BILINGUAL PUBLISHING GROUP

Journal of Computer Science Research

https://journals.bilpubgroup.com/index.php/jcsr

ARTICLE

Establishing the Forecasting Model with Time Series Data Based on Graph and Particle Swarm Optimization

Le Thi Luong, Tran Thi Thanh, Nghiem Van Tinh*[*], Bui Thi Thi

Faculty of Electronic, Thai Nguyen University of Technology, Thai Nguyen, 250000, Vietnam

ABSTRACT

In recent years, a wide variety of fuzzy time series (FTS) forecasting models have been created and recommended to handle the complicated and ambiguous challenges relating to time series data from real-world sources. However, the accuracy of a model is problem-specific and varies across data sets. But a model's precision varies between different data sets and depends on the situation at hand. Even though many models assert that they are better than statistics and a single machine learning-based model, increasing forecasting accuracy is still a challenging task. In the fuzzy time series models, the size of the intervals and the fuzzy relationship groups are thought to be crucial variables that affect the model's forecasting abilities. This study offers a hybrid FTS forecasting model that makes use of both the graph-based clustering technique (GBC) and particle swarm optimization (PSO) for adjusting interval lengths in the universe of discourse (UoD). The suggested model's forecasting results have been compared to those provided by other current models on a dataset of enrollments at the University of Alabama. For all orders of fuzzy relationships, the suggested model outperforms its counterparts in terms of forecasting accuracy.

Keywords: Forecasting; FTS; Fuzzy relationship group; GBC; Enrolments; COVID-19

1. Introduction

Forecasting the future of any phenomenon assists in making better judgments in uncertain situations. When forecasting particular events in the past, some researchers employed popular linear forecasting techniques, such as regression analysis, exponential moving averages, and autoregressive integrated moving averages (ARIMA). Nevertheless, conventional

*CORRESPONDING AUTHOR:
Nghiem Van Tinh, Faculty of Electronic, Thai Nguyen University of Technology, Thai Nguyen, 250000, Vietnam; Email: nghiemvantinh@tnut.edu.vn

ARTICLE INFO
Received: 2 February 2023 | Revised: 24 March 2023 | Accepted: 27 March 2023 | Published Online: 13 April 2023
DOI: https://doi.org/10.30564/jcsr.v5i2.5443

CITATION
Luong, L.T., Thanh, T.T., Tinh, N.V., et al., 2023. Establishing the Forecasting Model with Time Series Data Based on Graph and Particle Swarm Optimization. Journal of Computer Science Research. 5(2): 1-15. DOI: https://doi.org/10.30564/jcsr.v5i2.5443

time series forecasting models have the disadvantages of being unable to cope with forecasting problems in linguistic values historical data and insufficient data. To address the drawbacks of traditional time series forecasting approaches, Song and Chissom [1,2] proposed two FTS models based on fuzzy logic [3] and fuzzy mathematics approaches to forecasting the number of enrollments at the University of Alabama. However, in the defuzzification process, their methods included sophisticated max-min operations and the construction matrix of fuzzy relationships $R(t-1, t)$. As a consequence, Chen [4] has developed a first-order FTS model by using basic arithmetic computations in the defuzzification process, which takes less time than performing the max-min complex composition operation [1,2].

The calculation of interval lengths and the establishment of fuzzy relationships (FRs) are two crucial factors in the FTS forecasting model that can considerably impact predicting effectiveness. To achieve more accurate forecasting results, several researchers have proposed various improvements from Song's model [2] and Chen's model [4] in aspects of mining the lengths of intervals [5-22], building fuzzy relationships [22-31], and fuzzy relationship groups [10-12].

For selecting interval lengths, Huarng [5] found out that variable interval lengths in the UoD highly influence the forecasting performance of the model. As a result, he suggested two novel ways based on the average and distribution for determining interval lengths in the fuzzy time series model. Yolcu et al. [15] proposed a strategy based on ratio optimization for interval length. His model is used to anticipate enrollments and inventory demand at the University of Alabama. Furthermore, using an optimization approach with a single-variable constraint, Egrioglu et al. [21] introduced a new method for determining the proper interval length in high-order FTS.

Recently, soft computing techniques and evolutionary algorithms are considered crucial tools for calculating interval lengths in the fuzzy time series sector. For instance, Chen and Chung [10] suggested the high-order model utilizing a genetic algorithm (GA) to divide intervals in the UoD for forecasting the University of Alabama enrollments. Additionally, Lee et al. [30] used a genetic algorithm to establish the various interval lengths in the high-order forecasting model using temperature and TAIFEX data. Besides that, they also used the simulated annealing algorithm to determine optimum interval lengths. Eren Bas et al. [32] introduced a novel modified genetic algorithm to eliminate subjectivity in determining the duration of each interval in the FTS model for forecasting "dead in vehicle road accidents" in Belgium and enrollments at the University of Alabama. Furthermore, to determine interval lengths, several optimization techniques, including the Artificial Bee Colony Algorithm, Particle Swarm Optimization, memetic algorithms, and ant-colony optimization, have been utilized in the universe of discourse. The comparison findings reveal that the PSO algorithm outperforms other algorithms in terms of success rate and solution quality. Additionally, certain comparison studies have indicated that the PSO-based outcomes surpass the GA-based technique [33-36]. PSO applications in optimizing interval lengths in the FTS forecasting models have been demonstrated in research works [12,18,20,22-27,34-36]. Kuo et al. [12] introduced a unique forecasting model for improving predicting error by combining the PSO with FTS. By suggesting a new defuzzification rule, Kuo et al. [23] presented a novel FTS model for forecasting TAIFEX. Two research works [24,34] established two-factor high-order FTS models for predicting distinct situations from the same perspective of employing PSO in altering interval lengths. Furthermore, Chen and Bui [26] used the PSO not only to achieve optimal interval partitioning but also to generate optimal weight vectors for forecasting the TAIFEX and NTD/USD exchange rates. Moreover, clustering techniques also have been used in the FTS model to determine the optimal interval such as K-means [37], fuzzy C-means [38], and automatic clustering [11,39]. Cheng et al. [40] have presented a fuzzy forecasting model with two advantages: the PSO to achieve optimal intervals and the K-means method to split the subscripts of the fuzzy sets at the current states of FRs. Dincer [17] used the fuzzy k-medoid clustering technique to handle outliers and abnormal data time series for forecasting air pollution.

Another approach based on fuzzy time series, some of the authors in works [41,42] introduced intuitionistic fuzzy time series (IFTS) and established an IFTS forecasting model to forecast the University of Alabama enrolment and State Bank of India (SBI) market share price on the Bombay stock exchange (BSE). The authors in research work [43,44] introduced hesitant probabilistic fuzzy sets in time series forecasting to address the issues of non-stochastic non-determinism and apply for forecasting the enrolments of the University of Alabama and the share market prizes of the State Bank of India.

Based on the study shown above, it is clear that selecting the optimal lengths of intervals and building fuzzy relationships are difficult and have a major impact on the model's predicting ability. Despite substantial advances in leveraging the duration of each interval and establishing predicting output rules, these difficulties continue to occupy the attention of academics. There are still a lot of approaches in the universe of discourse to identify the duration of intervals and generate forecasting output values from fuzzified values. To address these concerns, we present a novel hybrid FTS forecasting model that combines the Graph-based clustering approach with PSO in the selection of optimum lengths for dealing with various problems. The suggested forecasting model's performance is assessed using several real-world data sets, including enrollment data from the University of Alabama and confirmed COVID-19 instances from Vietnam. The findings obtained are compared to those of other approaches. The suggested strategy offers more accurate forecasts for the future values of the given enrollment time series.

The remainder of the paper is organized as follows: Basic concepts of fuzzy time series and algorithms are provided in Section 2. Section 3 proposes an FTS forecasting model that combines the GBC technique with PSO. Section 4 explains the application of the forecasting model and experimental findings. In Section 5, conclusions are offered.

2. Preliminaries

In this section, we describe some basic concepts of fuzzy time series [1,2] and related algorithms to serve as a foundation for our work.

2.1 Basic definitions of FTS

Song and Chissom [1,2] introduced the FTS concept originally, then Chen [4] improved it with a straightforward defuzzification method that was more precise. Some basic definitions of FTS are as below:

Definition 1: Fuzzy time series [1,2]

Let $Y(t)$ (t = ..., 0, 1, 2...), a subset of real number R, be the universe of discourse by which $f_i(t)(i = 1, 2...)$ are defined and if $F(t)$ is a collection of $f_1(t)$, $f_2(t)$,..., then $F(t)$ is called a FTS definition on $Y(t)$ (t ..., 0, 1,2 ...).

Definition 2: Fuzzy relationships(FRs) [1,2]

Assume that $F(t) = A_k$ and $F(t-1) = A_i$, the first order relationship between $F(t-1)$ and $F(t)$ is denoted as a FLR: $A_i \rightarrow A_k$, where, A_k is the current state that dependent on previous state A_i.

Definition 3: m - order fuzzy relationships [1,2]

If $F(t)$ is dependent on previous state $F(t-1)$, $F(t-2)$,..., $F(t-m+1)$ $F(t-m)$, then $F(t-m)$,..., $F(t-2)$, $F(t-1) \rightarrow F(t)$ is called an m - order FLR. Time series with respect to this FLR is called an m - order FTS.

Definition 4: Fuzzy relationship groups (FRGs) [4]

The fuzzy relationships that Song and Chisom [1,2] presented in Definition 2, if they have the same left-hand side can be further grouped into an FRG. Assume the following FRs exist: $A_i \rightarrow A_k$, $A_i \rightarrow A_l$,.., $A_i \rightarrow A_m$; these FRs can be placed into the same FRG as : $A_i \rightarrow A_k, A_l \ldots, A_m$.

Definition 5: Time-variant fuzzy relationship groups [22]

The fuzzy relationship is presented by the relationship $F(t-1) \rightarrow F(t)$. If, let $F(t) = A_i(t)$ and $F(t-1) = A_j(t-1)$, the FR between $F(t-1)$ and $F(t)$ can be denoted as $A_j(t-1) \rightarrow A_i(t)$. Also at the time t, we only consider fuzzy relationships $A_j(t1-1) \rightarrow A_{i1}(t1)$; ...; $A_j(tn-1) \rightarrow A_{in}(tn)$ that occur before the fuzzy relationship $A_j(t-1) \rightarrow A_i(t)$ to make an FRG as $A_j(t-1) \rightarrow A_{i1}(t1), A_{i2}(t2), A_{in}(tn), A_i(t)$. It is referred to as first-order time-variant FRGs.

2.2 Graph-based clustering algorithm

Algorithms for graph-based clustering are effective at generating conclusions that are close to human intuition [45]. Building a graph on the set of data and using it as the basis for clustering is a frequent feature of graph-based clustering techniques developed in recent years [46]. In the GBC approach, each node in a graph represents a data object, and each object has connections with other objects. A set of things is said to form a cluster in this case if they are connected to one another but not to any other objects. Based on these ideas, our research [47] proposes a data clustering technique that uses a tree-like display of the information and constructs clusters automatically rather than the number of clusters user-predetermined. In particular, the graph-based clustering method is summarized into four procedures as follows:

(1) Root node location procedure (RNLP). This technique identifies the root node based on the provided data.

(2) Node insertion procedure (NIP). This technique inserts one element of the dataset and root node and places the elements in the proper position.

(3) Tree-making procedure (TMP). This procedure displays the tree from the provided data set and the root node.

(4) Clustering procedure (CP) based on nodes in the tree. This process makes logical node clustering using the tree that the TMP generated as input.

2.3 Particle swarm optimization algorithm

PSO was initially put up by Eberhart and Kennedy [48] for the purpose of locating the global optimal solution. A group of particles known as a swarm in PSO. Where each particle represents a potential solution and constantly navigates around the search space (d-dimensional space) in pursuit of the ideal answer. All particles (i.e., $Pmax$ particles) have fitness values to assess their performance during particle movement, and each particle recalls the best location from its own flight experience, which is called P_{best}. The best particle in the population as a whole is then referred to as G_{best}. Following Algorithm 1 is a brief summary of the PSO algorithm's steps:

Algorithm 1: The standard PSO algorithm

Step 1. Initialize random positions x_{ki}; random velocities v_{ki} in d dimensional space (i = 1,2,…,d);

- Positions of each k^{th} (k = 1,2, …, $Pmax$) particle's positions are randomly determined and kept in a X_{kd} given as follows:

$$X_{kd} = [x_{k,1}, x_{k,2}, \dots, x_{k,d}] \tag{1}$$

where x_{ki} denotes ith position of kth particle, N is the number of particles in a swarm.

- Velocities are generated at random and saved in the vector V_{kd} in Equation (2).

$$V_{kd} = [v_{k,1}, v_{k,2}, \dots, v_{k,d}] \tag{2}$$

Step 2: Based on the fitness function, P_{best_kd} and G_{best} particles given in Equations (1) and (2), respectively, are determined.

$$P_{best\_kd} = [\,p_{k,1}, p_{k,2}, \dots, p_{k,d}] \text{ and}$$
$$G_{best} = \text{minimum}\ (\,P_{best\_kd}).$$

Step 3: Similar to the ones shown [12], C_1 and C_2 are two learning factors, and ω is the time-varying inertia weight. In each iteration t, the parameter ω is calculated by using Equation (3) as follows:

$$\omega^t = \omega_{max} - \frac{t * (\,\omega_{max} - \omega_{min})}{iter\_max} \tag{3}$$

where, $iter\_max$ denotes the maximum iteration number.

Step 4: Values of velocities and positions are updated by using Equations (4) and (5), respectively.

$$V_{kd}^{t+1} = \omega^t * V_{kd}^t + c_1 * R1(\,) * \left(P_{best_{kd}} - X_{kd}^t\right) + c_2 * R2(\,) * (\,G_{best} - X_{kd}^t) \tag{4}$$

$$X_{kd}^{t+1} = X_{kd}^t + V_{kd}^{t+1} \tag{5}$$

Here, $R1(\,)$ and $R2(\,)$ are randomly generated values between [0, 1].

Step 5: Steps 2 to 4 are repeated until a predefined maximum number of iterations is reached.

3. An FTS forecasting model using graph-based clustering and PSO

This section's goal is to introduce a hybrid FTS model that combines PSO and GBC. This proposed model is named GBCFTS-PSO. The framework of the GBCFTS-PSO model is shown in **Figure 1**. It consists of three stages: (1) partitioning of intervals based on the GBC; (2) establishing the FTS model; and (3) choosing the best interval lengths by using the PSO.

3.1 Establishing the forecasting model using Graph-based clustering and fuzzy time series

The proposed FTS model's details are discussed in steps utilizing datasets of the number of new confirmed COVID-19 in Vietnam from June 15, 2021, to July 15, 2021, as shown in **Figure 2**. The followings are the steps of the suggested model:

Step 1: Using the GBC approach, divide data into C intervals.

In this step, four procedures from the Graph-based clustering algorithm in Section 2.2 are utilized to partition time series $X(t)$ into C clusters. The following are the short outcomes of these four procedures:

1) Root node location procedure (RNLP)

The dataset of newly confirmed COVID-19 instances is input $X(t)$ = (398, 414, 259, 471, ..., 2296, 2924, 1922); with (15/07/2021 $\leq t \leq$ 15/7/2021).

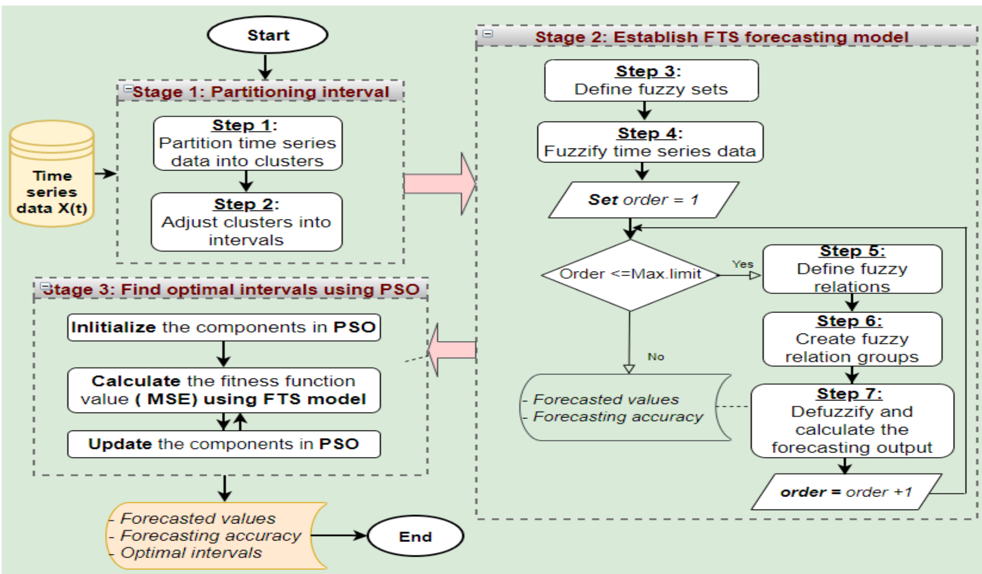

Figure 1. Framework of the proposed forecasting model based on GBC and PSO.

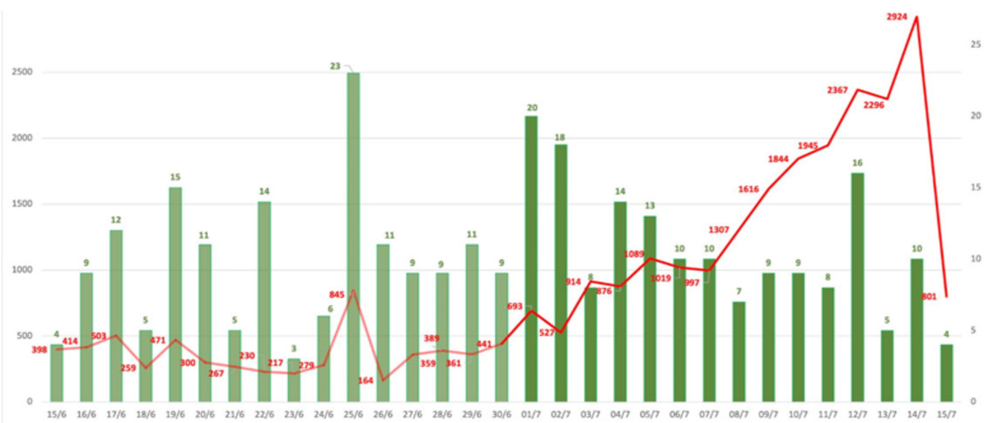

Figure 2. The dataset of number of confirmed cases of COVID-19 in Vietnam.

Source: https://ncov.moh.gov.vn

Calculate range $Rg = Xmax_{value} - Xmin_{value} = 2760$.

Calculate standard deviation of the time series as SD = 762.76.

$$w = \frac{R_g}{SD * N} = 0.12$$

Define universe of discourse (U) of the $X(t)$:

$U = [Xmin_{value} - w, Xmax_{value} + w] = [163.88, 2924.12]$

Calculate midpoint of U as: $Mid_u = (MIN_{value} + MAX_{value}) / 2 = 1544$.

Assign the Mid_u as root node: $Root = Mid_u = 1544$.

2) Tree make Procedure (TMP) and node insertion procedure (NIP)

For generating the tree, from the input dataset and Root node. We apply two procedures TMP and NIP to create tree and insert nodes into the tree. The results of these two procedures are shown in **Figure 3**.

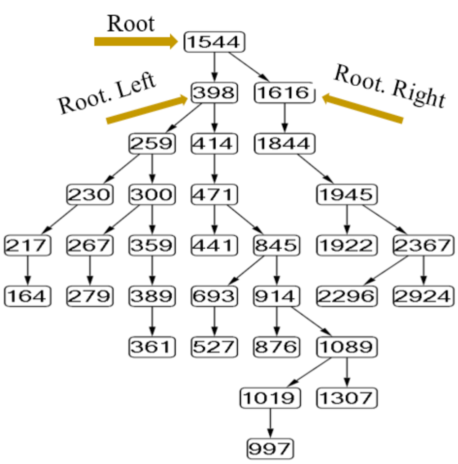

Figure 3. The tree depicts the input data of COVID-19 series based on two procedures TMP and NIP with root node of 1544.

3) Make the clusters from the tree in **Figure 3** based on Procedure 4 (CP)

Following the acquisition of the data tree depicted in **Figure 3**, the process of creating clusters is briefly described in accordance with the following conditions:

(1) Initially, check to see whether Root exists or not and check to see whether Root has left (Root. LEFT) and right (Root. RIGHT)

(2) Calculate the difference between the values of Root and Root. RIGHT, Root and Root. LEFT if both children exist for each Root. After that, create a cluster consisting of the Root and associated child that has the smallest difference.

(3) If there is just one child for each Root, create the cluster using either the Root and Root. LEFT or Root and Root. RIGHT.

(4) Continue repeating conditions 1-3 until all of the value from the tree's nodes has been added to the clusters.

From the above procedures, we obtain 15 clusters and their related cluster centers. Then, these clusters are then arranged in ascending order by the clustering centers, and the final results are listed in **Table 1**.

Step 2: Adjust these clusters into intervals

After applying the adjustment principles from clusters to intervals in the research work [49], we obtain 15 intervals u_i corresponding to the clusters in **Table 2**, and the midpoint values of these intervals are shown in **Table 3**.

Step 3: Identify the corresponding language term for each interval in Step 2

Each linguistic term can be defined by intervals that the historical time series data distributed among these intervals. For ten intervals in Step 2, we obtain 15 linguistic values of linguistic variables "*enrolments*" e.g., {{very very very few}, {very very few},

Table 1. Clusters and cluster centers.

No	Clusters C_i	Cluster center (V_i)	No	Clusters C_i	Cluster center (V_i)
1	(217,164)	190.5	9	(997)	997
2	(259,230)	244.5	10	(1089,1019)	1054
3	(300,267,279)	282	11	(1307)	1307
4	(359,389,361)	369.7	12	(1544,1616)	1580
5	(398,414)	406	13	(1844,1945,1922)	1903.7
6	(471,441)	456	14	(2367,2296)	2331.5
7	(693,527)	610	15	(2924)	2924
8	(845,914,876)	878.3			

Table 2. The intervals and midpoint values of them.

No	Intervals (u_i)	Midpoint	No	Intervals (u_i)	Midpoint
1	[164, 217.5]	190.8	9	[937.65, 1025.5]	981.6
2	[217.5, 263.25]	240.4	10	[1025.5, 1180.5]	1103
3	[263.25, 325.85]	294.6	11	[1180.5, 1443.5]	1312
4	[325.85, 387.85]	356.8	12	[1443.5, 1741.85]	1592.7
5	[387.85, 431]	409.4	13	[1741.85, 2117.6]	1929.7
6	[431, 533]	482	14	[2117.6, 2627.75]	2372.7
7	[533, 744.15]	638.6	15	[2627.75, 2924]	2775.9
8	[744.15, 937.65]	840.9			

{very few},{few}, …, {very many}, {too many}, {too many many}, {too many many many}}, which can be described with fuzzy sets A_i, e.g. {A_1, A_2, A_3,..., A_9, A_{15},}, respectively and calculated as follows:

$$A_i = {}^{a_{i1}}/_{u_1} + {}^{a_{i2}}/_{u_2} + \ldots + {}^{a_{ij}}/_{u_j} + \ldots + {}^{a_{i10}}/_{u_{15}} \quad (6)$$

where, the values $a_{ij} \in [0,1]$ denote the grade of membership of u_j in fuzzy set A_i and which is calculated in Equation (7). Here, the symbol '+' denotes the set union operator and the symbol '/' denotes the membership of u_j which belongs to A_i.

$$a_{ij} = \begin{cases} 1 & j = i \\ 0.5 & j = i-1 \; or \; j = i+1 \\ 0 & others \end{cases} \quad (7)$$

Step 4: Fuzzy all historical time series data

In order to fuzzy all historical time series, the typical method is to convert historical data which belongs to the interval U into fuzzy sets. If the maximum membership value of fuzzy set A_i occurs at u_i, then the fuzzified historical value is considered as A_i. For example, the COVID-19 data on day 15/6/2021 equal to 398 belongs to the interval u_5 = [387.85, 431] and the highest membership value of fuzzy set A_5 occurs at u_5. So, it is fuzzified into A_5. The similar way for next years, we complete the results of fuzzification of enrolments data for all years, as listed in **Table 3**.

Step 5: Establish all mth- order FRs between the fuzzified data values. ($m \geq 1$).

To build fuzzy relationships, we need to find any relationship having the type $F(t-m)$, $F(t-m+1)$,..., $F(t-1) \rightarrow F(t)$, where, the left-hand side of FLR ($F(t-m)$, $F(t-m+1)$,..., $F(t-1)$) is called the current state and the right-hand side of FLR. ($F(t)$) is called the next state, respectively. Then, the mth -order FLR is replaced by relationships in accordance with the corresponding fuzzy sets as: A_{im},, $A_{i(m-1)}$,…, A_{i2}, $A_{i1} \rightarrow A_k$.

In this way, we have achieved the first-order FRs for tall fuzzified data values, which are presented in column 4 of **Table 4**.

Here the linguistic value of F(16/7/2021) on the right-hand side of the last relationship is denoted by the symbol '#', which is used to represent the unknown linguistic value.

Step 6: Create all *m*—order time—variant FRGs

In this study, we develop fuzzy relation groups using the concept of a time-variant fuzzy relationship group [22], which was mentioned in Definition 5. Based on the present state of the FRs in **Table 4**, the FRs may be put into a group and entitled time vari-

Table 3. The results of fuzzification for COVID-19 time series.

Day	Actual data	Fuzzy sets	Maximum membership value	Linguistic value
15/6/2021	398	A_5	[**1** 0.5 0 0 0 0 0 0 0 0]	*"somewhat few"*
16/6/2021	414	A_5	[**1** 0.5 0 0 0 0 0 0 0 0]	*"somewhat few"*
-----	-----	----	------------------	----------------
14/7/2021	2624	A_{15}	[0 0 0 0 0 0 0 0 0 0.5 **1**]	*"too many many many"*
15/7/2021	1922	A_{13}	[0 0 0 0 0 0 0 0 0.5 **1** 0.5]	*"too many "*

ant-FRGs by examining the history of the emergence of the fuzzy sets on the future state of the FRs. From this approach, we obtain all first-order time-variant FRGs, as shown in **Table 5**. Where 30 groups (G1-G30) are in the training stage and one group (G31) is in the testing stage.

Table 4. The results of the first-order fuzzy relationships.

Day	No	Fuzzy set	First-order FRs
15/6/2021		A_5	
16/6/2021	1	A_5	$A_5 \rightarrow A_5$
	--	----	----------
14/7/2021	29	A_{10}	$A_{14} \rightarrow A_{15}$
15/7/2021	30	A_9	$A_{15} \rightarrow A_{13}$
16/7/2021	31	N/A	$A_{13} \rightarrow$ #

Table 5. The results of the first-order time-variant FRGs.

No	First-order Time-variant FRGs
G1	$A_5 \rightarrow A_5$
G2	$A_5 \rightarrow A_6$
----	--------------
G29	$A_{14} \rightarrow A_{14}, A_{15}$
G30	$A_{15} \rightarrow A_{13}$
G31	$A_{13} \rightarrow$ #

Step 7: Defuzzify the forecasting output values

Our defuzzified principles in the paper [49] are introduced to compute the forecasting value for all first-order and high-order time-variant FRGs in the training stage. In the testing, we apply a defuzzified principle [24] to compute with the unknown linguistic value. The forecasting rules are briefly represented as follows:

Principle 1: Calculate the forecasting value for known linguistic values

To obtain the forecasting results, we divide each corresponding interval concerning the linguistic value in the next state into four sub-intervals that have the same length and calculate the forecasted output value for each group based on Equation (8).

$$\text{Forecasted\_value} = \frac{1}{2*n}\sum_{i=1}^{n}(subm_{ik} + Value\_lu_{ik}) \quad (8)$$

where n is the total of fuzzy sets on FRG's next state.

- $\boldsymbol{subm_{ik}}$ is the midpoint value of one of four sub-intervals about the i-th linguistic value in the next state of FRG in which the real data at forecasting time falls.

- $\boldsymbol{Value\_lu_{ik}}$ is one of two values belonging to the lower and upper bounds of one of four sub-intervals with real data falling inside sub-interval u_{ik} at forecasting time.

Principle 2: Calculate the forecasting value with unknown linguistic values

In the testing phase, we determine the forecasting value for each FRG which has the unknown linguistic value occurring in the next state. Suppose there is the ^{m}th-order fuzzy relationship group whose next state is #, shown as follows: $A_{im}, A_{im-1}, \ldots, A_{i1} \rightarrow \#$.

Here the symbol '#' represents an unknown value, then the forecasted value of year i is calculated in accordance with [24] as follows:

$$\text{Forecasted value} = m_{i1} + \frac{\sum_{k=2}^{m} m_{i(k-1)} - m_{ik}}{2^{k-1}} \quad (9)$$

where, $m_{i1}, m_{i2}, \ldots, m_{ik}$ are midpoints of $u_{i1}, u_{i2}, \ldots$, and u_{ik} $(2 \leq k \leq m)$, respectively.

Based on the two forecasting principles above, we complete forecasting results for COVID-19 confirmed cases prediction from 15/6/2021 to 15/7/2021 with first-order TV-FRGs under fifteen intervals, which are given in **Table 6**.

Table 6. The forecasting results for COVID-19 based on 1st-order fuzzy relationships.

Day	COVID-19 data	Fuzzy sets	1st-order forecasted value
15/06/2021	398	A_5	Not forecasted
16/06/2021	414	A_5	414.6
----	------	----	-------
14/07/2021	2924	A_{15}	2614.7
15/07/2021	1922	A_{13}	1898.4

To confirm the effectiveness of the suggested forecasting model, the mean square error (*MSE*) is employed as assessment criteria in terms of the forecasted accuracy. The formulas of both indices are calculated as follows:

$$MSE = \frac{1}{n}\sum_{i=m}^{n}(F_i - R_i)^2 \quad (10)$$

where R_i, F_i represent the real value and forecasting value at year i, respectively; n is the total number of forecasted data, *m* means the order of the FLR.

3.2 Using PSO to determine appropriate interval lengths

In this section, we present a hybrid FTS forecasting model termed GBCFTS-PSO, which combines graph-based clustering and PSO to produce optimal interval lengths with the goal of improving forecasted accuracy. In GBCFTS-PSO model, each particle denotes the partitioning of historical data into intervals. The number of intervals is obtained from the GBC algorithm (e.g., d intervals). The lower bound and upper bound of the UoD be p_0 and p_d, respectively. Each particle symbolizes a vector with d-1 elements as p_1, p_2, …, p_{d-2} and p_{d-1}, where $p_i \leq p_{i+1} (1 \leq i \leq d-1)$. Based on these d-1 elements, obtain the intervals set as $u_1 = [p_0, p_1]$, $u_2 = [p_1, p_2]$,…, $u_i = [p_{i-1}, p_i]$,… and $u_d = [p_{d-1}, p_d]$, respectively. When a particle transfers from one place to another according to Equations (4) and (5), d-1 elements must be sorted in ascending order such that $p_1 \leq p_2 \leq \dots \leq p_{d-1}$. The function MSE (10) is employed to evaluate the forecasting accuracy of each particle. Algorithm 2 shows the entire process of the suggested model.

4. Experiment management and results evaluation

In this research, the GBCFTS-PSO model is applied for forecasting two time series datasets: The University of Alabama enrollment data [4] and the number of new confirmed COVID-19 in Vietnam from June 15, 2021, to July 15, 2021. Before implementing the suggested forecasting model, the time series datasets are briefly presented. The simulation and analysis findings for these data sets are then presented. The statistical properties of these time series data are provided below.

Algorithm 2: The GBCFTS-PSO algorithm

1**. Input**: Historical time series data
2**. Output**: The forecasting output and MSE value
Begin
3. **Choose** the initial intervals by using the GBC technique and using forecasted steps in Subsection 3.1 to get the forecasting accuracy (MSE)..
4. ***Initialize***: number of particles: *Pmax,* number of maximum iteration: *iter_max*
- ✓ The initial positions of all particles be limited by: $p_0 + Rand(\,) * (p_d - p_0)$; where, p_0 *and* p_d are the lower bound and upper bound of the universe of discourse UoD which is created by GBC; the intervals created by particle 1 are identical to the one created by GBC.
- ✓ The velocity V_{kd} of all particles be exceeded by: $V_{min} + Rand(\,) * (V_{max} = - V_{min})$; $V_{min} = - V_{max}$
- ✓ In order to find Gbest, all particles' beginning positions are set to their personal best positions.

5. **while** (*iter* ≤ *iter_max*) **do**
5.1**. foreach** particle *kd*, (1 ≤ *kd* ≤ *Pmax*) **do**
- ✓ Based on Step 3 in Subsection 3.1, determine all intervals defined by the present position of the particle *kd.*
- ✓ Fuzzify all historical data described in Step 4 of Subsection 3.1.
- ✓ Establish all *m* – order FRs using Step 5 in Subsection 3.1
- ✓ Build all *m* – order time-variant FRGs using Step 6 in Subsection 3.1
- ✓ Forecast and defuzzify output values using Step 7 in Subsection 3.1
- ✓ Calculate the MSE values for particle *kd* based on Equation (10)
- ✓ The new *Pbest* of particle *kd* is saved according to the MSE values.

end for
5.2. The new *Gbest* of all particles is saved according to the MSE values
6. **foreach** particle kd, (1 ≤ *kd* ≤ *Pmax*) **do**
✓ The particle kd is moved to another location according to (4) and (5)
end for
✓ change *ω* according to **(3)**
end while
End.

4.1 Data description and relevant parameters of each time series

1) *The enrolments data of university of Alabama* [4]: Enrolments time series is applied to quite a wide range of areas in literature from the beginning of research [1,2]. This dataset contained 22 observations during the period from 1971 to 1992. One of the outcomes of these studies is compared to the proposed model. The universe of discourse of the enrollment data series is determined by the clustering algorithm as $U = [Xmin_{value} - w, Xmax_{value} + w] = [13054.84, 19337.16]$, where w = 0.16, $Xmin_{value} = 13055$ and $Xmax_{value} = 19337$ are the data series' minimum and maximum values.

2) COVID-19 data sets in Vietnam: This data set contains 31 observed values for new confirmed cases of COVID-19 in Vietnam from June 15, 2021, to July 15, 2021, as shown in **Figure 2**. The minimum and maximum values of the time series are $Xmin_{value} =$ 1164 and $Xmax_{value} = 2924$. The universe of discourse of COVID-19 as U = [163.88, 2924.12] with w = 0.12.

For each dataset, the main parameters of the proposed model are intuitively determined, as shown in **Table 7**.

The suggested forecasting model has been executed 20 times with 10 intervals for enrolment data and 15 intervals for COVID-19 data. The best result from all runs is the final output.

4.2 Experimental results

Forecasting the enrolments of University of Alabama

The suggested model (GBCFTS-PSO) is used in this subsection to forecast enrolments with annual observations [4]. The results of five forecasting models in works [41-44,50] are chosen for comparison to demonstrate the performance of the suggested forecasting model based on first-order FTS under varied intervals. **Table 8** and **Figure 4** show the forecasted values and the forecasting accuracy between our suggested model and comparing models.

From the forecasting results in **Table 8** and **Figure 4**, it can be seen that the GBCFTS-PSO model gets the smallest MSE value of 39302.66 among all the compared models. The difference between the GBCFTS-PSO model and the models mentioned above is the way in which the fuzzy relationship group and

Table 7. The parameters of the GBCFTS-PSO model for forecasting enrolments and COVID-19.

Description for the parameters	Values of enrolments	Values of COVID-19
Number of particles	30	30
The max iteration number	150	150
The inertial weigh ω	0.9 to 0.4	0.9 to 0.4
The coefficient $C_1 = C_2$	2	2
The velocity in search range	[–100,100]	[–100,100]
The position in search range	[13054.84, 19337.16]	[163.88, 2924.12]

Table 8. A comparison of the forecasting results of the suggested model with its counterparts based on first-order fuzzy time series under 10 intervals.

Year	Actual	[41]	[43]	[44]	[50]	[42]	GBCFTS-PSO
1971	13055	-	-	-	-	-	
1972	13563	13693	13595.67	13680.75	13637	13682	13562.1
---	---	---	---	---	---	---	---
1991	19337	19500	19168.56	18972.15	18230	19311	19272.1
1992	18876	19500	19168.56	18972.15	18236	19311	19142.3
MSE value		243601	183723	186313	1800964	178665	39302.66

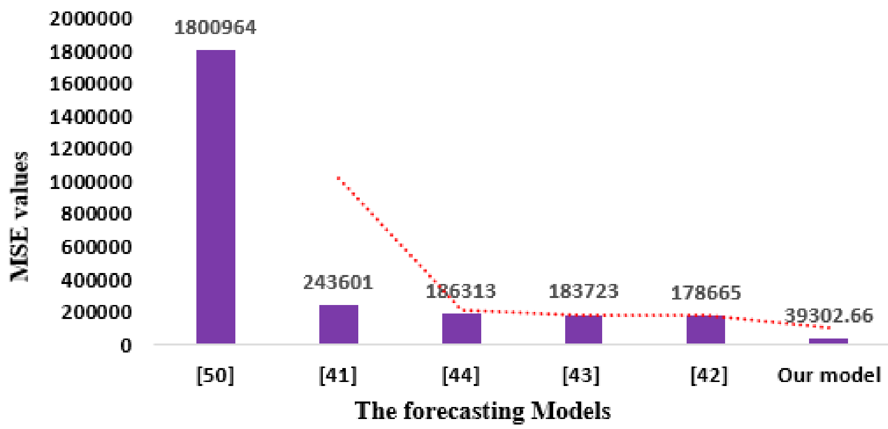

Figure 4. Comparing forecasting accuracy between our model and other models.

method of partitioning are applied to establish the forecasting model. This difference shows that the suggested forecasting model outperforms the previous model when used to evaluate enrollment data for the University of Alabam.

Additionally, in this study, we implement a forecasting model based on the high-order fuzzy relationship from orders 2 to 8 with a fixed number of intervals of 10. To verify the effectiveness of the forecasting model based on high-order fuzzy time series, four models listed in **Table 9** are selected for comparison. The comparison results in terms of MSE values listed in **Table 9** show that the GBCFTS-PSO model gives a lower MSE value than the models selected for comparison in all orders with 10 intervals. Among all fuzzy relationships done in the model, the GBCFTS-PSO model obtains the lowest MSE value of 71.5 with 4th-order fuzzy relationships.

Forecasting new confirmed cases of COVID-19

The GBCFTS-PSO model is also used to forecast the number of new COVID-19 cases that would be confirmed in Vietnam between June 15 and July 15, 2021. The GBCFTS-PSO model executes 20 separate runs for each order based on the parameters listed in **Table 7**, and the best outcome of these runs at each order of FRs is taken as the final forecasting result. The MSE value is used to measure the forecasting accuracy of the suggested model, which is based on the first-order FTS with various intervals. The obtained forecasting results from the proposed model are based on the first-order FTS with 15 intervals, which is shown in **Table 10**. For easy visualizing, the curves of actual value and forecasted value for the number of new confirmed cases each day are shown in **Figure 5**. From this figure, it can be seen that the forecasted value of the suggested model based on the first-order FLR is relatively close to the actual data.

In addition, to verify the superiority of the GBCFTS-PSO model based on the high-order FTS with the same number of intervals of 7. **Table 11** presents the forecasting results of the GBCFTS-PSO model in terms of the MSE value based on the

Table 9. A comparison of the results obtained between the GBCFTS-PSO model and its counterparts based on the various high-order FTS and different intervals.

Models	Number of orders						
	2	**3**	**4**	**5**	**6**	**7**	**8**
[10]	67834	31123	32009	24984	26980	26969	22387
[12]	67123	31644	23271	23534	23671	20651	17106
[25]	19594	31189	20155	20366	22276	18482	14778
[50]	8552	600.3	447.7	387.1	495.6	370.6	319.9
GBCFTS-PSO	2181	183.16	71.5	97.2	128.08	222.15	143

Table 10. The forecasting results of the first-order GBCFTS-PSO model with 15 intervals.

Date/month	Actual data	Forecasted values	Date/month	Actual data	Forecasted values
15/06/2021	398	----	01/7/32021	693	693.5
16/6/2021	414	404.7	2/7/2021	527	550.4
17/06/2021	503	455.2	3/7/2021	914	746.8
18/06/2021	259	262.1	4/7/2021	876	884.8
19/06/2021	471	464.1	5/7/2021	1089	1001.5
20/06/2021	300	286.5	6/7/2021	1019	1001.5
21/06/2021	267	390.8	7/7/2021	997	1001.5
22/06/2021	230	345.3	8/7/2021	1307	1067.6
23/06/2021	217	299.7	9/7/2021	1616	1622.9
24/06/2021	279	294.3	10/7/2021	1844	1835.5
25/06/2021	845	844.9	11/7/2021	1945	1944.7
26/06/2021	164	172.9	12/7/2021	2367	2369.2
27/06/2021	359	307	13/07/2021	2296	2296.5
28/06/2021	389	388.4	14/07/2021	2924	2918.4
29/06/2021	361	382.3	15/07/2021	1922	1924.3
30/06/2021	441	392.1			
MSE		4599.03			

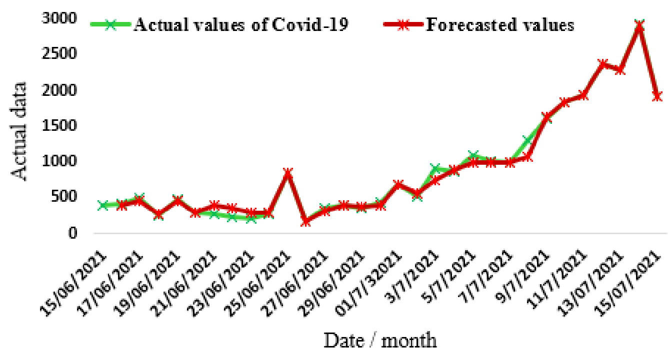

Figure 5. The comparison curves between the actual values and the forecasted values based on the first-order FRs.

Table 11. The forecasting accuracy of GBCFTS-PSO based on the high-order FTS with number of intervals equal to 15.

Orders of model	2nd-order	3rd-order	4th-order	5th-order	6th-order	7th-order	8th-order
MSE	930.05	628	218.06	268.3	356.5	423	415.08

various high-order FRs from 2nd-order to 8th-order. Among all orders of the suggested model with seven intervals, the best forecasting result has been found for the 4th-order FR which has the smallest MSE value equal to 218.06.

5. Conclusions

In this study, a hybrid FTS forecasting model for forecasting various issues is built by combining graph-based clustering and particle swarm optimization. The suggested model has solved two concerns that are seen to be significant and have a significant impact on forecasting accuracy, namely the issues with selecting interval length and how to create fuzzy relationship groups. The disadvantages of FTS models that use conventional fuzzy relationship groups are overcome by the recommended model, which uses the concept of a time-variant fuzzy relation group to produce the forecasting output results. By combining

GBC and PSO techniques in finding the optimal lengths of intervals from the universe of discourse, the forecasting efficiency of the proposed model can be significantly improved. Compared with conventional fuzzy time series models in this field, we employ a hybrid high-order fuzzy time series in order to forecast enrollment at the University of Alabama and the number of confirmed cases of COVID-19 in Vietnam. In many cases, the suggested model significantly outperforms other models on the dataset of enrolments and the simulated results on the dataset of confirmed cases of COVID-19 instances demonstrate that the suggested model gives remarkably better forecasting when using the high-order FTS. Details of the comparison are shown in **Tables 8 through 11**. Although compared to other forecasting models based on high-order FRs, our approach has higher forecasting capability. FLR determination in the high-order FTS model is more difficult and costly computationally than in the first-order. Therefore, in the near future, it would be wise to consider developing new techniques that can automatically decide the best order for high-order FRs.

Conflict of Interest

There is no conflict of interest.

Acknowledgement

The authors would like to thank the support of Thai Nguyen University of Technology (TNUT) to this research.

References

[1] Song, Q., Chissom, B.S., 1993. Fuzzy time series and its models. Fuzzy Sets and Systems. 54(3), 269-277.

[2] Song, Q., Chissom, B.S., 1993. Forecasting enrollments with fuzzy time series—Part I. Fuzzy Sets and Systems. 54(1), 1-9.

[3] Zadeh, L.A., 1965. Fuzzy Sets. Information and Control. 8(3), 338-353.

[4] Chen, S.M., 1996. Forecasting enrollments based on fuzzy time series. Fuzzy Sets and Systems. 81, 311-319.

[5] Huarng, K., 2001. Effective lengths of intervals to improve forecasting in fuzzy time series. Fuzzy Sets and Systems. 123(3), 387-394.

[6] Hwang, J.R., Chen, S.M., Lee, C.H., 1998. Handling forecasting problems using fuzzy time series. Fuzzy Sets and Systems. 100, 217-228.

[7] Yu, H.K., 2005. A refined fuzzy time-series model for forecasting. Physical A: Statistical Mechanics and Its Applications. 346(3-4), 657-681.

[8] Yu, H.K., 2005. Weighted fuzzy time series models for TAIEX forecasting. Physica A: Statistical Mechanics and Its Applications. 349(3-4), 609-624.

[9] Bosel, M., Mali, K., 2018. A novel data partitioning and rule selection technique for modelling high-order fuzzy time series. Applied Soft Computing. 63, 87-96.
DOI: https://doi.org/10.1016/j.asoc.2017.11.011

[10] Chen, S.M., Chung, N.Y., 2006. Forecasting enrolments using high-order fuzzy time series and genetic algorithms. International Journal of Intelligent Systems. 21, 485-501.

[11] Chen, S.M., Tanuwijaya, K., 2011. Fuzzy forecasting based on high-order fuzzy logical relationships and automatic clustering techniques. Expert Systems with Applications. 38, 15425-15437.

[12] Kuo, I.H., Horng, S.J., Kao, T.W., et al., 2009. An improved method for forecasting enrollments based on fuzzy time series and particle swarm optimization. Expert Systems with Applications. 36(3), 6108-6117.

[13] Loc, V.M., Nghia, P.T.H., 2017. Context-aware approach to improve result of forecasting enrollment in fuzzy time series. International Journal of Emerging Technologies in Engineering Research. 5(7), 28-33.

[14] Lu, W., Chen, X., Pedrycz, W., et al., 2015. Using interval information granules to improve forecasting in fuzzy time series. International Journal of Approximate Reasoning. 57, 1-18.

[15] Yolcu, U., Egrioglu, E., Uslu, V.R., et al., 2009. A new approach for determining the length of intervals for fuzzy time series. Applied Soft Computing. 9, 647-651.

[16] Liu, H.T., Wei, M.L., 2010. An improved fuzzy forecasting method for seasonal time series. Expert Systems with Applications. 37(9), 6310-6318.

[17] Dincer, N.G., Akkuş, Ö., 2018. A new fuzzy time series model based on robust clustering for forecasting of air pollution. Ecological Informatics. 43, 157-164.
DOI: http://dx.doi.org/10.1016/j.ecoinf.2017.12.001

[18] Huang, Y.L., Horng, S.J., Kao, T.W., et al., 2011. An improved forecasting model based on the weighted fuzzy relationship matrix combined with a PSO adaptation for enrollments. International Journal of Innovative Computing, Information and Control. 7(7), 4027-4046.

[19] Wang, L., Liu, X., Pedrycz, W., 2013. Effective intervals determined by information granules to improve forecasting in fuzzy time series. Expert Systems with Applications. 40(14), 5673-5679.

[20] Lu, W., Chen, X., Pedrycz, W., et al., 2015. Using interval information granules to improve forecasting in fuzzy time series. International Journal of Approximate Reasoning. 57, 1-18.

[21] Egrioglu, E., Aladag, C.H., Basaran, M.A., et al., 2011. A new approach based on the optimization of the length of intervals in fuzzy time series. Journal of Intelligent and Fuzzy Systems. 22, 15-19.

[22] Van Tinh, N., Dieu, N.C., 2018. Handling forecasting problems based on combining high-order time-variant fuzzy relationship groups and particle swam optimization technique. International Journal of Computational Intelligence and Applications. 17(2), 1-19.

[23] Kuo, I.H., Horng, S.J., Chen, Y.H., et al., 2010. Forecasting TAIFEX based on fuzzy time series and particle swarm optimization. Expert Systems with Applications. 37(2), 1494-1502.

[24] Hsu, L.Y., Horng, S.J., Kao, T.W., et al., 2010. Temperature prediction and TAIFEX forecasting based on fuzzy relationships and MTPSO techniques. Expert Systems with Applications. 37(4), 2756-2770.

[25] Huang, Y.L., Horng, S.J., He, M., et al., 2011. A hybrid forecasting model for enrollments based on aggregated fuzzy time series and particle swarm optimization. Expert Systems with Applications. 38(7), 8014-8023.

[26] Chen, S.M., Phuong, B.D.H., 2017. Fuzzy time series forecasting based on optimal partitions of intervals and optimal weighting vectors. Knowledge-Based Systems. 118, 204-216.

[27] Chen, S.M., Manalu, G.M.T., Pan, J.S., et al., 2013. Fuzzy forecasting based on two-factors second-order fuzzy-trend logical relationship groups and particle swarm optimization techniques. IEEE Transactions on Cybernetics. 43(3), 1102-1117.

[28] Chen, S.M., Chen, S.W., 2014. Fuzzy forecasting based on two-factors second-order fuzzy-trend logical relationship groups and the probabilities of trends of fuzzy logical relationships. IEEE Transactions on Cybernetics. 45(3), 391-403.

[29] Chen, S.M., Chung, N.Y., 2006. Forecasting enrollments of students by using fuzzy time series and genetic algorithms. International Journal of Information and Management Sciences. 17(3), 1-17.

[30] Lee, L.W., Wang, L.H., Chen, S.M., et al., 2006. Handling forecasting problems based on two-factors high-order fuzzy time series. IEEE Transactions on Fuzzy Systems. 14(3), 468-477.

[31] Lee, L.W., Wang, L.H., Chen, S.M., 2008. Temperature prediction and TAIFEX forecasting based on high-order fuzzy logical relationships and genetic simulated annealing techniques. Expert Systems with Applications. 34(1), 328-336.

[32] Bas, E., Uslu, V.R., Yolcu, U., et al., 2014. A modified genetic algorithm for forecasting fuzzy time series. Applied Intelligence. 41, 453-463.

[33] Panda, S., Padhy, N.P., 2007. Comparison of particle swarm optimization and genetic algorithm for FACTS-based controller design. Ap-

plied Soft Computing. 8(4), 1418-1427.

[34] Park, J.I., Lee, D.J., Song, C.K., et al., 2010. TAIFEX and KOSPI 200 forecasting based on two-factors high-order fuzzy time series and particle swarm optimization. Expert Systems with Applications. 37(2), 959-967.

[35] Samsami, R., 2013. Comparison between genetic algorithm (GA), particle swarm optimization (PSO) and ant colony optimization (ACO) techniques for NO emission forecasting in Iran. World Applied Sciences Journal. 28(12), 1996-2002.

[36] Chen, S.M., Zou, X.Y., Gunawan, G.C., 2019. Fuzzy time series forecasting based on proportions of intervals and particle swarm optimization techniques. Information Sciences. 500, 127-139.
DOI: http://dx.doi.org/10.1016/j.ins.2019.05.047

[37] Van Tinh, N., Dieu, N.C., 2019. Improving the forecasted accuracy of model based on fuzzy time series and K-Means clustering. Journal of Science and Technology: Issue on Information and Communications Technology. 3(2), 51-60.

[38] Egrioglu, E., Aladag, C.H., Yolcu, U., 2013. Fuzzy time series forecasting with a novel hybrid approach combining fuzzy c-means and neural networks. Expert Systems with Applications. 40, 854-857.

[39] Wang, W., Liu, X., 2015. Fuzzy forecasting based on automatic clustering and axiomatic fuzzy set classification. Information Sciences. 294, 78-94.

[40] Cheng, S.H., Chen, S.M., Jian, W.S., 2016. Fuzzy time series forecasting based on fuzzy logical relationships and similarity measures. Information Sciences. 327, 272-287.

[41] Kumar, S., Gangwar, S., 2015. Intuitionistic fuzzy time series: An approach for handling non-determinism in time series forecasting. IEEE Transactions on Fuzzy Systems. 24(6), 1270-1281.

[42] Pant, M., Kumar, S., 2022. Particle swarm optimization and intuitionistic fuzzy set-based novel method for fuzzy time series forecasting. Granular Computing. 7(2), 285-303.

[43] Bisht, K., Kumar, S., 2016. Fuzzy time series forecasting method based on hesitant fuzzy sets. Expert Systems with Applications. 64, 557-568.

[44] Gupta, K.K., Kumar, S., 2019. Hesitant probabilistic fuzzy set based time series forecasting method. Granular Computing. 4(4), 739-758.

[45] Jaromczyk, J.W., Toussaint, G.T., 1992. Relative neighborhood graphs and their relatives. Proceedings of the IEEE. 80(9), 1502-1517.

[46] Anand, R., Reddy, C.K., 2011. Graph-based clustering with constraints. Advances in Knowledge Discovery and Data Mining: 15th Pacific-Asia Conference, PAKDD 2011; 2011 May 24-27; Shenzhen, China. Berlin Heidelberg: Springer. Part II 15. p. 51-62.

[47] Tinh, N.V., Thanh, T.T., Thi, B.T., 2021. An interval partitioning approach using graph-based clustering in fuzzy time series forecasting model. Data Mining and Knowledge Discovery. 9, 142-150.

[48] Kennedy, J., Eberhart, R., 1995. Particle swarm optimization. Proceedings of the 1995 IEEE International Conference on Neural Networks; 1995 Nov 27-Dec ; Perth, WA, Australia. USA: IEEE. p. 1942-1948.

[49] Tinh, N.V., 2020. Enhanced forecasting accuracy of fuzzy time series model based on combined fuzzy C-mean clustering with particle swam optimization. International Journal of Computational Intelligence and Applications 19(2), 1-26.

[50] Pattanayak, R.M., Panigrahi, S., Behera, H.S., et al., 2020. High-order fuzzy time series forecasting by using membership values along with data and support vector machine. Arabian Journal for Science and Engineering. 45(12), 10311-10325.