International Journal for Housing Science and Its Applications
Publish September 11, 2025. Volume 47, Issue 1 Pages 852-864

https://doi.org/10.70517/ijhsa47174

Construction and Application Analysis of Macroeconomic Forecasting Models Based on Time Series Cluster Analysis

Dezhi Yang[1,*]
[1] College of Science, Eastern Liaoning University, Dandong, Liaoning, 118003, China
Corresponding authors: (e-mail: 18704155188@163.com).

Abstract Financial time series data exhibit characteristics such as complexity and high noise levels, rendering traditional clustering methods prone to limitations when processing such data. Consequently, this paper introduces rough set clustering, leveraging its advantages in financial time series analysis and forecasting to construct the macroeconomic prediction model required for research. Principal component analysis is employed to transform high-dimensional data into low-dimensional representations, thereby reducing the complexity of the prediction model. A comprehensive digital economy forecasting model is constructed, revealing its data processing workflow. This workflow emphasizes calculating the fitness of each solution and designing targeted fitness functions. Quantitative analysis across multiple datasets is conducted on the digital economy forecasting model, comparing traditional K-means clustering with the proposed model using economic indicators from multiple cities. Empirical applications demonstrate that the clustering results from the proposed macroeconomic forecasting model better reflect the developmental tiers of 36 cities, highlighting the significant value of rough clustering in financial time series analysis and forecasting.

Index Terms time series data; rough clustering; digital economy forecasting; principal component analysis

I. Introduction

In recent years, the global economic landscape has undergone profound adjustments, with increasing uncertainty and complexity [1]. The stable development of the macroeconomy is crucial for all nations, making accurate macroeconomic forecasting an urgent need for governments, businesses, and investors [2]-[4]. Macroeconomic forecasting models serve as tools to anticipate macroeconomic trends. By collecting various economic data—such as GDP, inflation rates, and unemployment rates—and applying specific algorithms and mathematical formulas, these models project future economic trajectories. This enables governments, businesses, and investors to gain advance insight into the general economic direction, thereby facilitating more informed decision-making [5]-[8]. Governments can adjust policies based on these forecasts, businesses can plan production and investments, and investors can determine capital allocation [9], [10].

As a classical data analysis method, time series clustering possesses unique advantages in handling data with temporal sequences and has been widely applied in macroeconomic forecasting [11-12]. Time series clustering is a method that categorizes similar time series data into the same group [13]. It helps us understand relationships between data, uncover hidden dynamic patterns, and classify and predict sequences [14-15]. Through clustering analysis, time series data can be partitioned into multiple groups, where sequences within each group exhibit greater similarity, while sequences across different groups show significant divergence [16], [17]. Macroeconomic forecasting models built upon time series clustering analysis predict outcomes based on the temporal patterns of economic data [18], [19]. It primarily analyzes characteristics such as the trend of economic variables over a past period, cyclical fluctuations, and seasonal variations [20], [21]. It assumes that past trends will continue into the future. For example, by observing GDP data from previous years and identifying certain seasonal fluctuations and long-term growth trends, the model can utilize these features of historical data to predict future GDP values [22]-[25]. Unlike econometric models, it does not require detailed analysis of complex causal relationships between economic variables. Instead, it focuses solely on the time-series characteristics of the data itself, employing relatively straightforward computational methods. This simplicity makes it widely applicable for short-term economic forecasting [26]-[28].

Reference [29] identifies limitations in current macroeconomic forecasting models and proposes a time-series clustering-based macroeconomic forecasting model. It demonstrates that this model effectively enhances decision-making accuracy, contributing to economic stability and development. Reference [30] introduces a similarity-based macroeconomic forecasting method, applying it to simulated data from a set of Monte Carlo experiments and to the forecasting of a series of key macroeconomic indicators. Results indicate this approach outperforms other time-varying forecasting methods. [31] examines a clustering proposal for autoregressive nonlinear time series data. Experiments demonstrate that this method yields identical results across different forecast horizons, while the dataset exhibits diverse cluster structures during periods of severe contraction in economic activity across all countries. Reference [32] analyzes multiple feature selection methods aimed at enhancing the predictive accuracy of macroeconomic forecasting models. Through comparative evaluation, it reveals that stepwise selection, tree-based, and similarity-based methods outperform others, emphasizing the significance of combining similarity measures with traditional feature selection techniques for improving model reliability. Reference [33] proposes a clustering-based causal feature selection algorithm for multi-time-series forecasting. By contrasting this approach with widely used dimensionality reduction and feature selection methods, it highlights the algorithm's improved predictive accuracy on test datasets. Reference [34] employs artificial neural networks to conduct macroeconomic simulation and forecasting analysis on industrial output value samples from a specific province over a defined period, revealing the neural networks' strong predictive precision. Reference [35] investigates the co-movement of macroeconomic variables within the European Union and its underlying drivers, examining variables such as GDP and employment. The above studies propose macroeconomic forecasting methods based on time series, similarity-based approaches, and multi-feature selection, demonstrating superior predictive efficiency and accuracy.

This paper analyzes the fundamental theories and methodological frameworks of clustering analysis, emphasizing the application steps of clustering methods and identifying commonly used inter-cluster distance retention metrics. It introduces rough clustering methods and leverages their advantages in various financial time series analysis and forecasting scenarios to construct a macroeconomic forecasting model framework. By integrating the dimension reduction characteristics of principal component analysis (PCA), a data processing workflow is designed for the forecasting model, leading to the development of a digital economy forecasting model. A multi-category time series dataset is selected for model validation using multiple clustering methods, highlighting the evaluation performance of the proposed model. Finally, the application effectiveness of this model is analyzed and compared using economic indicator data from multiple cities.

II. Fundamental Theory and Model Construction

II. A. Cluster Analysis

Cluster analysis is the process of dividing a set of data objects into several subsets according to specific rules. Each subset forms a cluster, or category, such that data objects within the same cluster are similar to each other, while those across different clusters are dissimilar. The collection of clusters generated by cluster analysis is referred to as a clustering [36], [37].

In this context, different clustering methods applied to the same dataset may yield different clusters. The partitioning is performed not by humans but by clustering algorithms.

As a data mining function, clustering analysis can serve as a standalone tool to gain insights into data distribution, observe the characteristics of each class, and focus further analysis on specific cluster sets.

II. A. 1) Basic Theory of Cluster Analysis

Providing an exact definition of clustering is challenging, as no unified definition currently exists. This paper employs a mathematical description to illustrate the definition of clustering. Let X denote a dataset, $X=\{x_1,x_2,\cdots,x_n\}$, R is a clustering defined on X that partitions X into m subsets $C_1,C_2,\cdots,C_m$ satisfying the following three conditions:

(a) $C_1 \neq \varnothing, i=1,\cdots,m$.

(b) $\bigcup_{i=1}^{m} C_i = X$.

(c) $C_i \bigcap C_j = \varnothing, i,j=1,\cdots,m$ for all $i \neq j$.

The first principle states that each subclass set is non-empty. The second principle states that the union of all subclass sets equals the entire set. The third principle states that the intersection of any two subclasses is empty. Therefore, every data point in the dataset must belong to exactly one class.

Typically, clustering analysis algorithms consist of four steps:

Step 1: Feature extraction and selection, which aims to obtain data distinguishing different attributes between objects and reduce redundancy.

Step 2: Calculate similarity measures. This involves computing the similarity between objects by calculating the distance between their features.

Step 3: Grouping. Objects are categorized based on similarity or distance, grouping similar or proximate objects together while assigning dissimilar or distant objects to separate clusters.

Step 4: Presentation of clustering results. The grouping information of objects is output, and clustering results can also be visualized graphically.

In cluster analysis, the conventional approach to describing “differences” involves measuring distances between clusters. The similarity between two cases is inversely proportional to their distance $d(i,j)$. Suppose there are n cases. The m attribute values describing the i th case correspond to variable values $x_{i1}, x_{i2}, \cdots, x_{im}$; and the m attribute values describing the j th case correspond to variable values $x_{j1}, x_{j2}, \cdots, x_{jm}$. The distance between them is $d(i,j)$. Several methods exist for calculating this distance, including the Ming distance, defined by the following formula:

$$d(i,j)=\sqrt[q]{\sum_{k=1}^{m}| x_{ik}-x_{jk} |^{q}} \tag{1}$$

When q=1, it is called the absolute distance.

When q=2, it is called the Euclidean distance.

When q approaches infinity, it is called the Chebyshev distance.

Among these, the Euclidean distance is the most commonly used; most clustering methods default to this distance metric.

A frequently used similarity measure is the similarity coefficient, defined as:

$$x=(x_1,x_2,\cdots,x_m)', y=(y_1,y_2,\cdots,y_m)' \tag{2}$$

(1) Cosine angle coefficient: The similarity between samples can be measured by their cosine angle, as shown in the following formula:

$$\cos(x,y)=\frac{x^{'}y}{[(x^{'}x)(y^{'}y)]^{1/2}} \tag{3}$$

(2) Pearson correlation coefficient: The correlation coefficient is the cosine of the angle between standardized data, as shown in the following formula:

$$r(x,y)=\frac{(x-\overline{x})^{'}(y-\overline{y})^{'}}{[(x-\overline{x})^{'}(x-\overline{x})(y-\overline{y})^{'}(y-\overline{y})]^{1/2}} \tag{4}$$

In the similarity measures above, the greater the similarity between two samples, the higher the value, with a maximum value of 1.

II. A. 2) Methodological Framework for Cluster Analysis

Types of cluster analysis include systematic clustering, two-step clustering, K-means clustering, and fuzzy clustering, among others.

Systematic clustering comprises two approaches: agglomerative and divisive. Divisive clustering begins with all data points as a single cluster. It then recursively decomposes the cluster based on similarity between data points, forming progressively smaller clusters. This process continues until each cluster contains only one data point, or until a predetermined termination criterion is met. Agglomerative clustering operates in the opposite manner, starting with each data point as its own cluster. It progressively aggregates data points upward based on similarity between clusters, forming increasingly larger clusters until a single cluster encompassing all data points is achieved, or clustering terminates upon meeting specific conditions.

Regardless of whether agglomerative or divisive methods are used, a crucial metric is the distance between two clusters. Commonly used inter-cluster distance measures include several approaches:

Shortest Distance Method: Defines the distance between classes C_i and C_j, where $| p-p |$ is the distance between point p and $p^{'}$. The formula is as follows:

$$dist_{\min}(C_i,C_j)=\min_{p\in C_i, p^{'}\in C_j}\{| p-p^{'} |\} \tag{5}$$

When an algorithm uses maximum distance to measure inter-class distance, it is called the maximum distance method, with the formula as follows:

$$dist_{\max}(C_i, C_j) = \max_{peC_i, p^{'}eC_j}\{|\, p - p^{'}\,|\} \tag{6}$$

The minimum and maximum distance measures represent two extremes of inter-class distance, tending to be overly sensitive to outliers or noisy data. Using the mean distance or average distance offers a compromise between minimum and maximum distances, helping mitigate sensitivity to outliers. m_i denotes the mean of class C_i, while n_i represents the number of objects in class C_i.

The mean distance is calculated as follows:

$$dist_{mean}(C_i, C_j) = |\, m_i - m_j\,| \tag{7}$$

Average distance, calculated using the following formula:

$$dist_{avg}(C_i, C_j) = \frac{1}{n_i n_j}\sum_{p \in C_i, p^{'} \in C_j} |\, p - p^{'}\,| \tag{8}$$

II. B.Application of Rough Fuzzy Clustering in Macroeconomic Forecasting

Rough clustering is an unsupervised clustering method based on rough set theory. It handles uncertainty in data by dividing data objects into imprecise clusters. Unlike traditional clustering methods such as K-means, rough clustering allows data objects to belong to multiple clusters, thereby better reflecting the inherent structure of data in certain scenarios.

The core idea of rough clustering is to partition data objects into a set of equivalence classes and construct upper and lower approximate sets for each cluster based on these classes. The upper approximate set contains all objects potentially belonging to the cluster, while the lower approximate set includes only those objects definitively belonging to it. By calculating the discrepancy between upper and lower approximations, the incompleteness, uncertainty, and fuzziness of a cluster can be assessed.

Fuzzy clustering finds applications in various scenarios of financial time series analysis and forecasting, including macroeconomic forecasting, stock market analysis, and financial crisis early warning. Analyzing these use cases provides valuable insights for decision-making in finance-related fields.

Rough clustering can be applied to analyze and forecast macroeconomic time series data. For instance, a series of time series indicators related to national economic growth—such as GDP growth rate, unemployment rate, and inflation rate—are closely monitored by governments and economists. Rough clustering enables grouping of economic indicators with similar characteristics, thereby helping policymakers and economists gain a more comprehensive understanding of the macroeconomic landscape and formulate more effective economic policies.

In macroeconomic forecasting applications, rough set theory can partition time series data into lower approximation sets, upper approximation sets, and boundary regions. This identifies economic indicators with similar trends and periodicity, enabling macroeconomic analysis and prediction. By integrating other forecasting methods—such as quantile regression with random forests (QRF) and long short-term memory neural networks (LSTM)—further analysis can be conducted on rough clustering results. For instance, these methods can forecast future economic growth rates, unemployment rates, and inflation rates, providing governments and economists with insights into upcoming macroeconomic trends.

II. C.Principal Component Analysis Method

Principal Component Analysis (PCA) focuses on non-random variables. Today, PCA has become a widely adopted dimension reduction technique, gaining extensive application across numerous fields due to its operational simplicity and intuitive nature. The core objective of PCA is to explore a new coordinate system that aims to reduce the number of features while preserving as much valuable information from the original data as possible [38], [39].

The representation of raw data in a new coordinate system is essentially regarded as a linear transformation process. Let the original basis space be of dimension n and the new basis space be of dimension m. This transformation can be expressed as follows:

$$\begin{cases} u_1 = a_1^{(1)} x_1 + a_2^{(1)} x_2 + \ldots + a_n^{(1)} x_n \\ u_2 = a_1^{(2)} x_1 + a_2^{(2)} x_2 + \ldots + a_n^{(2)} x_n \\ \ldots\ldots \\ u_m = \alpha_1^{(m)} x_1 + \alpha_2^{(m)} x_2 + \ldots + \alpha_n^{(m)} x_n \end{cases} \tag{9}$$

$x^T = (x_1, x_2, \ldots, x_n)$ denotes the coordinates before transformation, $u^T = (u_1, u_1, \ldots, u_m)$ denotes the transformed coordinates. The parameter vector $(a_1^{(i)}, a_2^{(i)}, \ldots, a_n^{(i)})$ represents the basis for the i th dimension of the transformed coordinate axes. For notational convenience, this relationship is denoted as:

$$\begin{gathered} a^{(i)T} = (a_1^{(i)}, a_2^{(i)}, \ldots, a_n^{(i)}), \\ x^T = (x_1, x_2, \ldots, x_n), \\ u_i = a^{(i)T} x. \end{gathered} \tag{10}$$

At the same time, a matrix can be constructed using the basis vectors of the new coordinate system:

$$A = \begin{pmatrix} a_1^{(1)} & a_2^{(1)} & \ldots & a_n^{(1)} \\ a_1^{(2)} & a_2^{(2)} & \ldots & a_n^{(2)} \\ \ldots\ldots & & & \\ a_1^{(m)} & a_2^{(m)} & \ldots & a_n^{(m)} \end{pmatrix} \tag{11}$$

Next, a common method for evaluating how well principal components preserve initial information is to observe how data points are distributed in the new coordinate system.

Therefore, the optimization objective transforms into: maximizing variance in each dimension while ensuring the covariance between any two dimensions is zero. Based on Equation (11), the following can be derived:

$$var(u_1) = var(\alpha^{(1)T} x) = \alpha^{(1)T} \Sigma \alpha^{(1)} \tag{12}$$

Here, Σ represents the covariance matrix of the original coordinate vector x. The objective is to compute the parameter vector $a^{(1)}$ that maximizes equation (11). However, it is important to note that if no constraints are imposed on the values of the parameter vectors $a^{(i)}$, the value of equation (12) can become infinitely large, rendering the analysis meaningless. Therefore, the parameter vectors $a^{(i)}$ are constrained to unit vectors, transforming the problem of calculating the first principal component into an optimization problem, as shown below:

$$\begin{cases} \max\limits_{a} a^{(1)T} \sum a^{(1)} \\ s.t. a^{(1)T} a^{(1)} = 1 \end{cases} \tag{13}$$

Applying the Lagrange method to the above equation yields Equations (14) and (15):

$$\Sigma a^{(1)} = \lambda a^{(1)} \tag{14}$$

$$a^{(1)T} \Sigma a^{(1)} = a^{(1)T} \lambda a^{(1)} = \lambda. \tag{15}$$

From equation (15), it can be deduced that one eigenvalue of the covariance matrix Σ is λ. The eigenvector corresponding to the eigenvalue λ is $a^{(1)}$. Multiplying both sides of equation (15) by $a^{(1)T}$ from the left yields equation (16).

The second parameter vector $a^{(2)}$ can be obtained using a similar approach, simply by replacing the calculation target with the eigenvector corresponding to the second-largest eigenvalue. It can be mathematically proven that eigenvectors associated with different eigenvalues are orthogonal to each other. Clearly, they are uncorrelated, thus satisfying the required constraints. To determine the number of principal components to retain, one must further consider the variance distribution, analyzing each specific case individually. That is:

$$\lambda_i / \sum_{j=1}^{n} \lambda_j \tag{16}$$

$$\sum_{i=1}^{m}\lambda_i / \sum_{j=1}^{n}\lambda_j \tag{17}$$

Equation (16) provides the calculation method for variance contribution rate (for the i th principal component), while Equation (17) represents the cumulative variance contribution rate of the first m principal components, which are also the m principal components with the highest variance shares. To fully demonstrate the effectiveness of PCA, the cumulative variance contribution rate is typically required to reach 90% or higher.

The specific steps for performing PCA are as follows:

(1) To completely eliminate misleading effects from dimensions, first normalize the initial data and calculate the covariance matrix Σ of the normalized data.

(2) Solve for all eigenvalues of Σ and sort them as $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n > 0$.

(3) Determine the number of principal components m required. This step is primarily achieved by calculating the cumulative variance contribution rate.

(4) Select the m largest eigenvalues and use their corresponding eigenvectors as the coordinate axes $a^{(1)},\ldots,a^{(m)}$ of the new coordinate system.

(5) Calculate the expressions for the m principal components using Equation (17).

II. D.Design of Digital Economy Forecasting Methods

II. D. 1) Model Processing Procedure

This paper integrates several previously introduced methods to construct a digital economy scale prediction model. The data processing workflow of the model is illustrated in Figure 1.

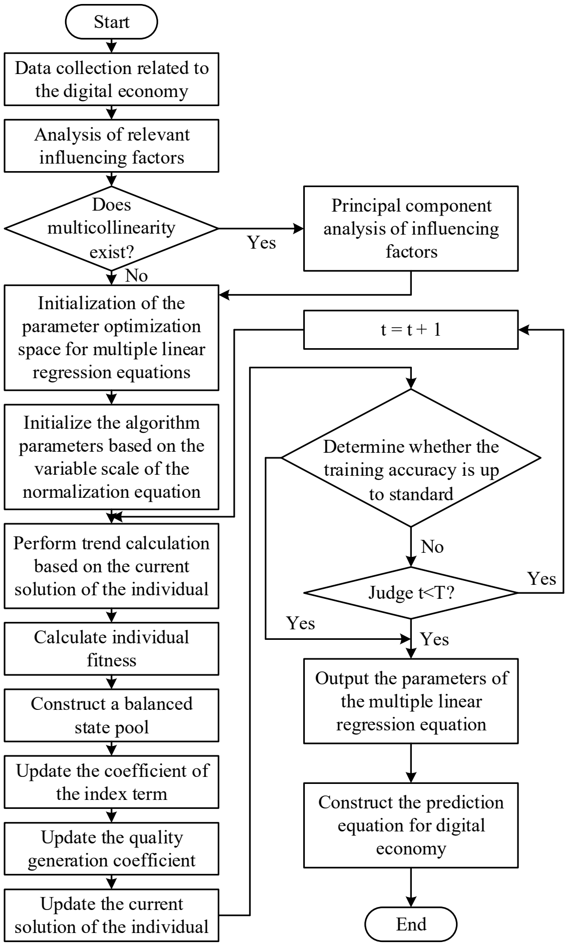
Start
Data collection related to the digital economy
Analysis of relevant influencing factors
Does multicollinearity exist?
Yes
No
Principal component analysis of influencing factors
Initialization of the parameter optimization space for multiple linear regression equations
t = t + 1
Initialize the algorithm parameters based on the variable scale of the normalization equation
Determine whether the training accuracy is up to standard
Perform trend calculation based on the current solution of the individual
No
Yes
Judge t<T?
Yes
Yes
Calculate individual fitness
Output the parameters of the multiple linear regression equation
Construct a balanced state pool
Update the coefficient of the index term
Construct the prediction equation for digital economy
Update the quality generation coefficient
Update the current solution of the individual
End

Figure 1: Data processing process for digital economic scale forecasting model

A crucial step in the data processing workflow of the digital economy scale prediction model involves calculating the fitness of each solution (individual). This fitness calculation requires designing a targeted fitness function. The primary objective of the EO algorithm within the prediction model is to optimize the parameters of the multiple linear regression equation, minimizing the error between the fitted values and actual values to meet the target. The designed fitness function is as follows:

$$Fit(\vec{C})=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(observed_i-predicted_i)^2} \tag{18}$$

Observing the fitness function reveals that this expression represents the formula for calculating the root mean square error (RMSE). RMSE serves as a crucial criterion for assessing the accuracy of a multiple linear regression equation's fit. This paper aims to leverage it to enable the EO algorithm to discover optimal parameter combinations for the multiple linear regression equation. Here, $observed_i$ denotes the actual value of the digital economy scale. $predicted_i$ denotes the predicted value of the digital economy scale based on the current solution $\vec{C}$ in the constructed multiple linear regression equation. N represents the number of samples in the training data.

II. D. 2) Multivariate Linear Prediction Model for Digital Economy Scale

To forecast the future trajectory of the digital economy's scale, this paper designed a multiple linear forecasting model:

$$y=\beta_0+\beta_1\overrightarrow{x_1}+\beta_2\overrightarrow{x_2}+\beta_3\overrightarrow{x_3}+\beta_4\overrightarrow{x_4}+\beta_5\overrightarrow{x_5}+\varepsilon \tag{19}$$

Among these, $\overrightarrow{x_1},\overrightarrow{x_2},\cdots,\overrightarrow{x_5}$ represent the five principal components. As seen from Equation (19), the model includes seven parameters requiring determination: $\beta_0,\beta_1,\cdots,\beta_5$ and ε.

To obtain these seven parameter values, this paper employs a balanced optimizer for parameter optimization, with the specific process illustrated in Figure 1. During the optimization process, the population size $n=60$ was set for the Balanced Optimizer, with the upper bound of the search interval defined as $\vec{C}_{\max}=1$ and the lower bound as $\vec{C}_{\min}=0$. Additionally, to prevent overfitting, the fitting target $\sigma=0.07$ was established.

Through multiple iterations of the balanced optimizer, the values of these seven parameters were determined, establishing the normalized digital economy scale prediction model, expressed as follows:

$$y=0.0508+0.3327\overrightarrow{x_1}+0.0027\overrightarrow{x_2}+0.0065\overrightarrow{x_4}+0.4852 \tag{20}$$

Table 1: Detailed information on using data sets in experiments

Dataset	c	n	z	q
Beef (BEE)	6	50	420	100.1
Meat (MEA)	4	150	436	10.1
Wine (WIN)	3	102	220	3.1
MiddlePhalanxOutlineAgeGroup (MPA)	2	500	60	10.1
MiddlePhalanxTW (MPT)	6	551	80	100.1
OSULeaf (OSU)	5	450	430	100.1
ProximalPhalanxOutlineAgeGroup (PPA)	3	600	60	100.1
ProximalPhalanxTW (PPT)	5	600	80	100.1
SwedishLeaf (SWL)	12	1220	120	2.0
WordSynonyms (WDS)	22	900	260	10.1
DiatomSizeReduction (DSR)	3	350	320	100.1
Wafer (WAF)	2	7000	140	2.1
ChlorineConcentration (CHL)	3	4200	155	2.1
Lightning2 (LI2)	2	122	626	100.1
SmoothSubspace (SMT)	3	300	14	2.1
BME (BME)	2	150	120	2.1

III. Model Validation and Empirical Analysis

III. A. Model Validation

III. A. 1) Experimental Dataset

To validate the superiority and effectiveness of the proposed model, experimental evaluations were conducted on multiple widely used UCR time series datasets.

Detailed information about the datasets used in the experiments is presented in Table 1. The following table lists the experimental datasets and their parameter settings for the proposed model. Parameters were determined by algorithmically traversing the range with a step size of 1 to identify the optimal values as their final settings.

III. A. 2) Comparison Algorithm Description

To evaluate the clustering performance of the proposed algorithm, this paper compares it with several existing clustering algorithms. These include two classical clustering algorithms: FCM and K-means. Two traditional time series clustering algorithms: TS3C and R-Clustering. Three time series clustering algorithms based on deep learning frameworks: SDCN, DSC, and DTC. Additionally, the study examines U-shapelet, a time series subsequence clustering algorithm, and FTRR, a subspace clustering algorithm for high-dimensional data. Both FCM and K-means algorithms utilize the scikit-learn library available in Python.

Following the parameter specifications outlined in the respective literature for each comparison algorithm, 100 tests were conducted for each algorithm. The average results obtained under the optimal parameters were recorded as the final outcomes.

III. A. 3) Cluster Evaluation Metrics

In the experiment, the classical evaluation metric Rand Index (RI) was used to compare the clustering results of different algorithms, including:

$$RI = \frac{TP + TN}{TP + TN + FN + FP} \tag{21}$$

True positive (TP) refers to two similar objects being assigned to the same cluster. True negative (TN) refers to two dissimilar objects being assigned to different clusters. False positive (FP) and false negative (FN) refer respectively to two dissimilar objects being assigned to the same cluster and two similar objects being assigned to different clusters.

The metric ranges from [0,1]. A higher value indicates greater similarity between the clustering results and the true categories, signifying higher clustering quality.

III. A. 4) Experimental Analysis

The clustering results of the proposed model and comparison algorithms on multiple real-world datasets are shown in Figure 2. The figure displays clustering results from various algorithms on the Image dataset, where the maximum and minimum values correspond to the proposed model (0.9825) and SDCN (0.0918), respectively.

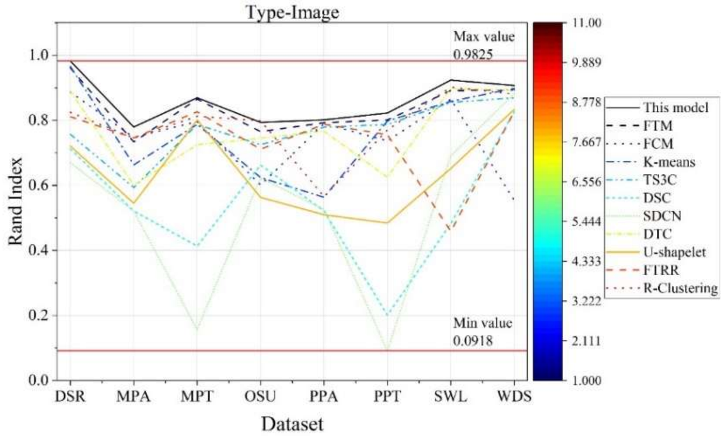

Figure 2: The clustering results of the comparison algorithm on multiple data sets

Figure 3 shows the quantitative evaluation metrics of different clustering algorithms on the Spectro-type time series dataset, where the proposed model achieves the optimal values on this dataset.

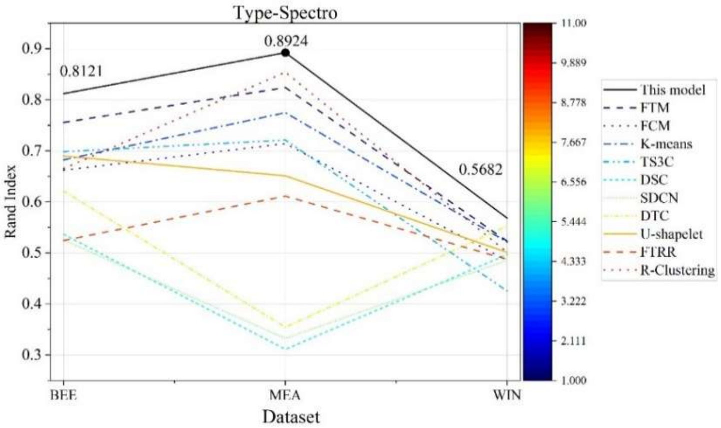

Figure 3: Quantitative evaluation indicators on the Spectro series data set

Figure 4 shows the quantitative evaluation metrics of different clustering algorithms on the Sensor-type time series dataset. The proposed model consistently achieves the best performance on this dataset. By evaluating the performance of various clustering models across Image-type, Spectro-type, and Sensor-type time series datasets, the universality of the proposed model is validated.

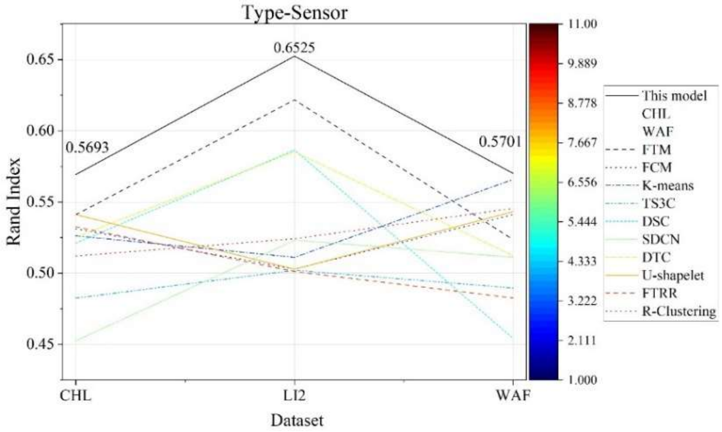

Figure 4: Quantitative evaluation indicators on the Spectro series data set

III. B. Cluster Analysis Based on Principal Component Analysis and Rough Sets

III. B. 1) Perform k-means clustering on the experimental data

The primary data analyzed in this study consists of economic indicators from multiple cities, encompassing 36 samples across provincial capitals, municipalities directly under the central government, and cities with independent planning status. The dataset comprises 22 attributes, including annual total population and urban gross domestic product.

The sample data from the 36 cities was imported into SPSS software for K-means clustering. SPSS is a multifunctional statistical software capable of performing various data analyses. Before importing the data into Excel, the first row must be designated as the attribute row. Upon importing into SPSS, a prompt will appear. Select the first row as the attribute row. After importing into SPSS, the first row will directly appear in the attribute row of the SPSS interface. In this sample dataset, there are 36 samples representing 36 cities, including provincial capitals and cities with independent planning status. There are 22 attributes, representing 22 economic indicators for the 36 cities.

In the SPSS analysis, K-means clustering was directly applied to the sample data. The final cluster centers are shown in Table 2. K-means clustering yielded the final clusters and clustering results, dividing the 36 cities into 3 categories.

Table 2: Final cluster center

Attribute	Clustering		
	1	2	3
Population	1253	1233	645
Regional GDP	13200	6842	2502
First crop	117.5	235.3	142.3
Second production	4523.1	3211.5	1241.9
Tertiary industry	9053.6	3526.7	1156.8
Freight volume	49863.2	41222.3	14563.2
Local budget revenue	23631215	6452311	2013984
Local budget expenditure	27545123	8954273	2965141
Total investment in fixed assets	52613895	36985411	16583421
Total savings of urban and rural residents	144568459	54900536	18323678
The average salary of the workers in the field	62321.3	43652.2	35521.21
The number of post offices of late year	725	765	314
Number of users at the end of the year	952.3	486.5	198.2
Total social commodity retail	52565836	24525463	11235568
Total import and export	25632385	80213614	1142793
There are number of vehicles at the end of the year	19563	11214	3658
Theatre number	135	40	22
The number of students in the average higher school	556523	487545	366319
Hospital health	782	565	243
Practitioner number	52563	32126	14204
The acquisition of product output	112436	401186	71189

The K-means clustering results are shown in Table 3. The data from 36 cities were grouped into three categories: Category 1 includes Beijing and Shanghai. Category 2 includes Tianjin, Hangzhou, Guangzhou, Shenzhen, Chengdu, and Chongqing.

Table 3: K-mean clustering results

Class number	City	Case number
First class	Beijing, Shanghai	2.0
Second type	Tianjin, Hangzhou, Guangzhou, Shenzhen, Chengdu, Chongqing	6.0
Third class	Shijiazhuang, Taiyuan, Hohhot, Shenyang, Dalian, Changchun, Harbin, Nanjing, Ningbo, Hefei, Fuzhou, Xiamen, Nanchang, Jinan, Qingdao, Zhengzhou, Wuhan, Changsha, Nanning, Haikou, Guiyang, Kunming, Lhasa, Xi'an, Lanzhou, Xining, Yinchuan, Urumqi.	28.0

III. B. 2) Clustering Model for Experimental Data

In this experiment, 36 sample data points were first subjected to principal component analysis (PCA) for dimensionality reduction. The original dataset comprised 22 attributes. Following PCA, the dimensions were reduced to three principal components. Subsequently, dynamic clustering based on rough sets was performed on the 36 cities, yielding the dynamic clustering centers and the final clustering results, including the upper and lower approximate sets of each cluster.

The final cluster centers are shown in Table 4. For the first cluster, the lower approximate set is no less than 2, while the upper approximate set exceeds 1.997.

Table 4: Final cluster center

Principal component		Clustering		
		1	2	3
First class	Lower approximation set	2.564	2.215	2.003
	Upper approximation set	2.236	2.193	1.997
Second type	Lower approximation set	1.993	1.625	1.483
	Upper approximation set	1.326	1.212	0.896
Third class	Lower approximation set	1.725	1.236	0.851
	Upper approximation set	0.901	0.972	0.683

Using the prediction model proposed in this paper, clustering results based on principal components and rough sets were obtained. The clustering model results are shown in Table 5.

In the clustering model, cities are divided into three categories. The first category represents developed cities, with 6 cities in the lower approximate cluster and 6 cities in the upper approximate cluster. The second category comprises moderately developed cities, with 25 cities in the lower approximate cluster and 29 cities in the upper approximate cluster. The third category consists of underdeveloped cities, with 7 cities in the lower approximate cluster and 10 cities in the upper approximate cluster.

Table 5: Cluster model results

Class number		Object number
First class	Lower approximation set	6
	Upper approximation set	6
Second type	Lower approximation set	25
	Upper approximation set	29
Third class	Lower approximation set	7
	Upper approximation set	10

Based on the number of objects in the model clustering of this paper, specific city classifications are obtained. The clustering results of this model are shown in Table 6.

The K-means clustering results categorize cities into three groups. From this outcome, it is evident that the second category includes cities such as Tianjin, Hangzhou, Guangzhou, Shenzhen, Chengdu, and Chongqing. These cities form a cluster that is more developed than the first category. The third category comprises the remaining cities. This clustering reveals some inconsistencies: cities like Nanjing and Urumqi in Xinjiang exhibit distinct developmental trajectories, yet K-means clustering groups them together. This classification fails to reflect the hierarchical development levels among cities.

Table 6: Clustering results of this model

Class		Cluster result
First class	Lower approximation set	Beijing, Shanghai, Guangzhou, Shenzhen, Chongqing, Chengdu
	Upper approximation set	Beijing, Shanghai, Guangzhou, Shenzhen, Hangzhou, Dalian
Second type	Lower approximation set	Shijiazhuang, Tianjin, Chongqing, Chengdu, Taiyuan, Shenyang, Dalian, Changchun, Harbin, Nanjing, Hangzhou, Ningbo, Hefei, Fuzhou, Xiamen, Nanchang, Jinan, Qingdao, Zhengzhou, Wuhan, Changsha, Nanning, Haikou, Guiyang, Kunming, Xi 'an
	Upper approximation set	Shijiazhuang, Tianjin, Chongqing, Chengdu, Taiyuan, Shenyang, Dalian, Changchun, Harbin, Nanjing, Hangzhou, Ningbo, Hefei, Fuzhou, Xiamen, Nanchang, Jinan, Qingdao, Zhengzhou, Wuhan, Changsha, Nanning, Haikou, Guiyang, Kunming, Xi 'an, Urumqi, Lanzhou
Third class	Lower approximation set	Hohhot, Lhasa, Lanzhou, Xining, Yinchuan, Urumqi
	Upper approximation set	Hohhot, Lhasa, Lanzhou, Xining, Yinchuan, Urumqi, Nanning, Nanchang, Guiyang, Zhengzhou

The clustering results of this paper's model divide the 36 cities into three categories: the first category represents developed regions, the second category represents moderately developed regions, and the third category represents underdeveloped regions. Furthermore, based on rough set-based clustering analysis, the upper approximate set and lower approximate set for each category are obtained. The clustering results clearly distinguish the development levels of cities and align with actual conditions, validating the effectiveness of this clustering model.

IV. Conclusion

This paper develops the application of rough set clustering analysis in macroeconomic forecasting and designs a digital economy forecasting method by integrating principal component analysis. After validating the performance of the digital economy forecasting model, it incorporates economic data from multiple cities to predict their respective economic development.

(1) Multiple time-series datasets encompassing Image, Spectro, and Sensor categories were selected to compare this model against various clustering approaches. Quantitative evaluation metrics were applied across different time-series datasets, validating the model's universality. On the Image dataset, the clustering results achieved a maximum value of 0.9825, with an average exceeding 0.85.

(2) Comparing the proposed digital economy clustering model with traditional K-means clustering analysis, the clustering results of the economic forecasting model better reflect the economic development tiers of 36 cities, thereby validating the effectiveness of the proposed clustering model.

References

[1] Neuendorf, F., von Haaren, C., & Albert, C. (2018). Assessing and coping with uncertainties in landscape planning: an overview. Landscape Ecology, 33(6), 861-878.

[2] Plakandaras, V., Gupta, R., & Wohar, M. E. (2019). Persistence of economic uncertainty: a comprehensive analysis. Applied Economics, 51(41), 4477-4498.

[3] Akyüz, Y. (2017). Global economic landscape and prospects. Presentation made at the briefing for developing countries on Global Trends and Linkages to Geneva Multilateral Processes). Geneva, 13.

[4] Chen, S., & Ranciere, R. (2019). Financial information and macroeconomic forecasts. International Journal of Forecasting, 35(3), 1160-1174.

[5] Bovi, M., & Cerqueti, R. (2016). Forecasting macroeconomic fundamentals in economic crises. Annals of Operations Research, 247(2), 451-469.

[6] Knotek II, E. S., & Zaman, S. (2019). Financial nowcasts and their usefulness in macroeconomic forecasting. International Journal of Forecasting, 35(4), 1708-1724.

[7] Martinsen, K., Ravazzolo, F., & Wulfsberg, F. (2014). Forecasting macroeconomic variables using disaggregate survey data. International Journal of Forecasting, 30(1), 65-77.

[8] Tanaka, M., Bloom, N., David, J. M., & Koga, M. (2020). Firm performance and macro forecast accuracy. Journal of Monetary Economics, 114, 26-41.

[9] Leon-Gonzalez, R. (2021). Forecasting macroeconomic variables in emerging economies. Journal of Asian Economics, 77, 101403.

[10] Hall, A. S. (2018). Machine learning approaches to macroeconomic forecasting. The Federal Reserve Bank of Kansas City Economic Review, 103(63), 2.

[11] Alqahtani, A., Ali, M., Xie, X., & Jones, M. W. (2021). Deep time-series clustering: A review. Electronics, 10(23), 3001.

[12] Paparrizos, J., & Gravano, L. (2017). Fast and accurate time-series clustering. ACM Transactions on Database Systems (TODS), 42(2), 1-49.

[13] Ma, Q., Zheng, J., Li, S., & Cottrell, G. W. (2019). Learning representations for time series clustering. Advances in neural information processing systems, 32.

[14] Aghabozorgi, S., Shirkhorshidi, A. S., & Wah, T. Y. (2015). Time-series clustering–a decade review. Information systems, 53, 16-38.

[15] Montero, P., & Vilar, J. A. (2015). TSclust: An R package for time series clustering. Journal of Statistical Software, 62, 1-43.

[16] Javed, A., Lee, B. S., & Rizzo, D. M. (2020). A benchmark study on time series clustering. Machine Learning with Applications, 1, 100001.

[17] Imalin, S., Anithakumari, V., & Arul Flower Mary, V. M. (2024). A Forecasting Method Based On K-Means Clustering and First Order Fuzzy Time Series. Journal of Computational Analysis & Applications, 33(6).

[18] Özkoç, E. E. (2020). Clustering of time-series data. Data Mining-Methods, Applications and Systems, 87.

[19] Kembe, M. M., & Onoja, A. A. (2017). Cluster Analysis of macroeconomic indices. Journal of Statistics and Mathematical Sciences, 3(1).

[20] Pang, C. (2022). Construction and analysis of macroeconomic forecasting model based on biclustering algorithm. Journal of Mathematics, 2022(1), 7768949.

[21] Jo, Y., & Lim, Y. (2025). Forecasting GDP time series via the-means based factor model. Communications for Statistical Applications and Methods, 32(2), 173-180.

[22] Zheng, T., Fan, X., Jin, W., & Fang, K. (2024). Words or numbers? Macroeconomic nowcasting with textual and macroeconomic data. International Journal of Forecasting, 40(2), 746-761.

[23] Mori, U., Mendiburu, A., & Lozano, J. A. (2015). Similarity measure selection for clustering time series databases. IEEE Transactions on Knowledge and Data Engineering, 28(1), 181-195.

[24] Serra, A. P., & Zárate, L. E. (2015). Characterization of time series for analyzing of the evolution of time series clusters. Expert systems with applications, 42(1), 596-611.

[25] Gružauskas, V., Čalnerytė, D., Fyleris, T., & Kriščiūnas, A. (2021). Application of multivariate time series cluster analysis to regional socioeconomic indicators of municipalities. Real estate management and valuation., 29(3), 39-51.

[26] Abbasimehr, H., & Noshad, A. (2025). Big time series data forecasting based on deep autoencoding and clustering. Cluster Computing, 28(4), 220.

[27] Khochiani, R., & Hosseini, S. M. (2020). Clustering Based on Forecasting Density: Case Study of Unemployment Rate in Iran's Provinces. Regional Planning, 10(37), 1-16.

[28] Iyetomi, H., Aoyama, H., Fujiwara, Y., Souma, W., Vodenska, I., & Yoshikawa, H. (2020). Relationship between macroeconomic indicators and economic cycles in US. Scientific reports, 10(1), 8420.

[29] Shi, Z. (2024, October). Construction and Application of Macroeconomic Forecasting Model Based on Time Series Clustering Analysis. In 2024 First International Conference on Software, Systems and Information Technology (SSITCON) (pp. 1-5). IEEE.

[30] Dendramis, Y., Kapetanios, G., & Marcellino, M. (2020). A similarity-based approach for macroeconomic forecasting. Journal of the Royal Statistical Society Series A: Statistics in Society, 183(3), 801-827.

[31] La Rocca, M., Giordano, F., & Perna, C. (2023). Time Series Clustering Based on Forecast Distributions: An Empirical Analysis. Statistical Models and Methods for Data Science, 81.

[32] Goldani, M. (2024). Evaluating Feature Selection Methods for Macro-Economic Forecasting, Applied for Iran’s Macro-Economic Variables. Journal of Sciences, Islamic Republic of Iran, 35(3), 243-256.

[33] Hmamouche, Y., Przymus, P., Casali, A., & Lakhal, L. (2017). GFSM: a feature selection method for improving time series forecasting. International Journal On Advances in Systems and Measurements.

[34] Li, B. (2023, April). Research on Economic Forecasting Model Based on Artificial Neural Network. In 2023 International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE) (pp. 1-7). IEEE.

[35] Batóg, J. A. C. E. K., & Batóg, B. (2019). Synchronization of Business Cycles in the EU: Time Series Clustering. WSEAS Trans. Bus. Econ, 16, 298-305.

[36] Víctor Dugo,David Gálvez Ruiz & Pilar Díaz Cuevas. (2025). The sustainable energy development dilemma in European countries: a time-series cluster analysis. Energy, Sustainability and Society,15(1),36-36.

[37] Alberto Manuel Garcia Navarro,Celine Eid,Vera Rocca,Christoforos Benetatos,Claudio De Luca,Giovanni Onorato & Riccardo Lanari. (2025). Integrated Analysis of Satellite and Geological Data to Characterize Ground Deformation in the Area of Bologna (Northern Italy) Using a Cluster Analysis-Based Approach. Remote Sensing,17(15),2645-2645.

[38] Dingsha Jin,Asif Iqbal,Mengshu Huang,Yuqing Liu,Yage Zhang,Youzhen Lin... & Xiaoning Wang. (2025). Comprehensive screening of salt-tolerant rice germplasm using a fuzzy membership and PCA-based evaluation model. Euphytica,221(9),142-142.

[39] Zuoling Zhang,Yijun Yuan,Tingting Hu,Mengling Wang,Weijie Zhang,Yao Nie... & Bingjie Zou. (2025). Proximity cleavage assay (PCA): A single-step, 20-minute platform for protein detection in 1-μL biofluids. Biosensors & bioelectronics,289,117864.