feature based approach. The approach has proven itself well in other tasks, but it is not able to identify nonlinear connections between companies, which is why it should not be used in this task. Figure 6 clearly shows the absence of a clear boundary between clusters.

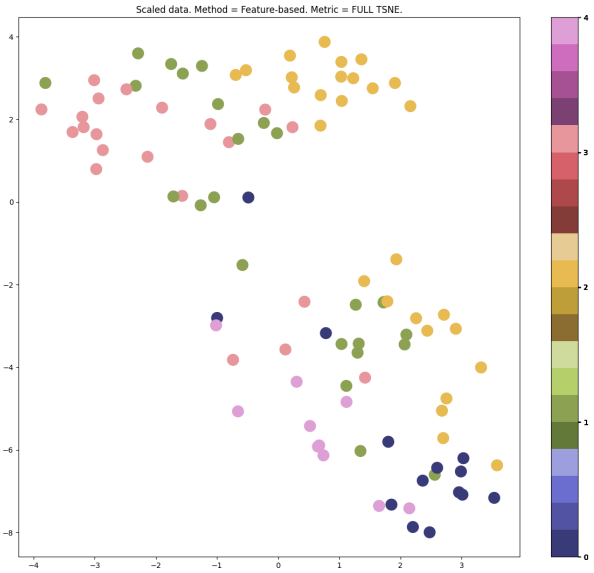

Figure 6: Visualization of clusters as a result of feature based clustering using tsne

Summing up the results of the stage. The results of experiments with clustering showed that companies from the NASDAQ-100 index can be divided into 4-5 clusters. Hierarchical clustering showed the most representative results. Its results will become the basis for the creation of synthetic indexes.

Tables 2-5 show the final distribution of companies by cluster, which will be used in further experiments.

Company name	Sector	Industry	Country
Apple Inc.	Technology	Consumer Electronics	United States
Analog Devices, Inc.	Technology	Semiconductors	United States
Automatic Data Processing, Inc.	Industrials	Staffing & Employment Services	United States
American Electric Power Company	Utilities	Utilities—Regulated Electric	United States
Amgen Inc.	Healthcare	Drug Manufacturers—General	United States
ANSYS, Inc.	Technology	Software—Application	United States
Activision Blizzard, Inc.	Communication Services	Electronic Gaming & Multimedia	United States
Broadcom Inc.	Technology	Semiconductors	United States
Astrazeneca PLC	Healthcare	Drug Manufacturers—General	United Kingdom
Biogen Inc.	Healthcare	Drug Manufacturers—General	United States
Booking Holdings Inc. Common St	Consumer Cyclical	Travel Services	United States
Cadence Design Systems, Inc.	Technology	Software—Application	United States
Costco Wholesale Corporation	Consumer Defensive	Discount Stores	United States
Copart, Inc.	Consumer Cyclical	Auto & Truck Dealerships	United States
Cisco Systems, Inc.	Technology	Communication Equipment	United States
CSX Corporation	Industrials	Railroads	United States
Cintas Corporation	Industrials	Specialty Business Services	United States
Dollar Tree, Inc.	Consumer Defensive	Discount Stores	United States
Electronic Arts Inc.	Communication Services	Electronic Gaming & Multimedia	United States
Exelon Corporation	Utilities	Utilities—Regulated Electric	United States
Fastenal Company	Industrials	Industrial Distribution	United States
Fiserv, Inc.	Technology	Information Technology Services	United States
Fortinet, Inc.	Technology	Software—Infrastructure	United States
Gilead Sciences, Inc.	Healthcare	Drug Manufacturers—General	United States
Honeywell International Inc.	Industrials	Conglomerates	United States
Keurig Dr Pepper Inc.	Consumer Defensive	Beverages—Non-Alcoholic	United States
The Kraft Heinz Company	Consumer Defensive	Packaged Foods	United States
KLA Corporation	Technology	Semiconductor Equipment & Materials	United States
Marriott International	Consumer Cyclical	Lodging	United States
Microchip Technology Incorporat	Technology	Semiconductors	United States
Mondelez International, Inc.	Consumer Defensive	Confectioners	United States
Monster Beverage Corporation	Consumer Defensive	Beverages—Non-Alcoholic	United States
Old Dominion Freight Line, Inc.	Industrials	Trucking	United States
O'Reilly Automotive, Inc.	Consumer Cyclical	Specialty Retail	United States
Palo Alto Networks, Inc.	Technology	Software—Infrastructure	United States
Paychex, Inc.	Industrials	Staffing & Employment Services	United States
PACCAR Inc.	Industrials	Farm & Heavy Construction Machinery	United States
Pepsico, Inc.	Consumer Defensive	Beverages—Non-Alcoholic	United States
Regeneron Pharmaceuticals, Inc.	Healthcare	Biotechnology	United States
Ross Stores, Inc.	Consumer Cyclical	Apparel Retail	United States
Starbucks Corporation	Consumer Cyclical	Restaurants	United States
Seagen Inc.	Healthcare	Biotechnology	United States
Sirius XM Holdings Inc.	Communication Services	Entertainment	United States
Synopsys, Inc.	Technology	Software—Infrastructure	United States
T-Mobile US, Inc.	Communication Services	Telecom Services	United States
Texas Instruments Incorporated	Technology	Semiconductors	United States
Verisk Analytics, Inc.	Industrials	Consulting Services	United States
VeriSign, Inc.	Technology	Software—Infrastructure	United States
Vertex Pharmaceuticals Incorporated	Healthcare	Biotechnology	United States
Walgreens Boots Alliance, Inc.	Healthcare	Pharmaceutical Retailers	United States
Xcel Energy Inc.	Utilities	Utilities—Regulated Electric	United States

Table 2: Cluster 1

Company name	Sector	Industry	Country
Charter Communications, Inc.	Communication Services	Telecom Services	United States
Illumina, Inc.	Healthcare	Diagnostics & Research	United States
Intel Corporation	Technology	Semiconductors	United States
Meta Platforms, Inc.	Communication Services	Internet Content & Information	United States
Netflix, Inc.	Communication Services	Entertainment	United States
Splunk Inc.	Technology	Software—Infrastructure	United States

Table 3: Cluster 2

Company name	Sector	Industry	Country
Adobe Inc.	Technology	Software—Infrastructure	United States
Autodesk, Inc.	Technology	Software—Application	United States
Applied Materials, Inc.	Technology	Semiconductor Equipment & Materials	United States
Amazon.com, Inc.	Consumer Cyclical	Internet Retail	United States
ASML Holding N.V. - New York Re	Technology	Semiconductor Equipment & Materials	Netherlands
Comcast Corporation	Communication Services	Telecom Services	United States
Cognizant Technology Solutions	Technology	Information Technology Services	United States
DexCom, Inc.	Healthcare	Medical Devices	United States
eBay Inc.	Consumer Cyclical	Internet Retail	United States
Alphabet Inc.	Communication Services	Internet Content & Information	United States
Alphabet Inc.	Communication Services	Internet Content & Information	United States
IDEXX Laboratories, Inc.	Healthcare	Diagnostics & Research	United States
Intuit Inc.	Technology	Software—Application	United States
Intuitive Surgical, Inc.	Healthcare	Medical Instruments & Supplies	United States
Lam Research Corporation	Technology	Semiconductor Equipment & Materials	United States
lululemon athletica inc.	Consumer Cyclical	Apparel Retail	Canada
MercadoLibre, Inc.	Consumer Cyclical	Internet Retail	Uruguay
Micron Technology, Inc.	Technology	Semiconductors	United States
NetEase, Inc.	Communication Services	Electronic Gaming & Multimedia	China
NVIDIA Corporation	Technology	Semiconductors	United States
NXP Semiconductors N.V.	Technology	Semiconductors	Netherlands
QUALCOMM Incorporated	Technology	Semiconductors	United States
Skyworks Solutions, Inc.	Technology	Semiconductors	United States
Workday, Inc.	Technology	Software—Application	United States

Table 4: Cluster 3

Company name	Sector	Industry	Country
Align Technology, Inc.	Healthcare	Medical Devices	United States
Advanced Micro Devices, Inc.	Technology	Semiconductors	United States
Baidu, Inc.	Communication Services	Internet Content & Information	China
CrowdStrike Holdings, Inc.	Technology	Software—Infrastructure	United States
Datadog, Inc.	Technology	Software—Application	United States
DocuSign, Inc.	Technology	Software—Application	United States
JD.com, Inc.	Consumer Cyclical	Internet Retail	China
Moderna, Inc.	Healthcare	Biotechnology	United States
Marvell Technology, Inc.	Technology	Semiconductors	United States
Match Group, Inc.	Communication Services	Internet Content & Information	United States
Okta, Inc.	Technology	Software—Infrastructure	United States
Pinduoduo Inc.	Consumer Cyclical	Internet Retail	China
PayPal Holdings, Inc.	Financial Services	Credit Services	United States
Atlassian Corporation	Technology	Software—Application	Australia
Tesla, Inc.	Consumer Cyclical	Auto Manufacturers	United States
Zoom Video Communications, Inc.	Technology	Software—Application	United States
Zscaler, Inc.	Technology	Software—Infrastructure	United States

Table 5: Cluster 4

4.3.2 Synthetic index

Each cluster is a separate synthetic index. On graphs 7 - 10 the blue line indicates the cluster index line. The index is calculated as the average value between time series that fall into the same cluster.

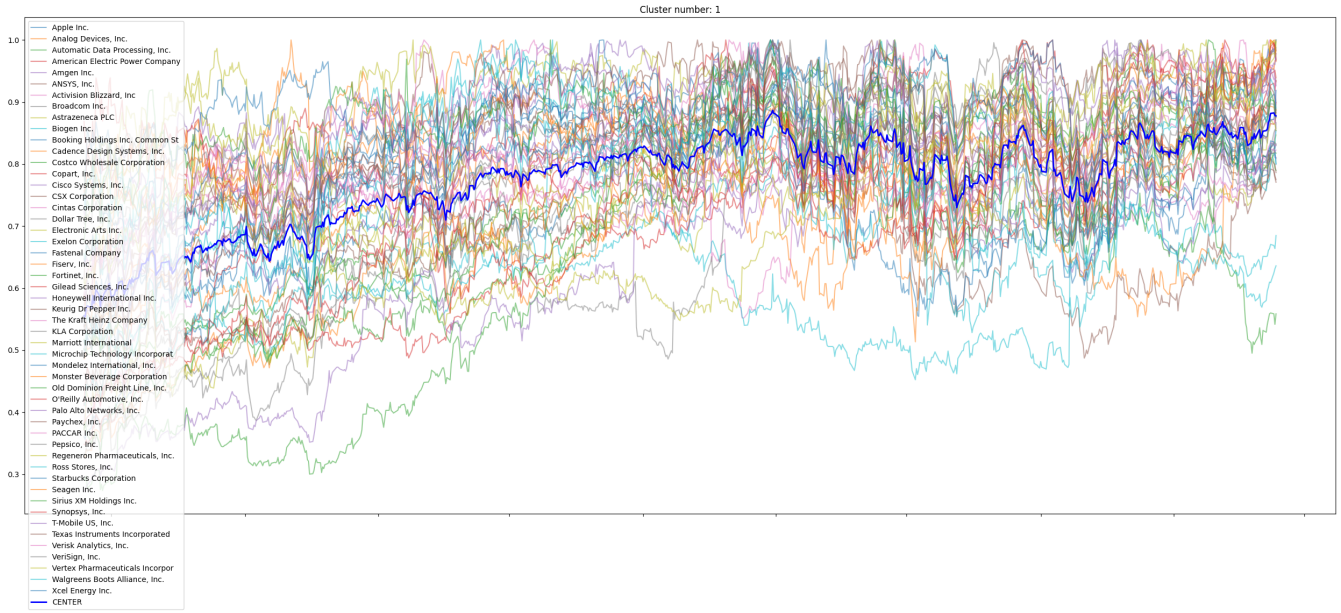

Figure 7: Cluster 1

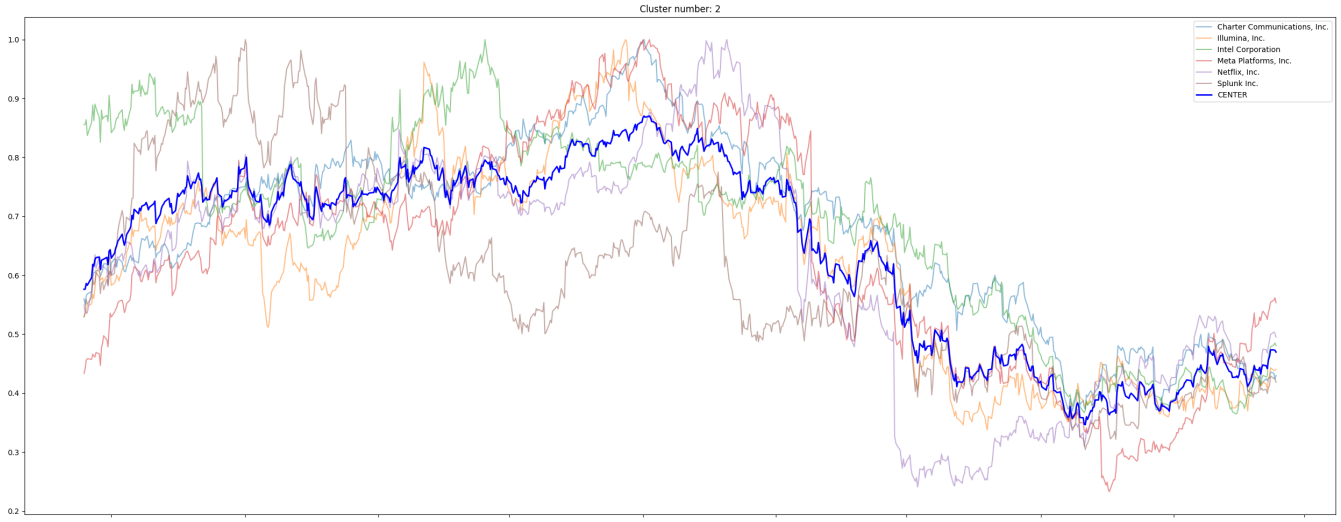

Figure 8: Cluster 2

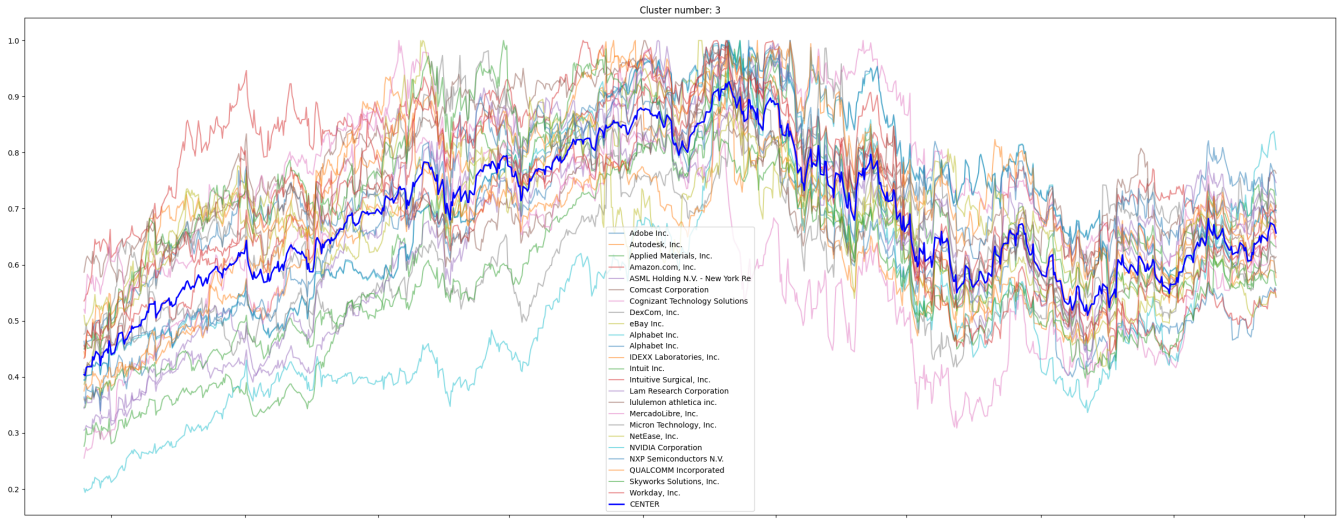

Figure 9: Cluster 3

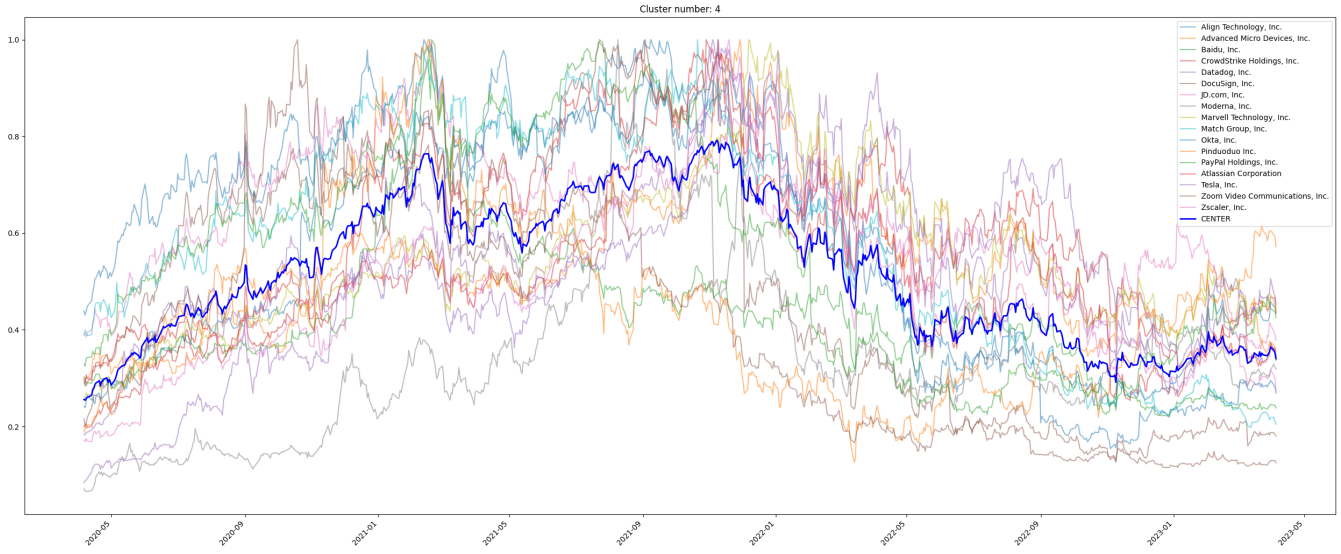

Figure 10: Cluster 4

4.3.3 Stock price prediction

The first task of this stage was to train the basic TFT model on one time series. A number of launches were carried out, which showed that the method with training on 7 days and prediction of 1 is the most effective. For the final prediction, the historical_forecasts method is used, which, unlike the standard predict, takes into account the validation sample, in this case the prediction horizon is 3 days.

Further research has shown that the index can be used to improve the forecasts of quotations of companies. The most effective results were shown by the method with training not just on a synthetic index, but with training on a list of time series of companies from one index.

To make sure the hypothesis was correct, the experiments were repeated for 12 random companies (for 3 companies from each index). The final results showed that, on average, the approach developed in this scientific article shows more accurate results. For some companies, a slight improvement is also demonstrated by models trained on a random index, but it is worth noting that for most of these models, the forecast either improved slightly or, conversely, worsened.

The following table shows a comparison of indicators for a model trained on only one time series (BASE column), for a model trained on time series of companies from one cluster (CLUSTER column), and for a model trained on data from companies from random clusters (RANDOM column).

Company name	**BASE**			**CLUSTER**			**RANDOM**		
	MAPE	MAE	SMAPE	MAPE	MAE	SMAPE	MAPE	MAE	SMAPE
Cluster 1									
Fortinet, Inc.	4.64	0.04	4.8	**2.67**	**0.02**	**2.72**	3.06	0.02	3.15
Automatic Data Processing, Inc.	2.73	0.02	2.78	**1.95**	0.02	**1.96**	2.58	0.02	2.59
Apple Inc.	3.69	0.03	3.8	**1.94**	*0.02*	**1.98**	3.69	0.03	3.75
Cluster 2									
Charter Communications, Inc.	3.31	0.01	3.34	**2.79**	0.01	**2.82**	2.98	0.01	3.01
Intel Corporation	5.35	0.02	5.53	**3.33**	**0.01**	**3.35**	3.95	0.02	4.09
Netflix, Inc.	4.65	0.02	4.82	**3.46**	0.02	3.51	4.08	0.02	4.19
Cluster 3									
Intuit Inc.	3.74	0.02	3.77	3.7	0.02	3.74	*4.13*	0.02	*4.15*
Adobe Inc.	3.72	0.02	3.8	**3.43**	0.02	**3.48**	3.62	0.02	3.66
Skyworks Solutions, Inc.	3.87	0.02	3.95	**3.14**	0.02	**3.19**	*4.37*	0.02	*4.53*
Cluster 4									
Marvell Technology, Inc.	5.41	0.02	5.5	**4.83**	0.02	**4.91**	*5.43*	*0.03*	*5.6*
Baidu, Inc.	4.76	0.02	4.66	4.4	0.02	4.38	4.52	0.02	4.51
Match Group, Inc.	6.18	0.01	5.99	4.57	0.01	4.52	4.77	0.01	4.74

Table 6: Prediction results

Graphs 11 - 46 illustrate actual and predicted stock prices of 12 companies listed in Table 6.

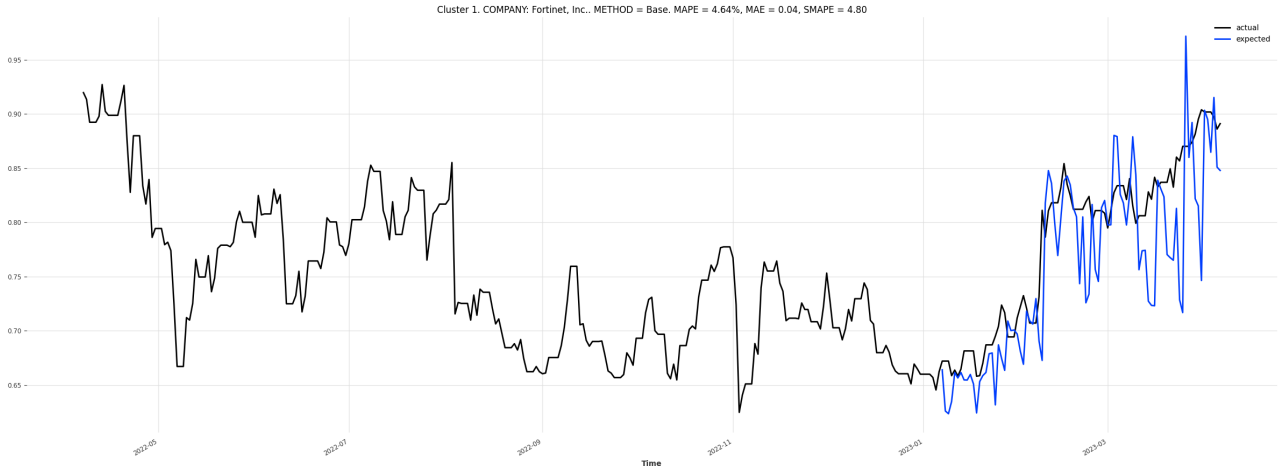

Figure 11: Cluster 1: Fortinet, Inc. [Model trained on only one time series]

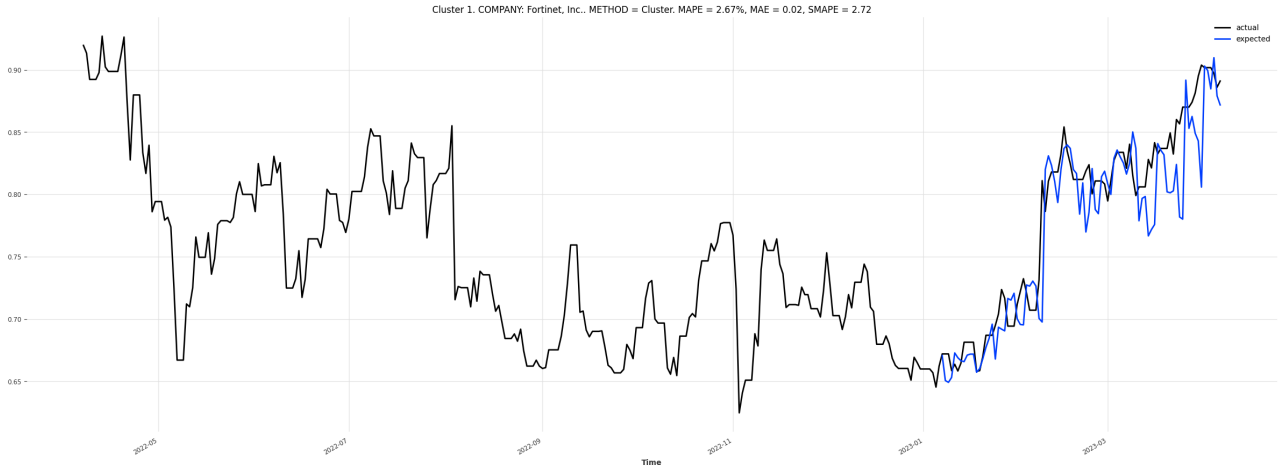

Figure 12: Cluster 1: Fortinet, Inc. [Model trained on data of companies from the same cluster]

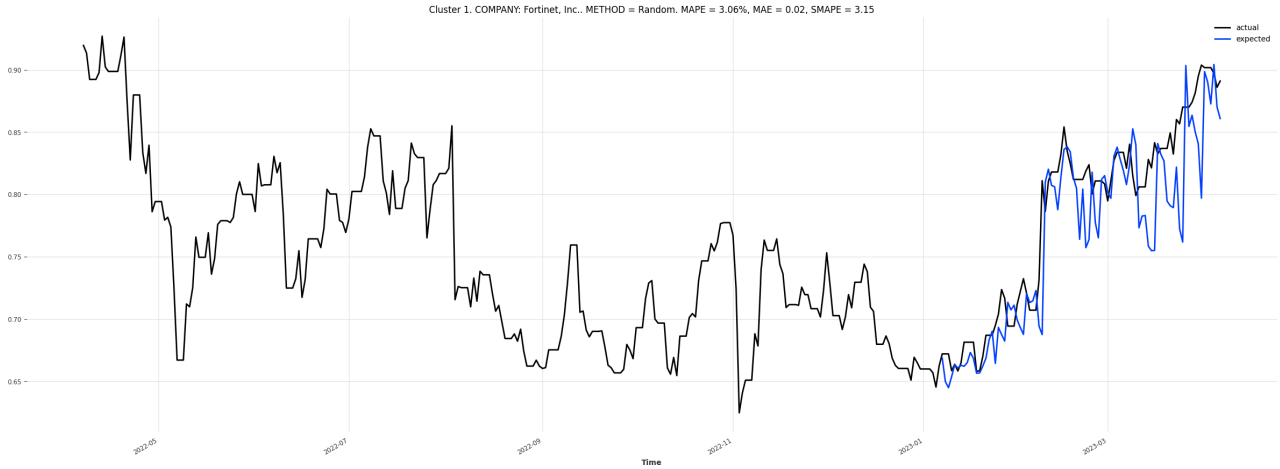

Figure 13: Cluster 1: Fortinet, Inc. [Model trained on data of companies from random clusters]

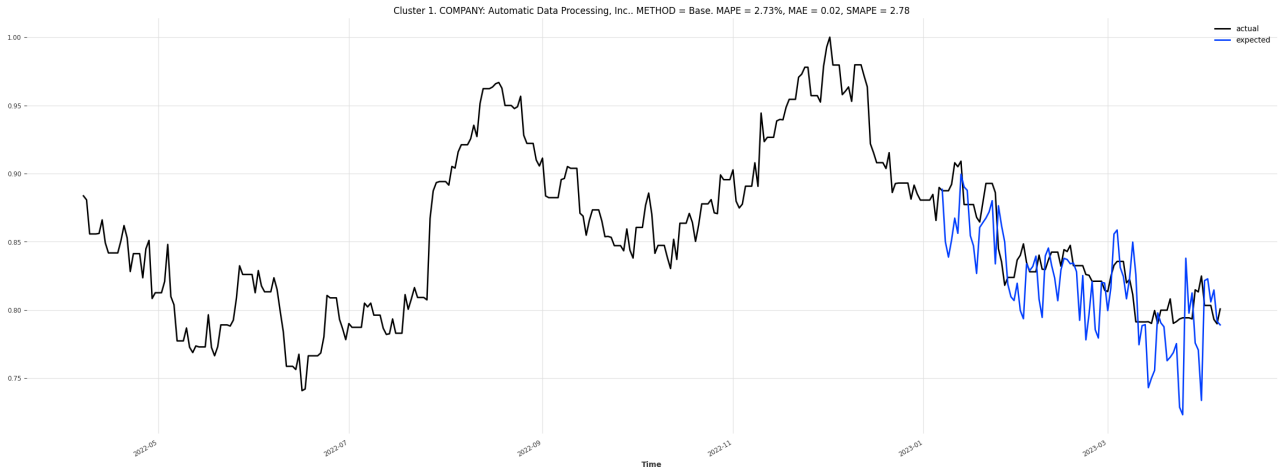

Figure 14: Cluster 1: Automatic Data Processing, Inc. [Model trained on only one time series]

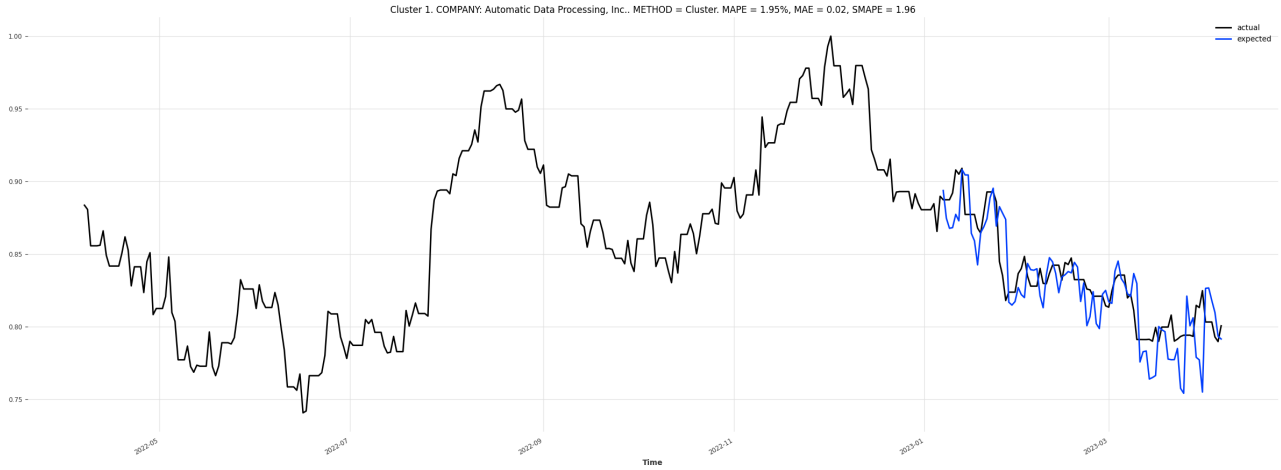

Figure 15: Cluster 1: Automatic Data Processing, Inc. [Model trained on data of companies from the same cluster]

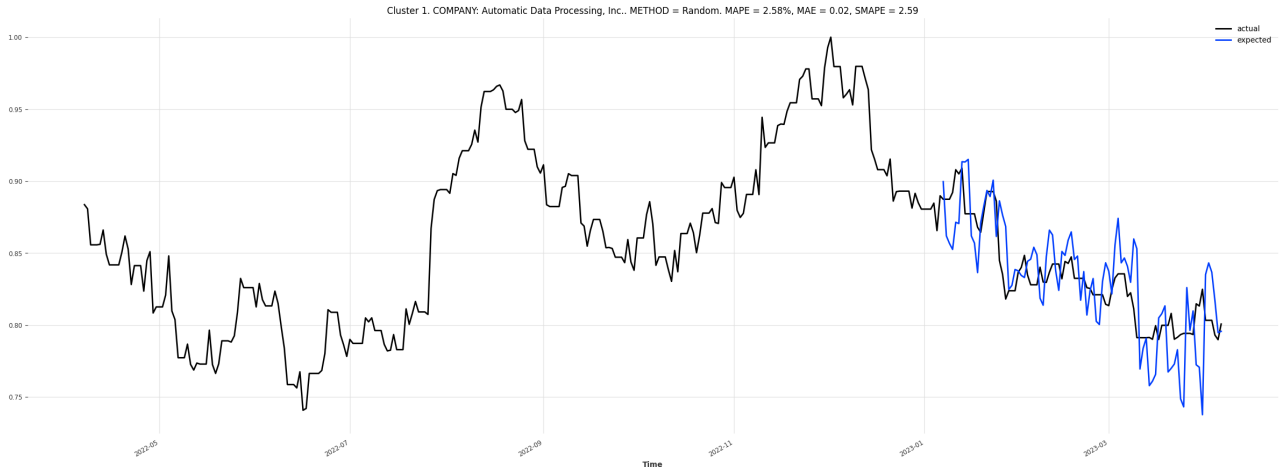

Figure 16: Cluster 1: Automatic Data Processing, Inc. [Model trained on data of companies from random clusters]

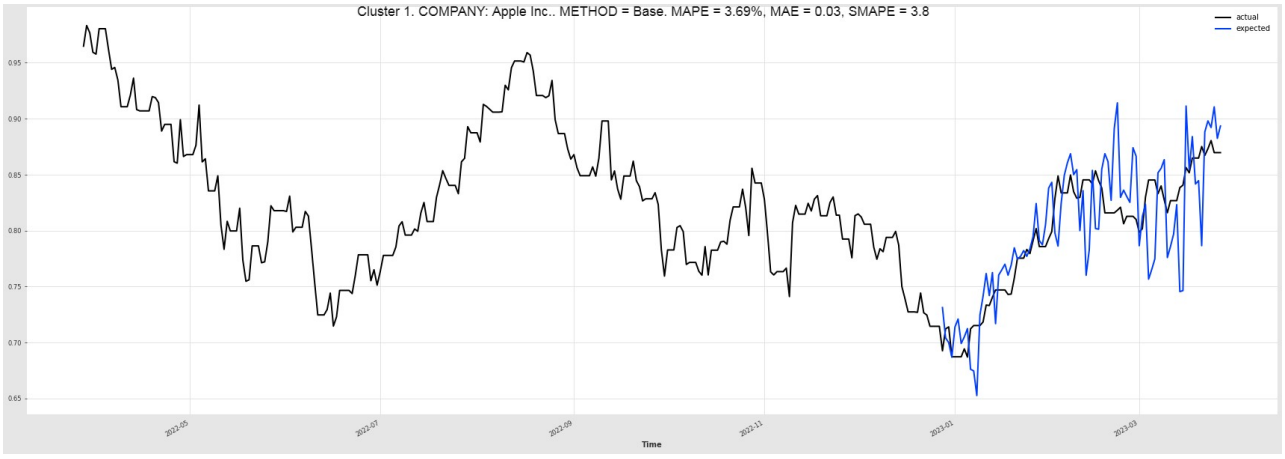

Figure 17: Cluster 1: Apple Inc. [Model trained on only one time series]

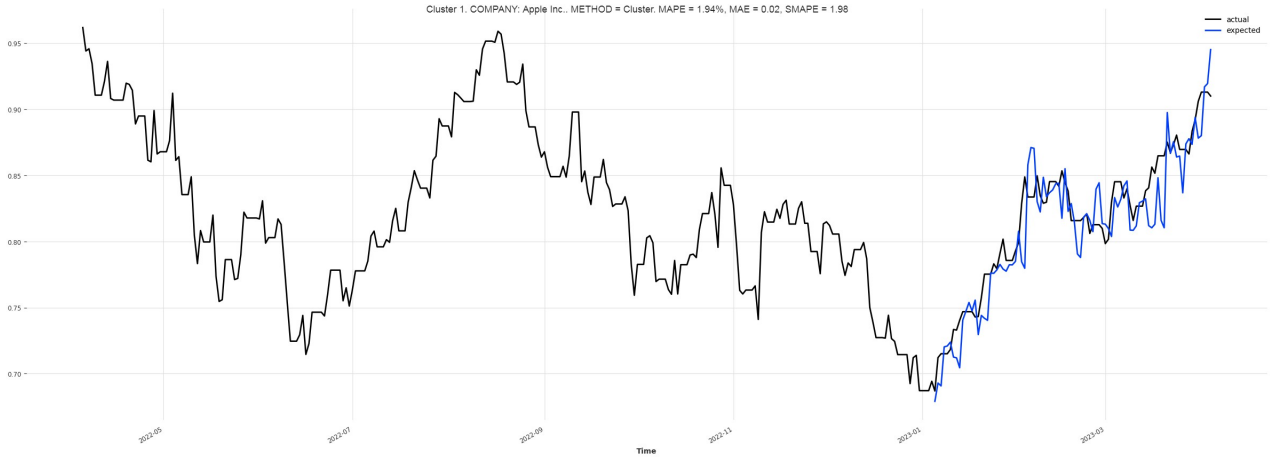

Figure 18: Cluster 1: Apple Inc. [Model trained on data of companies from the same cluster]

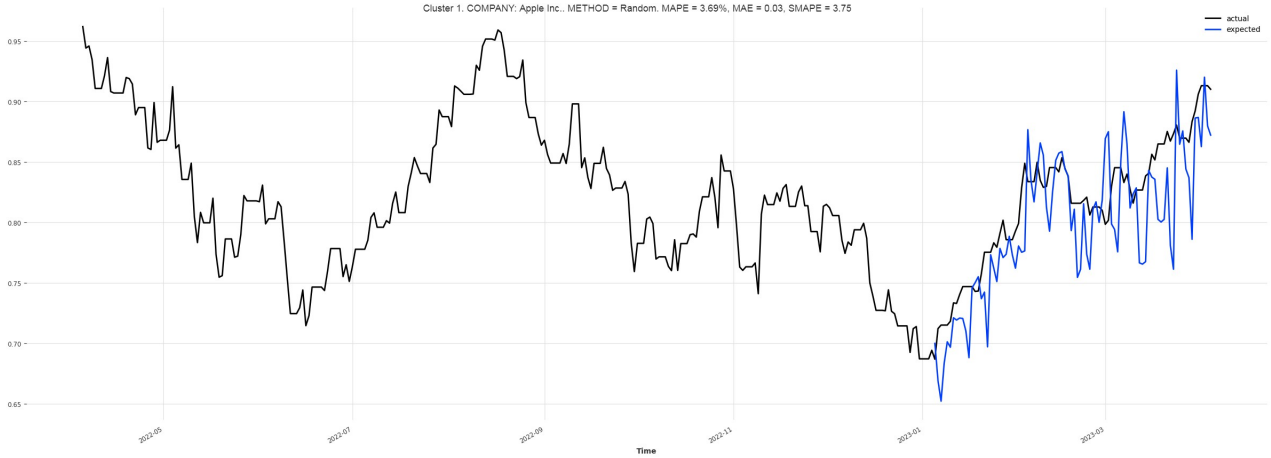

Figure 19: Cluster 1: Apple Inc. [Model trained on data of companies from random clusters]

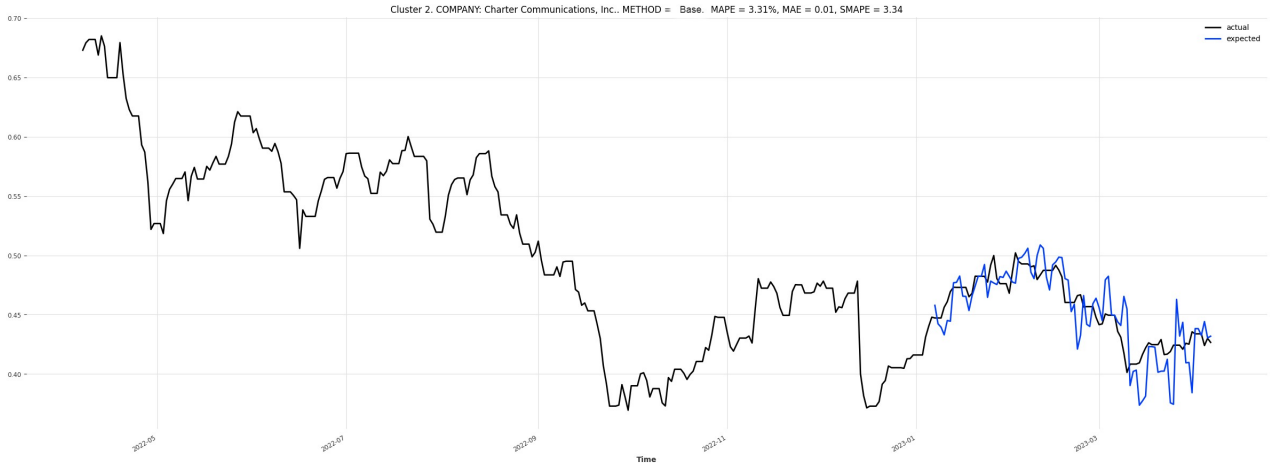

Figure 20: Cluster 2: Charter Communications, Inc. [Model trained on only one time series]

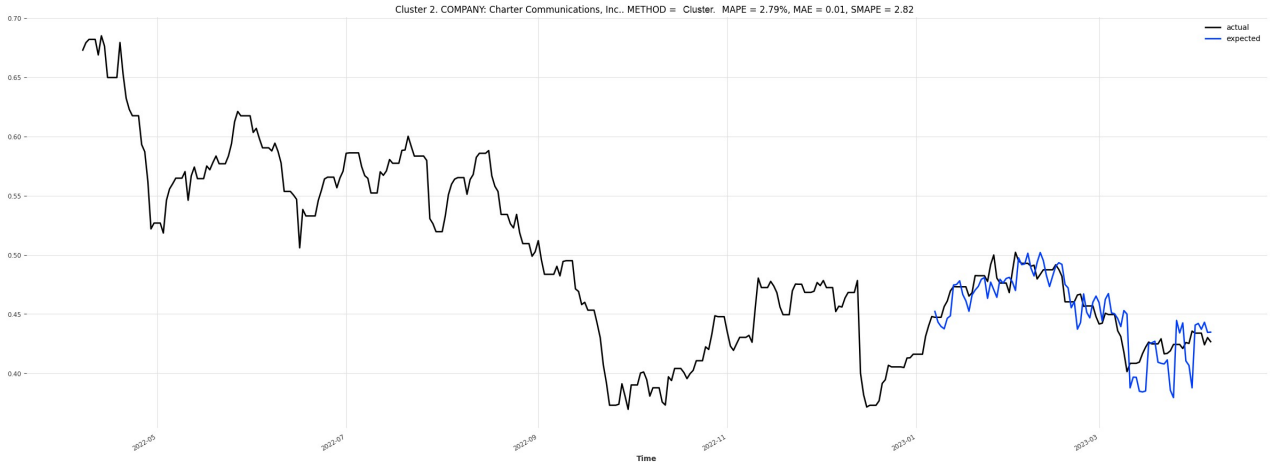

Figure 21: Cluster 2: Charter Communications, Inc. [Model trained on data of companies from the same cluster]

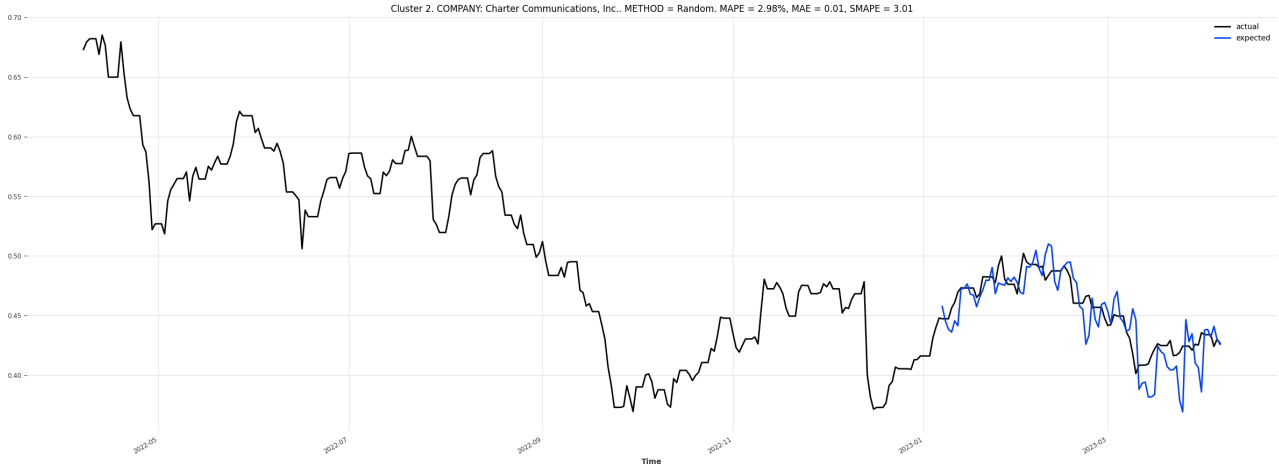

Figure 22: Cluster 2: Charter Communications, Inc. [Model trained on data of companies from random clusters]

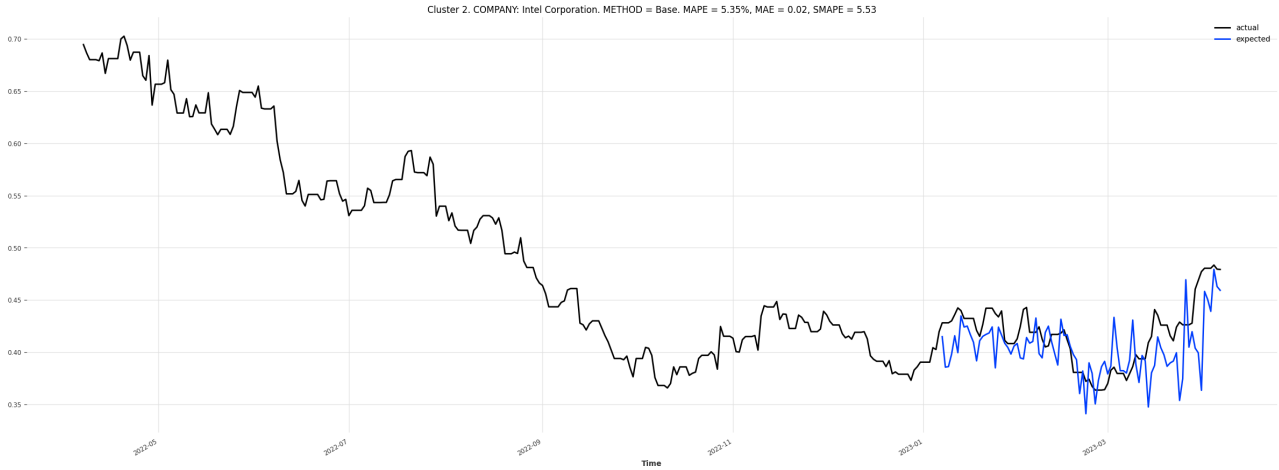

Figure 23: Cluster 2: Intel Corporation [Model trained on only one time series]

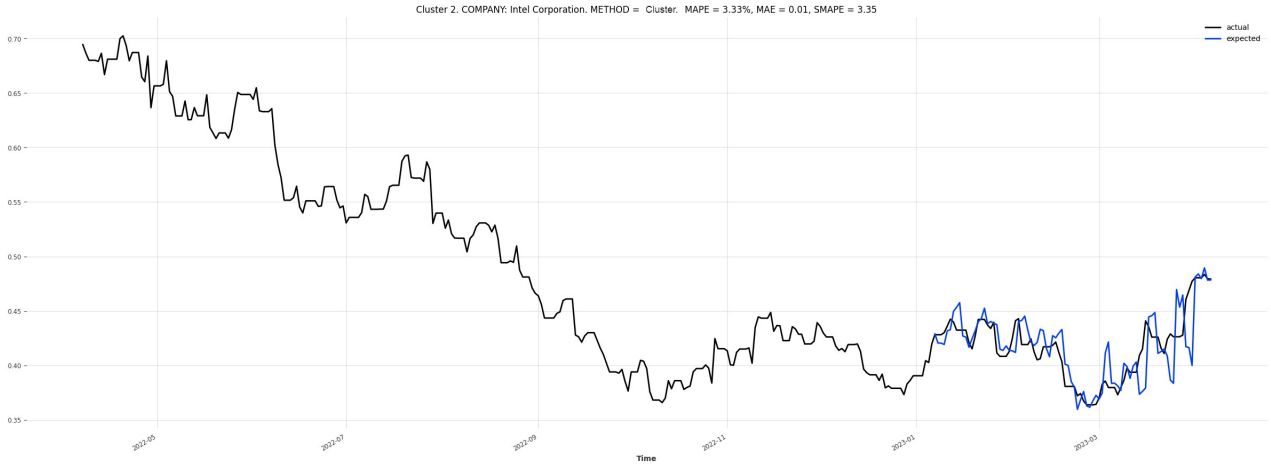

Figure 24: Cluster 2: Intel Corporation [Model trained on data of companies from the same cluster]

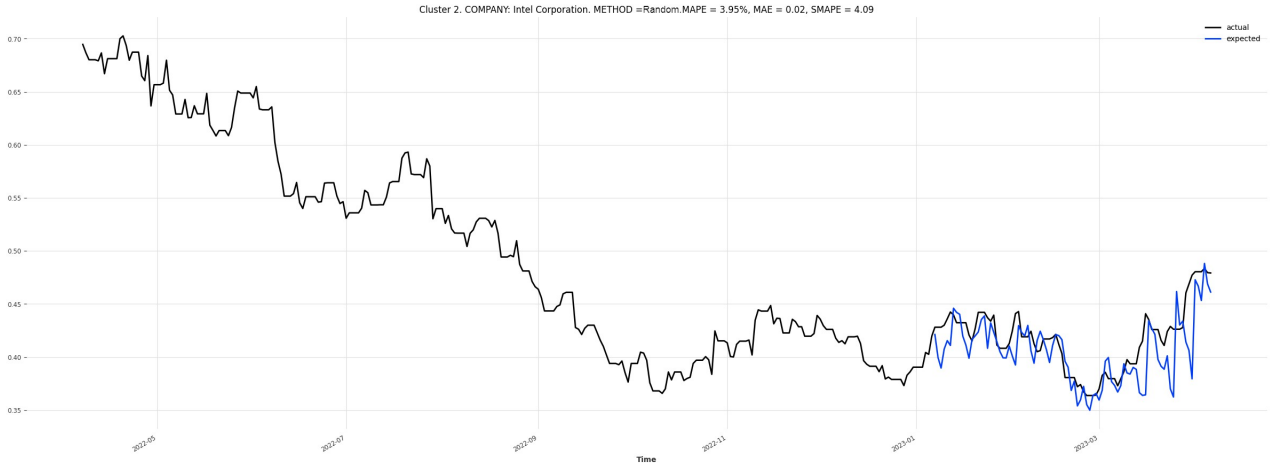

Figure 25: Cluster 2: Intel Corporation [Model trained on data of companies from random clusters]

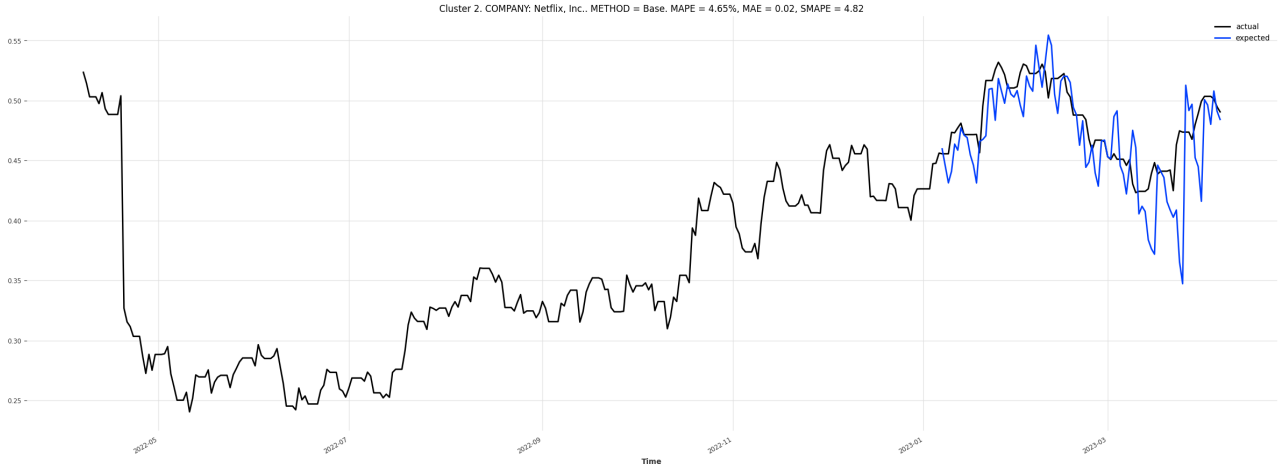

Figure 26: Cluster 2: Netflix, Inc. [Model trained on only one time series]

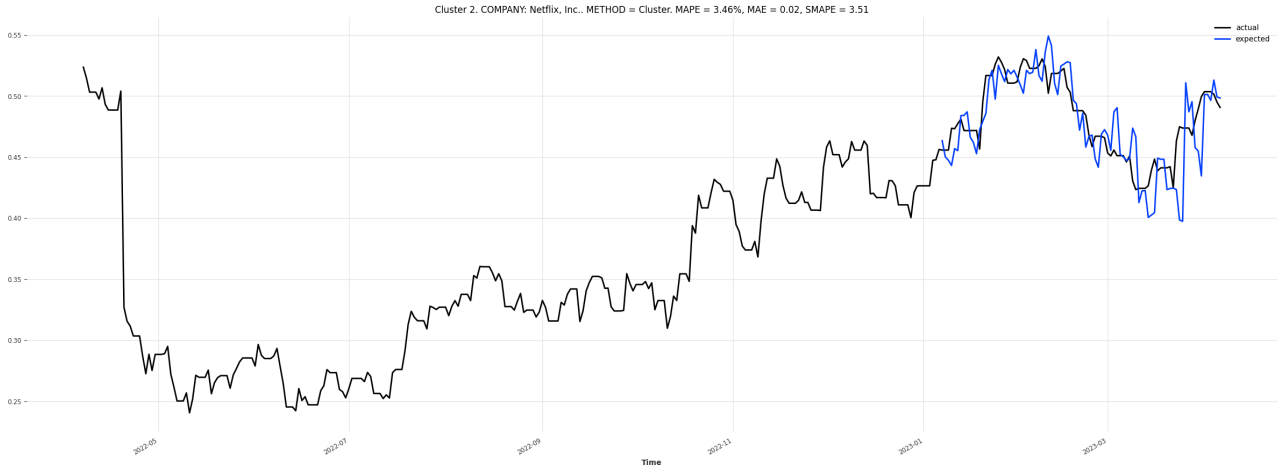

Figure 27: Cluster 2: Netflix, Inc. [Model trained on data of companies from the same cluster]

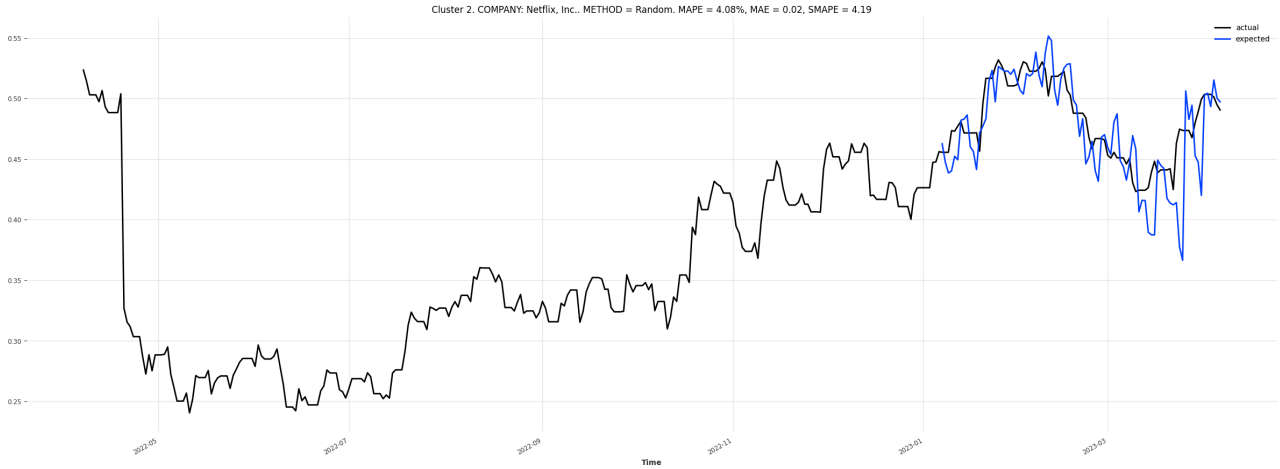

Figure 28: Cluster 2: Netflix, Inc. [Model trained on data of companies from random clusters]

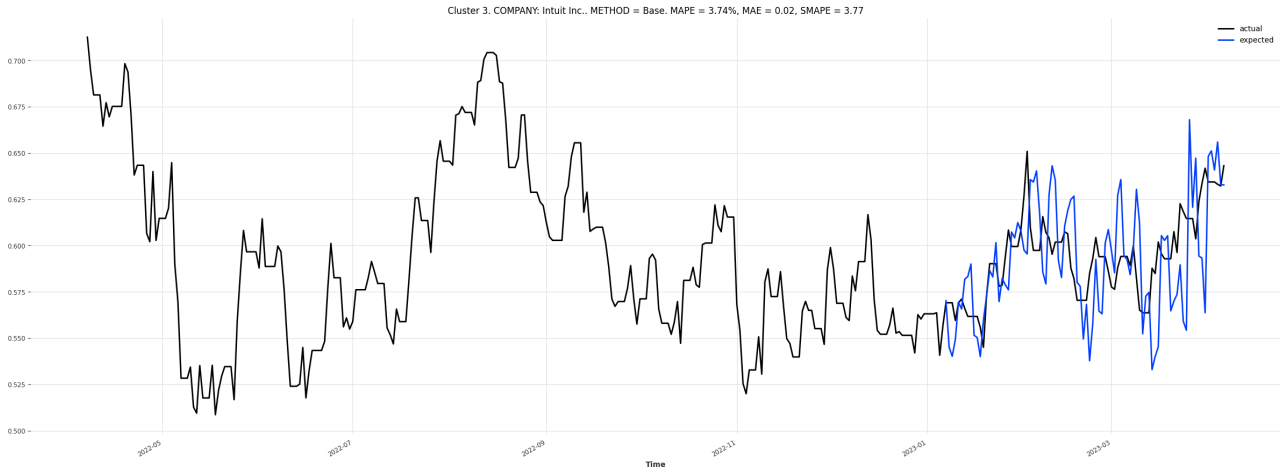

Figure 29: Cluster 3: Intuit Inc. [Model trained on only one time series]

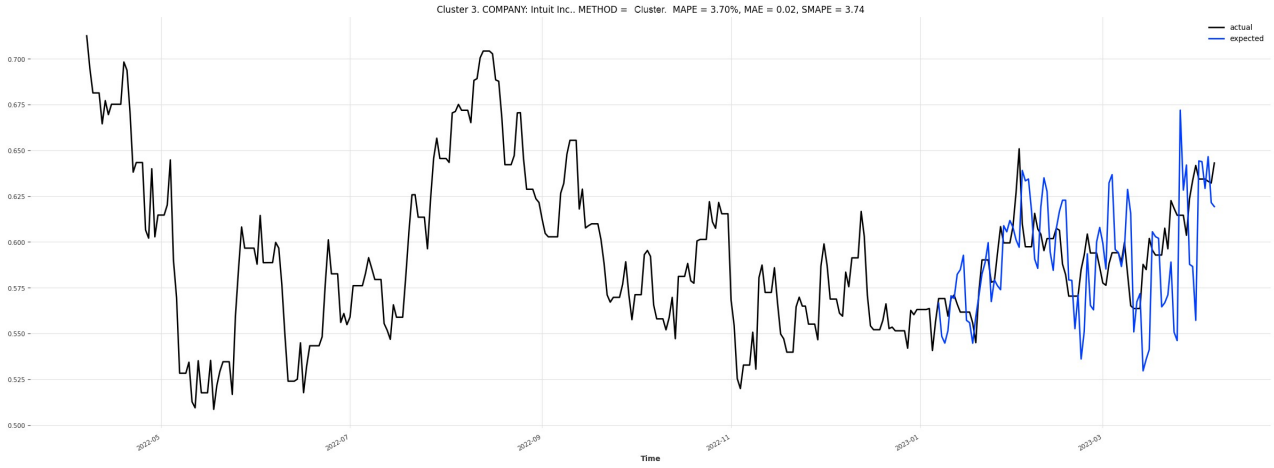

Figure 30: Cluster 3: Intuit Inc. [Model trained on data of companies from the same cluster]

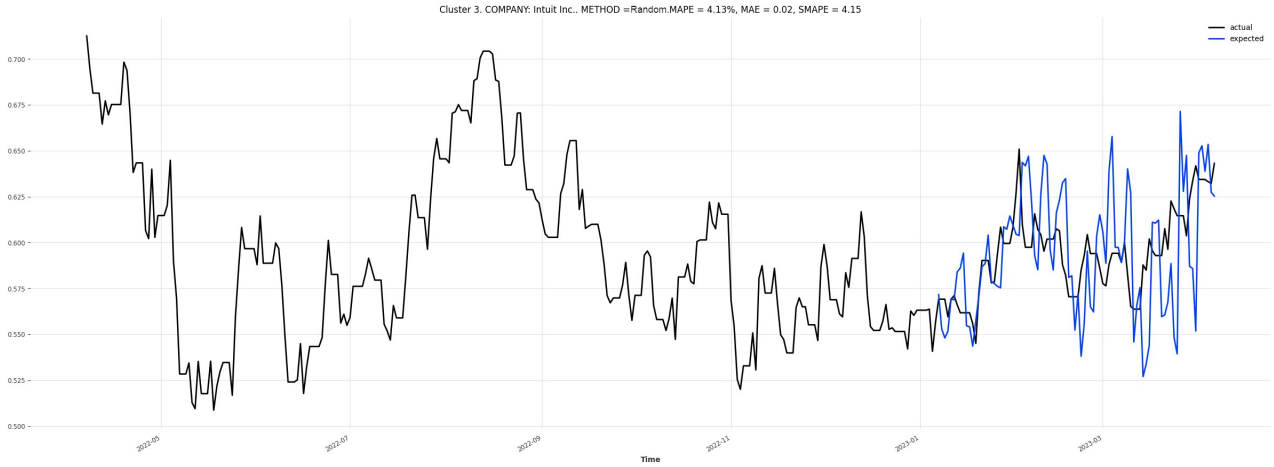

Figure 31: Cluster 3: Intuit Inc. [Model trained on data of companies from random clusters]

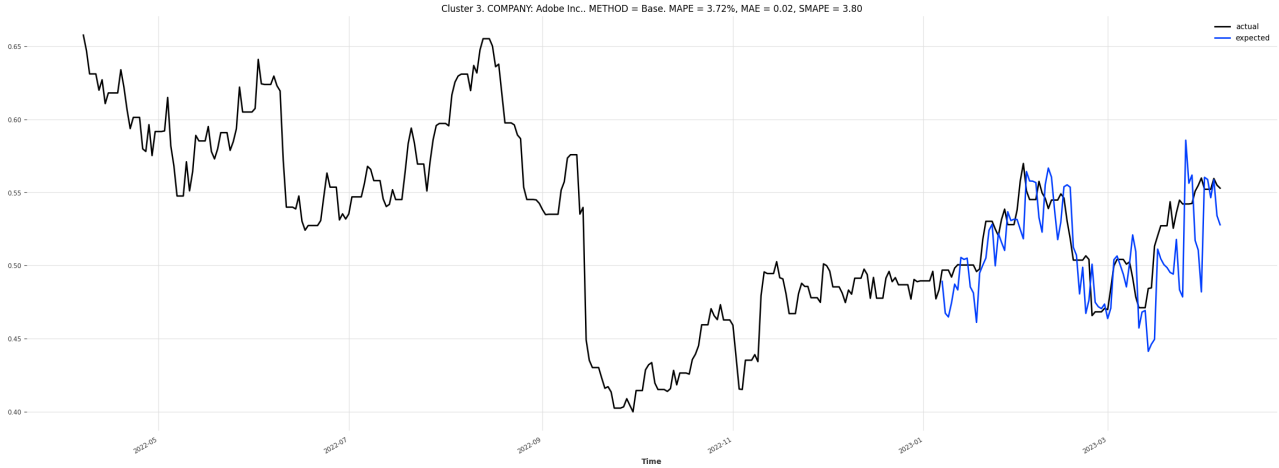

Figure 32: Cluster 3: Adobe Inc. [Model trained on only one time series]

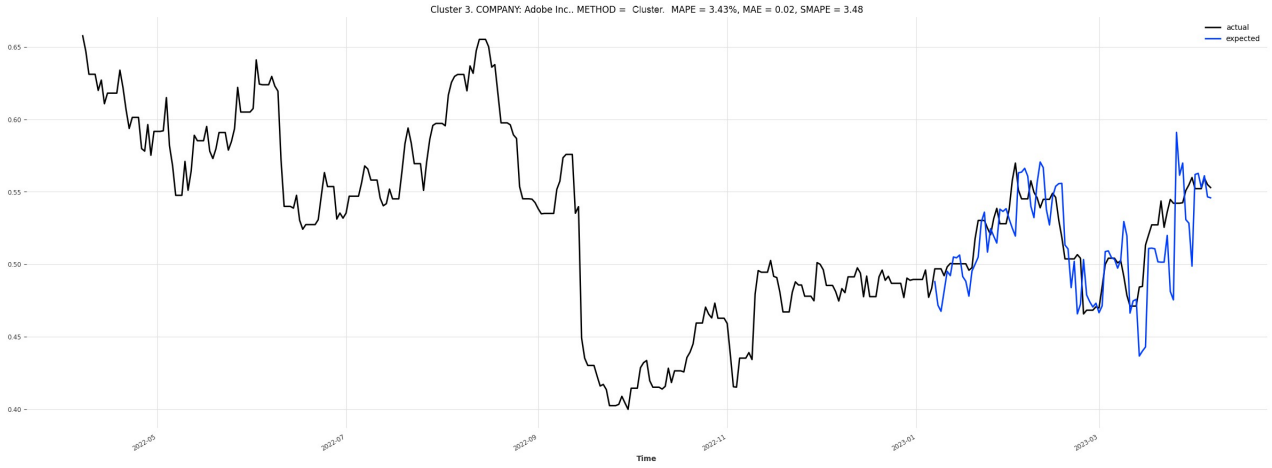

Figure 33: Cluster 3: Adobe Inc. [Model trained on data of companies from the same cluster]

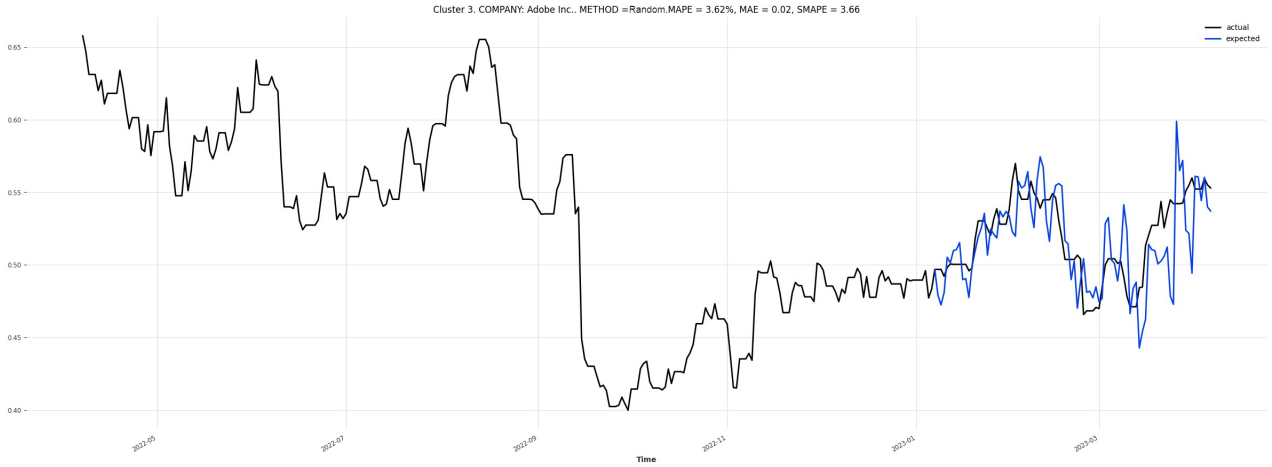

Figure 34: Cluster 3: Adobe Inc. [Model trained on data of companies from random clusters]

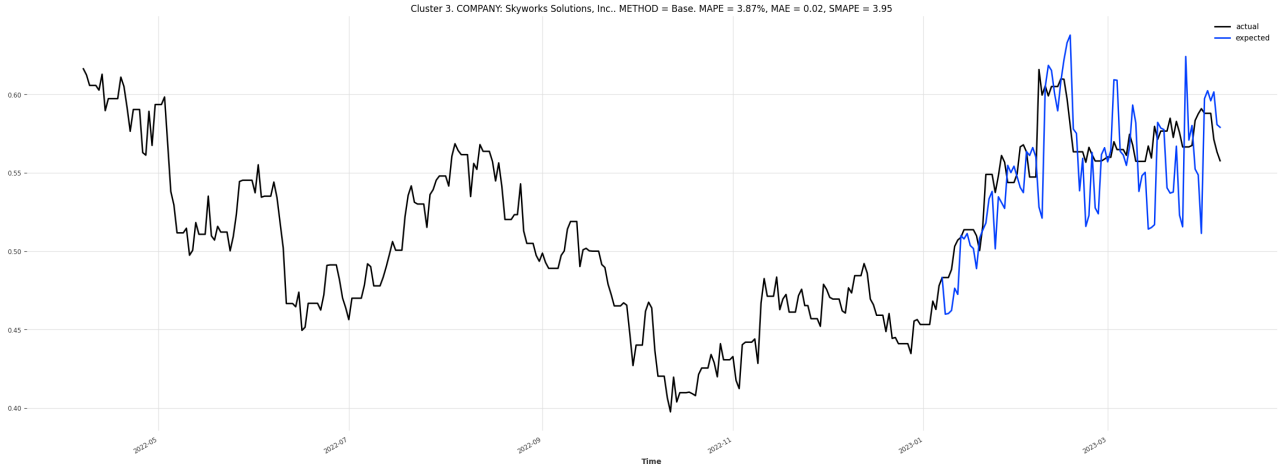

Figure 35: Cluster 3: Skyworks Solutions, Inc. [Model trained on only one time series]

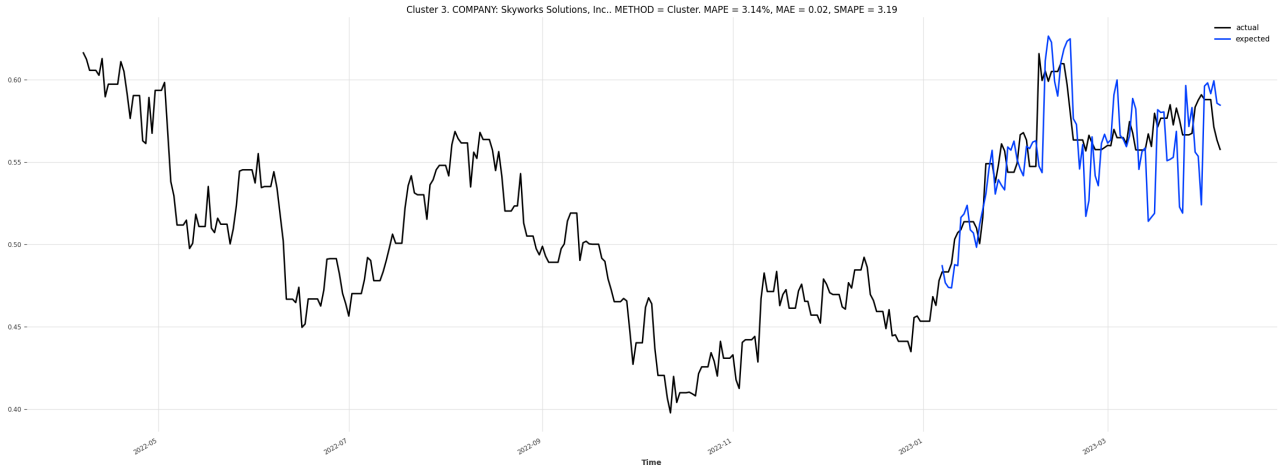

Figure 36: Cluster 3: Skyworks Solutions, Inc. [Model trained on data of companies from the same cluster]

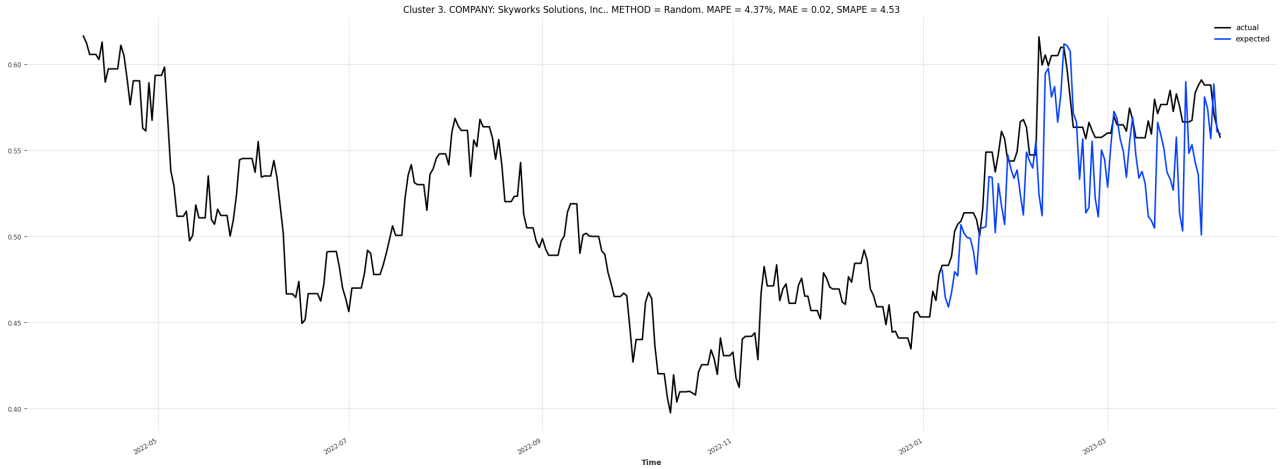

Figure 37: Cluster 3: Skyworks Solutions, Inc. [Model trained on data of companies from random clusters]

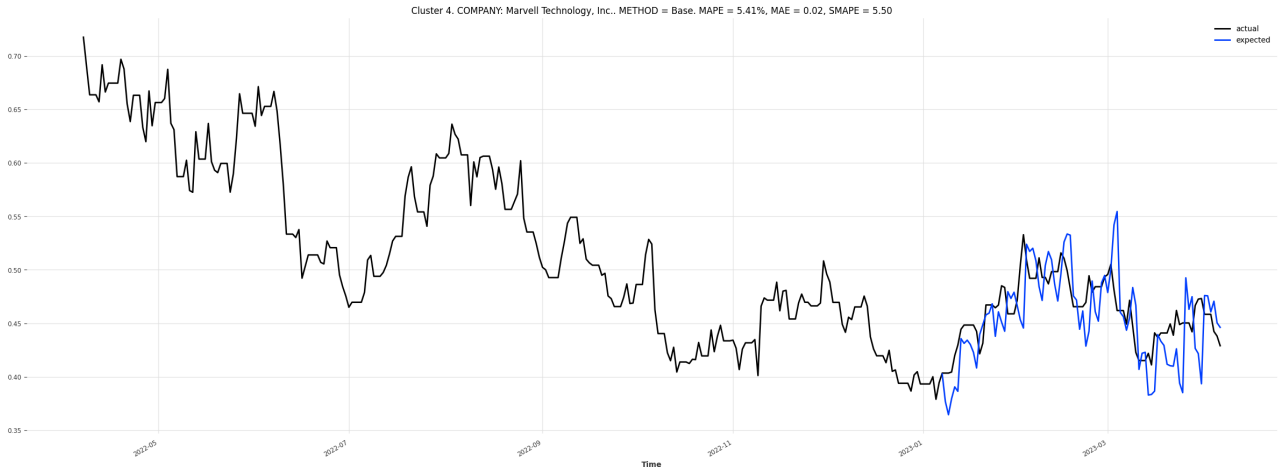

Figure 38: Cluster 4: Marvell Technology, Inc. [Model trained on only one time series]

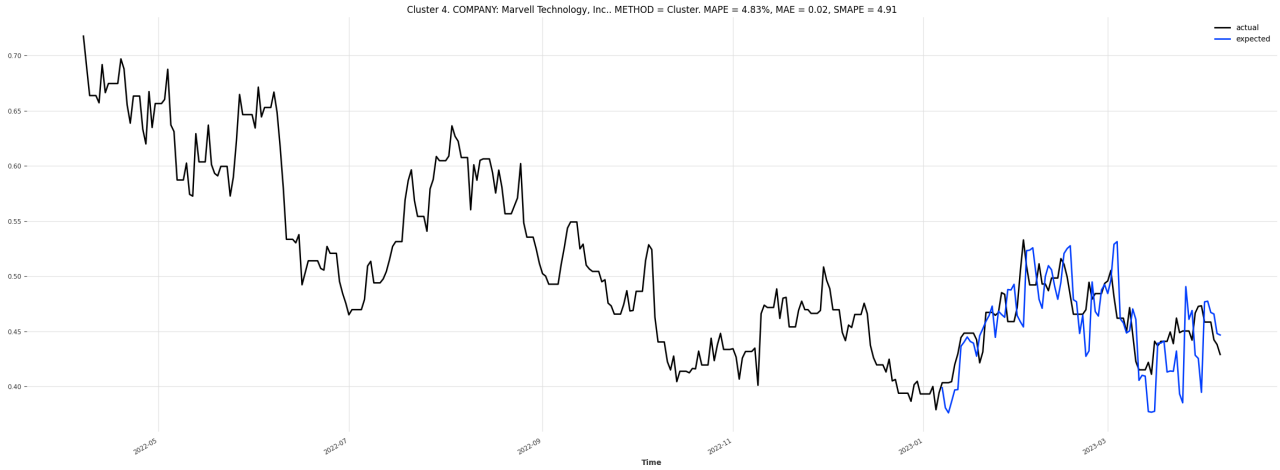

Figure 39: Cluster 4: Marvell Technology, Inc. [Model trained on data of companies from the same cluster]

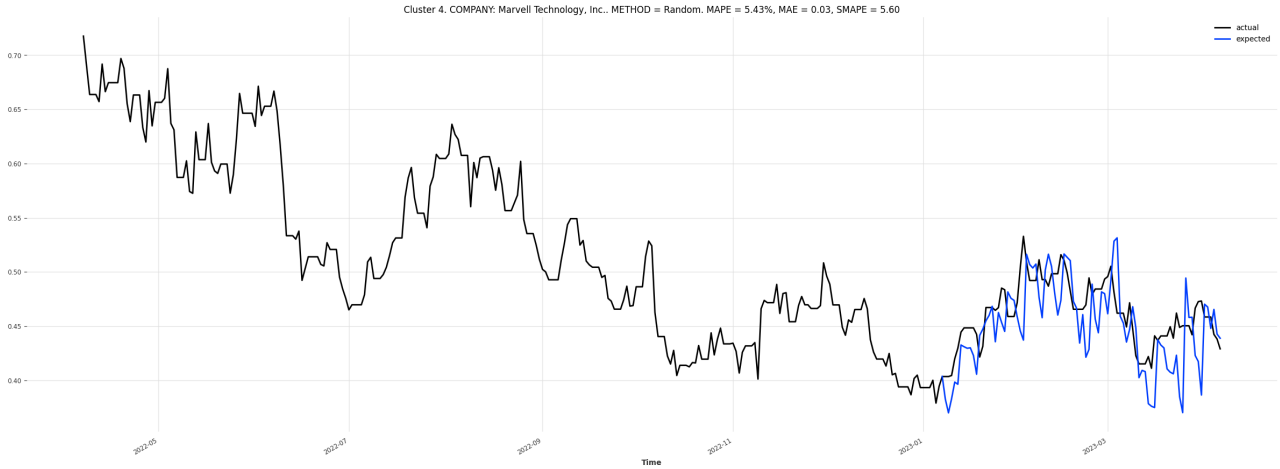

Figure 40: Cluster 4: Marvell Technology, Inc. [Model trained on data of companies from random clusters]

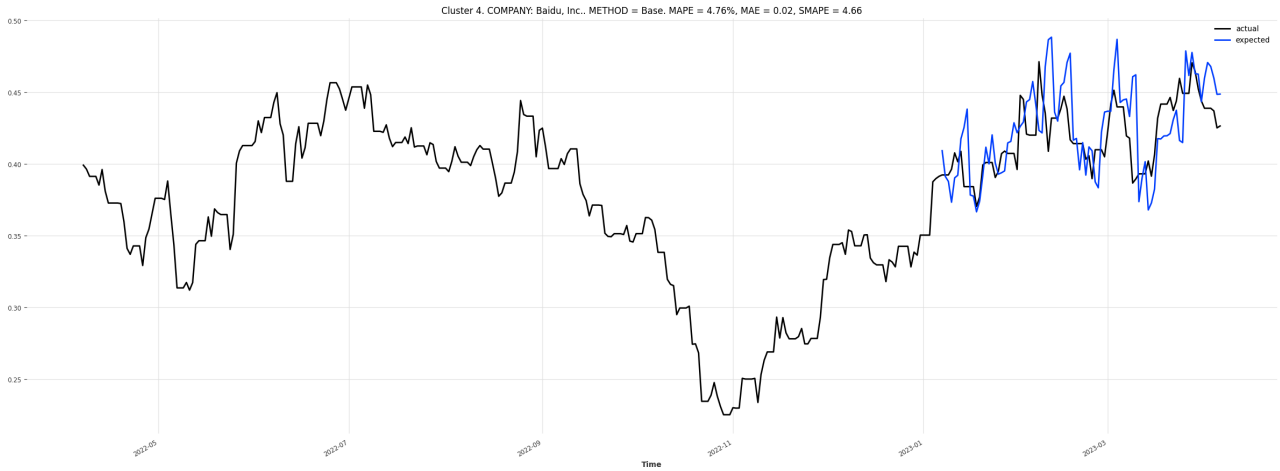

Figure 41: Cluster 4: Baidu, Inc. [Model trained on only one time series]

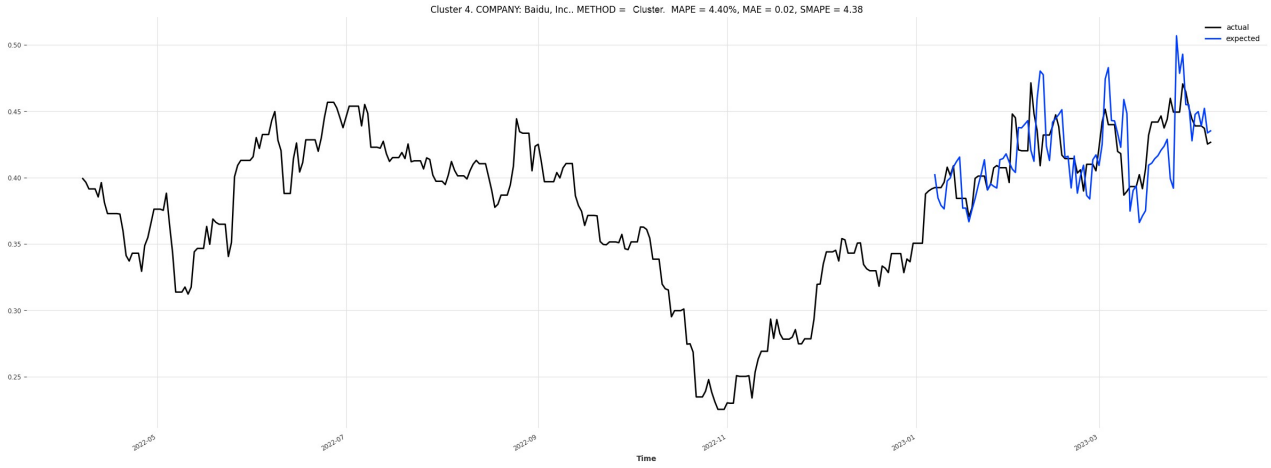

Figure 42: Cluster 4: Baidu, Inc. [Model trained on data of companies from the same cluster]

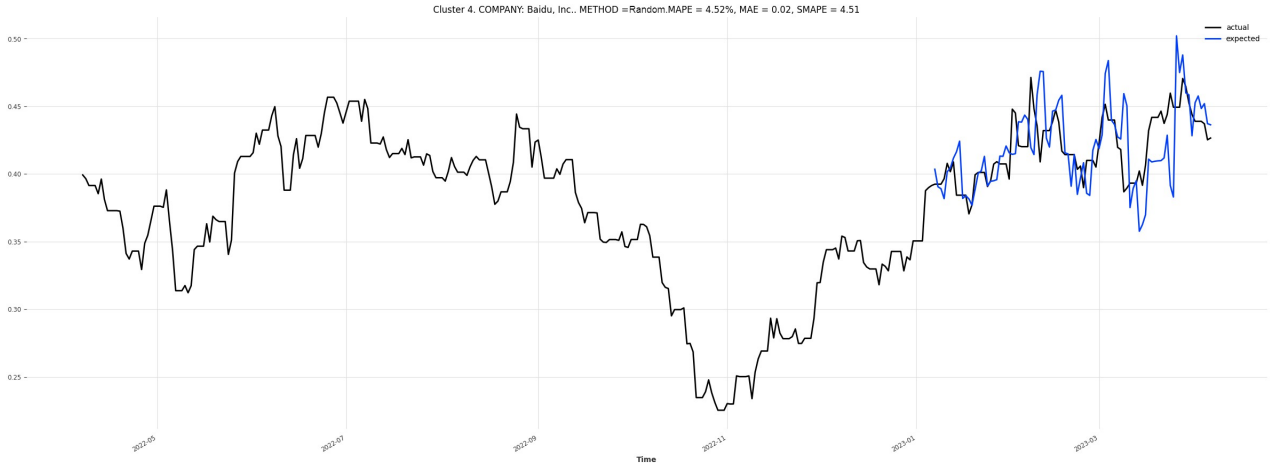

Figure 43: Cluster 4: Baidu, Inc. [Model trained on data of companies from random clusters]

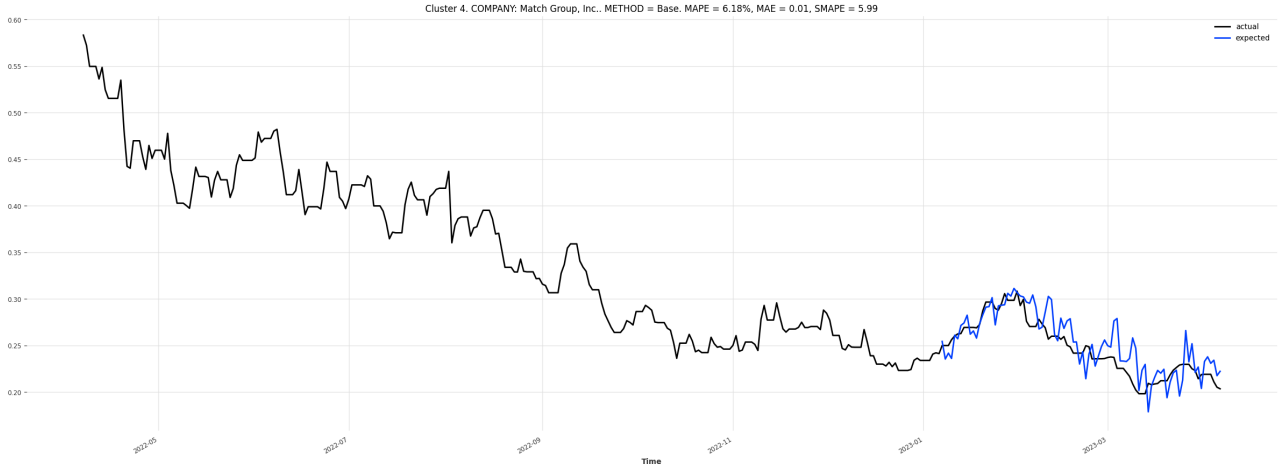

Figure 44: Cluster 4: Match Group, Inc. [Model trained on only one time series]

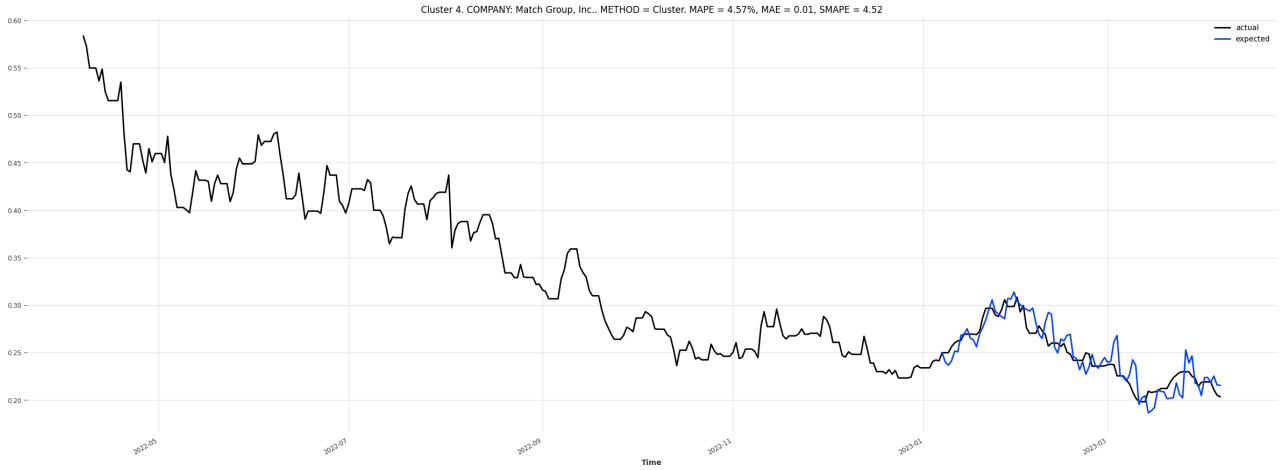

Figure 45: Cluster 4: Match Group, Inc. [Model trained on data of companies from the same cluster]

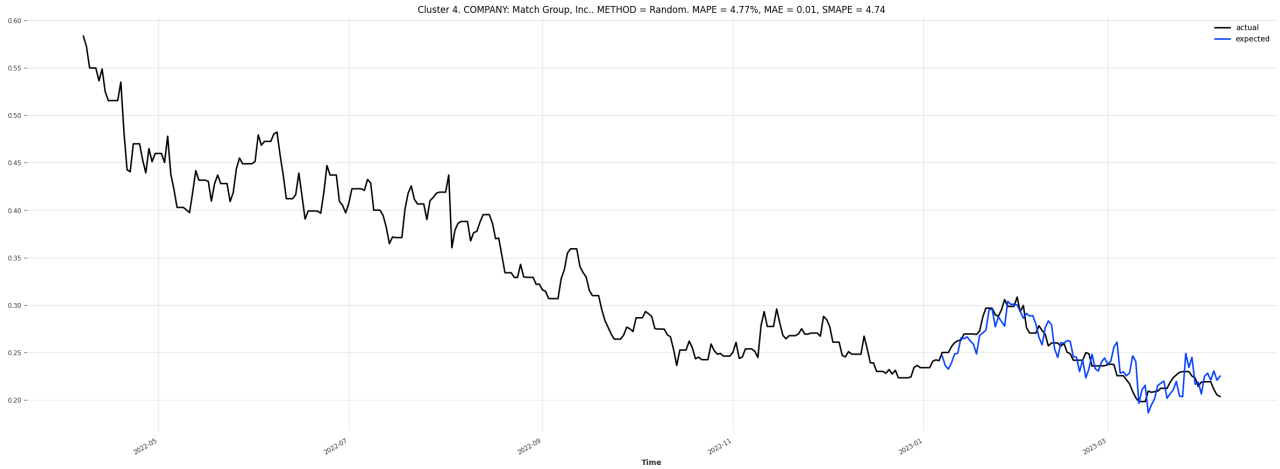

Figure 46: Cluster 4: Match Group, Inc. [Model trained on data of companies from random clusters]

Thus, the conducted work has shown that clustering of companies based on their time series can be of great practical importance for analyzing financial markets and predicting trends in them in the future.

5 Conclusions

In this paper, a study was conducted on the topic of clustering of time series in order to improve predictions. Various methods of clustering time series, as well as prediction algorithms were considered.

Among the clustering algorithms considered, the DTW-based hierarchical clustering method proved to be particularly effective. The incomparable advantage of the algorithm is that the parameter for regulating the number of clusters in this case is the average distance between clusters. By changing this parameter, it is much easier to achieve the optimal number of clusters. The distance metric also plays a huge role in any clustering algorithm, as it is important that it takes into account the shift in two time series when calculating the distance.

The hypothesis of this study was confirmed. Training on companies from the same index can really improve the forecast. In this case, additional time series become augmentations to the original series.

Overall, the results of this study show the importance of using clustering to improve prediction accuracy in machine learning. Further research can be carried out on a large data set and using other clustering methods and distance metrics between rows to achieve higher prediction accuracy.

References

[1] S. Aghabozorgi, A. S. Shirkhorshidi, and T. Wah. Time-series clustering - a decade review. *Information Systems*, 53, 05 2015.

[2] D. Gunopulos and G. Das. Time series similarity measures and time series indexing (abstract only). *ACM SIGMOD Record*, 30:624, 06 2001.

[3] T. Liao. Clustering time series data — a survey. *Pattern Recognition*, 38:1857–1874, 11 2005.

[4] B. Lim, S. Arık, N. Loeff, and T. Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37, 06 2021.

[5] S. Rani and G. Sikka. Recent techniques of clustering of time series data: A survey. *International Journal of Computer Applications*, 52:1–9, 08 2012.

[6] S. Selvin, R. Vinayakumar, E. A. Gopalakrishnan, V. K. Menon, and K. P. Soman. Stock price prediction using lstm, rnn and cnn-sliding window model. 2017.