Kỷ yếu Hội nghị Quốc gia lần thứ X về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), Đà Nẵng, ngày 17-18/08/2017
DOI: 10.15625/vap.2017.0002

A NOVEL FORECASTING MODEL BASED ON COMBINING TIME-VARIANT FUZZY LOGICAL RELATIONSHIP GROUPS AND K-MEANS CLUSTERING TECHNIQUE

Nghiem Van Tinh[1], Nguyen Cong Dieu[2]

[1]Thai Nguyen University of Technology, Thai Nguyen University, Thai Nguyen, Viet Nam
[2]Thang Long University, Ha Noi, Viet Nam

nghiemvantinh@tnut.edu.vn, ncdieu@yahoo.com

ABSTRACT: *Most of previous the forecasting approaches based on fuzzy time series (FTS) used the same length of intervals. The weakness of the static length of intervals is that the historical data are roughly put into intervals, although the variance of them is not high. In this paper, a new forecasting model based on combining the fuzzy time series and K-mean clustering algorithm with three computational methods, K-means clustering technique, the time - variant fuzzy logical relationship groups and defuzzification forecasting rules, is presented. Firstly, the authors use the K-mean clustering algorithm to divide the historical data into clusters and adjust them into intervals with different lengths. Then, based on the new intervals obtained, the proposed method is used to fuzzify all the historical data and create the time -variant fuzzy logical relationship groups based on the new concept of time – variant fuzzy logical relationship group. Finally, Calculate the forecasted output value by the improved defuzzification technique in the stage of defuzzification. To evaluate performance of the proposed model, two numerical data sets are utilized to illustrate the proposed method and compare the forecasting accuracy with existing methods. The results show that the proposed model gets a higher average forecasting accuracy rate to forecast the Taiwan futures exchange (TAIFEX) and enrollments of the University of Alabama than the existing methods based on the first – order and high-order fuzzy time series.*

Keywords*: Fuzzy time series, time – variant fuzzy logical relationship groups, K-mean clustering, enrollments, TAIFEX.*

I. INTRODUCTION

In the past decades, many forecasting models based on the concept of fuzzy time series have been proposed to deal with various domain problems, such as the enrollments forecasting [1] - [4], crop forecast [5], [6], stock markets [7], [8], temperature prediction [8], [9]. There is the matter of fact that the traditional forecasting methods cannot deal with the forecasting problems in which the historical data are represented by linguistic values. Song and Chissom proposed the time-invariant FTS model [1] and the time-variant FTS model [2] which use the max–min operations to forecast the enrollments of the University of Alabama. However, the main drawback of these methods are which required a lot of computation time when a fuzzy logical relationship matrix is large. Then, Chen [3] used simplified arithmetic operations avoiding the complicated max–min operations, and their method produced better results. Afterward, fuzzy time series has been widely studied to improve the accuracy of forecasting in many applications. Huarng [10] presented a new method for forecasting the enrollments of the University of Alabama and the TAIFEX by adding a heuristic function to get better forecasting results. Chen also extended his previous work [3] to present several forecast models based on the high-order FTS to deal with the enrollments forecasting problem [4], [11]. Yu have shown models of refinement relation [12] and weighting scheme [7] for improving forecasting accuracy. Both the stock index and enrollments are used as the targets in the empirical analysis. Huang [13] shown that different lengths of intervals may affect the accuracy of forecast. He modified previous method [10] by using the ratio-based length to get better forecasting accuracy. Recently, in [14] - [16] presented a new hybrid forecasting model which combined particle swarm optimization with fuzzy time series to find proper length of each interval by which adjust interval lengths. In addition, Dieu N. C et al. [17] introduced the concept of time-variant fuzzy logical relationship group and combined it with PSO algorithm for forecasting in fuzzy time series model. N. Van Tinh and N. C Dieu [18] extended our previous work [17] to a high-order fuzzy time series model to forecast stock market indices of TAIFEX. Some other techniques for determining best intervals and interval lengths based on clustering techniques such as; automatic clustering techniques are found [19], the K-means clustering combining the FTS in [20] and the fuzzy c-means clustering in [21]. Another way, a high-order algorithm for Multi-Variable FTS [22] based on fuzzy clustering is presented to deal various forecasting problems such as: enrollment forecasting, Gas forecasting, Rice produce prediction.

In this paper, a new hybrid forecasting model based on combining the K-mean clustering algorithm for partitioning the universe of discourse and the time – variant fuzzy logical relationship groups (FLRGs) is presented. Although the idea of using K-means clustering algorithm for partitioning historical dataset into intervals of different lengths is not novel as can be seen in [20], combining with the time – variant FLRGs in the determining of fuzzy logical relationships stage and novel forecasted rules in the defuzzification stage can help to improve the forecasting result significantly. From this view point, the proposed method is different from the approaches which also using clustering algorithms [20], [21] and [22] in the way where the fuzzy logical relationship groups and forecasted rule are created. In case study, the proposed method was applied to forecast the enrollments of the University of Alabama and the TAIFEX. The experimental results showed that the proposed method gets a higher average forecasting accuracy compared to the existing methods. In addition, the empirical results also showed that the high-order FTS model outperformed the first-order FTS model with a lower forecast error.

II. FUZZY TIME SERIES AND ALGORITHMS

2.1. Fuzzy time series

In 1993, Song and Chissom proposed the definitions of fuzzy time series, where the values of fuzzy time series are represented by fuzzy sets. Let U={$u_1, u_2, \dots, u_n$ } be an universe of discourse; a fuzzy set A of U is defined as A={ fA(u_1)/ u_1+…+fA(u_n)/ u_n },where fA is a membership function of a given set A, fA :U→[0,1], fA(u_i) indicates the grade of membership of u_i in the fuzzy set A, fA(u_i) ϵ [0, 1], and 1≤ i ≤ n . General definitions of fuzzy time series are given as follows:

Definition 1: **Fuzzy time series** [1]-[2]

Let Y(t) (t = ..., 0, 1, 2 …), a subset of R, be the universe of discourse on which fuzzy sets $f_i(t)$ (i = 1,2…) are defined and let F(t) be a collection of $f_i(t)$) (i = 1, 2…). Then, F(t) is called a fuzzy time series on Y(t) (t = . . ., 0, 1,2, . . .).

Definition 2: Fuzzy logical relationship (FLR) [3]

If there exists a fuzzy relationship R(t-1,t), such that F(t) = F(t-1)∗ R(t-1,t), where "∗ " is an max–min composition operator, then F(t) is said to be caused by F(t-1). The relationship between F(t) and F(t-1) can be denoted by F(t-1)→ F(t). Let A_i = F(t) and A_j = F(t-1), the relationship between F(t) and F(t -1) is denoted by fuzzy logical relationship $A_i \rightarrow A_j$ where A_i and A_j refer to the current state or the left - hand side and the next state or the right-hand side of fuzzy relations.

Definition 3: Fuzzy logical relationship groups (FLRGs) [3]

Fuzzy logical relationships, which have the same fuzzy set located in the left-hand side of the fuzzy logical relationships, can be grouped into a FLRG. Suppose there are exists fuzzy logical relationships as follows: $A_i \rightarrow A_{k1}$, $A_i \rightarrow A_{k2}$,.., $A_i \rightarrow A_{km}$; they can be grouped into an FLRG as : $A_i \rightarrow A_{k1}, A_{k2}, \dots, A_{km}$.

The repeated FLRs in the FLRGs are discarded by Chen [3], [14] and counted only once, but according to Yu model [7], this repeated FLRs can be accepted.

Definition 4: The λ- order fuzzy logical relationships [4]

Let *F(t)*be a fuzzy time series. If F(t) is caused by F(t-1), F(t-2),…, F(t-λ+1) F(t-λ) then this fuzzy relationship is represented by F(t-λ), …, F(t-2), F(t-1)→ F(t) and is called an $\boldsymbol{\lambda}$- order fuzzy time series.

Definition 5: Time-variant fuzzy relationship groups [17]

The relationship between F(t) and F(t-1) is determined by F(t-1)→ F(t). Let $F(t) = A_i(t)$ and $F(t-1) = A_j(t-1)$, we will have the relationship $A_j(t-1) \rightarrow A_i(t)$. At the time t, we have the following fuzzy relationships: $A_j(t-1) \rightarrow A_i(t)$; $A_j(t1-1) \rightarrow A_{i1}(t1); \dots$; $A_j(tp-1) \rightarrow A_{ip}(tp)$ with $t1, t2, \dots, tp \leq t$. It is noted that $A_i(t_1)$ and $A_i(t_2)$ has the same linguistic value as A_i, but appear at different times t_1 and t_2, respectively. It means that if the fuzzy relations occurred before $A_j(t-1) \rightarrow A_i(t)$, we can group the fuzzy logic relationship to be $A_j(t-1) \rightarrow A_{i1}(t1), A_{i2}(t2), A_{ip}(tp), A_i(t)$. It is called first – order time-variant fuzzy logical relationship group.

2.2. The time – variant FLRGs algorithm

Suppose there are fuzzy time series F(t), t =1, 2 ,…, q which it is presented by fuzzy sets as follows:

$A_{i1}, A_{i2}, \dots, A_{iq}$.

Based on the Definition 5 of the time - variant FLRGs, an algorithm is proposed as follows :

Algorithm 2.2: The λ - order time – variant fuzzy logical relationship groups algorithm

1: initialize the λ-order time – variant FLRGs t=λ ; F(1),F(2),..,F(λ-1) → F(λ) or $A_{j2},\dots,A_{j\lambda} \rightarrow A_{k1}(\lambda)$

*2: **for** t: = λ to q **do***

* **for** h: = λ **down to** 1 **do***

* Create all λ- order FLRs $A_{j2}(t-\lambda),\dots,A_{j\lambda}(t-1) \rightarrow A_{k1}(t)$*

* **end for***

*3: **for** v: = 1 to t-1 **do***

* **for** h = 1 to v **do***

* if there is fuzzy logical relation $A_{j2},\dots,A_{jm} \rightarrow A_{k2}(h)$ at the same left - hand side, then add A_{k2} into FLRGs as follows:*

* $A_{j2},\dots,A_{j\lambda} \rightarrow A_{k1}, A_{k2}$*

* end **for***

*4. end **for***

2.3. K-Means clustering algorithm

K-means clustering introduced in [20] is one of the simplest unsupervised learning algorithms for solving the well-known clustering problem. K-means clustering method groups the data based on their closeness to each other according to Euclidean distance. The result depends on the number of cluster.

The algorithm is consists of the following major steps

*Step 1:*Choose *k centroids* $\{z_1, z_2, \dots, z_k\}$

Step 2: Assign each object x to the clusters C_i*:* $x \in C_i$ *if* $d(x, z_i) < d(x, z_j)$, $j \neq i$

Step 3:update $\{z_i\}$ *to minimize* $J_i = \sum_{x \in C_i} |x - z_i|^2$, *i=1..k*

$$z_i = \frac{1}{N_i} \sum_{C_i} x = m_i$$

Step 4: Reassign the objects using the new centroids

Step 5: Repeat Steps 2, 3 and 4 until the centroids no longer move.

III. FORECASTING MODEL BASE ON K-MEANS CLUSTERING AND TIME – VARIANT FLRGS

In this section, a novel method based on combining the time - variant FLRGs and K-means clustering algorithm for forecasting the enrolments of University of Alabama, is presented. Firstly, K-means clustering algorithm is applied to classify the collected data of enrolments into clusters and adjusted these clusters into contiguous intervals for generating intervals from the enrolment data in Subsection 3.1. Then, based on the defined intervals, we fuzzify all historical enrolments data and establish time - variant FLRGs. Finally, the forecasting results are calculated from the time - variant fuzzy logical relationship groups and the proposed forecasting rules, shown in Subsection 3.2. To verify the effectiveness of the proposed model, all historical enrollments [3] (*the enrollment data at the University of Alabama from 1971s to 1992s*) are used to illustrate the first - order fuzzy time series forecasting process.

3.1. The K-Mean algorithm for generating intervals from historical dataset

The algorithm composed of 3 steps is introduced step-by-step with the same dataset [3]:

Step 1: Apply the K-means clustering algorithm to partition the historical time series data into **c** clusters and sort the data in clusters in an ascending sequence. In this paper, suppose **c** =14 clusters, the results of clusters are as follows:

{13055}, {13563, 13867}, {14696, 15145, 15163}, {15311}, {15433, 15460, 15497}, {15603}, {15861, 15984}, {16388}, {16807}, {16859}, {16919}, {18150}, {18970, 18876}, {19328, 19337}

Furthermore, the number of clusters is selected by an any way that do not exceed the total amount of data in the time series, such as c is 7, 8,9 11, ..., 22.

Step 2: Calculate the cluster centres

In this step, we use automatic clustering techniques [19] to generate cluster center ($Center_k$) from clusters as follows:

$$\boldsymbol{Center_k} = \frac{\sum_{i=1}^{n} d_i}{n} \quad (1)$$

where d_i is a datum in $cluster_k$, n denotes the number of data in $cluster_k$ and $1 \leq k \leq c$.

Step 3: Adjust the clusters into intervals according to the following rules.

Assume that $Center_k$ and $Center_{k+1}$ are adjacent cluster centers, then the upper bound $Cluster\_UB_k$ of $cluster_k$ and the lower bound $cluster\_LB_{k+1}$ of $cluster_{k+1}$ can be calculated as follows:

$$Cluster\_UB_k = \frac{Center_k + Center_{k+1}}{2} \quad (2)$$

$$Cluster\_LB_{k+1} = Cluster\_UB_k \quad (3)$$

where *k =1,.., c - 1*. Because there is no previous cluster before the first cluster and there is no next cluster after the last cluster, the lower bound $Cluster\_LB_1$ of the first cluster and the upper bound $Cluster\_UB_c$ of the last cluster can be calculated as follows:

$$Cluster\_LB_1 = \boldsymbol{Center_1} - (\boldsymbol{Center_1} - Cluster\_UB_1) \quad (4)$$

$$Cluster\_UB_c = \boldsymbol{Center_c} + (\boldsymbol{Center_c} - Cluster\_LB_c) \quad (5)$$

Then, assign each cluster Cluster_k form an interval $interval_k$, which means that the upper bound $Cluster\_UB_k$ and the lower bound $Cluster_LB_k$the cluster $cluster_k$ are also the upper bound $interval\_UB_k$ and the lower bound

$interval\_LB_k$ of the interval $interval_k$, respectively. Calculate the middle value Mid_value_k of the interval $interval_k$ as follows:

$$Mid\_value_k = \frac{interval\_LB_k + interval\_UB_k}{2} \quad (6)$$

where $interval\_LB_k$ and $interval\_UB_k$ are the lower bound and the upper bound of the interval $interval_k$, respectively, with $k = 1,..,c$.

3.2. Forecasting model based on the first – order time – variant FLRGs

In this section, we present a new method for forecasting enrollments based on the K-mean clustering algorithm and time - variant fuzzy logical relationship groups. The proposed method is now presented as follows:

Step 1: Partition the universe of discourse U into intervals

After applying the procedure K-mean clustering, we can get 14 intervals corresponding to the clusters obtained in subsection 3.1 and calculate the middle value of the intervals are listed in Table 1.

Table 1: The intervals and their midpoints obtained by K-means clustering technique

No	Intervals	MidPoint	No	Intervals	MidPoint
1	[12725, 13385)	13055	8	[16155, 16597.5)	16376.25
2	[13385, 14358)	13871.5	9	[16597.5, 16833)	16715.25
3	[14358, 15156)	14757	10	[16833, 16889)	16861
4	[15156, 15387)	15271.5	11	[16889, 17534.5)	17211.75
5	[15387, 15533)	15460	12	[17534.5, 18536.5)	18035.5
6	[15533, 15762.5)	15647.75	13	[18536.5, 19127.5)	18832
7	[15762.5, 16155)	15958.75	14	[19127.5, 19536.5)	19332

***Step 2*:** Define the fuzzy sets A_i for observations (historical data).

Each interval in Step 1 represents a linguistic variable of enrollments. For 14 intervals, there are 14 linguistic variables. Each linguistic variable represents a fuzzy set A_i ($1 \le i \le 14$). All possible values of these linguistic variables are A_1={very very very very few }, A_2={very very very few}, A_3={very very few}, A_4 ={very few }, A_5 ={few }, 6 ={moderate}, A_7={many}, A_8={many many}, A_9={very many}, A_{10}={too many}, A_{11}={too many many}, A_{12}={too many many many}, A_{13}={too many many many many} and A_{14}={too many many many many many} which can represent different intervals in the universe of discourse, and its definition is described according to Eq.(7)

$$\begin{aligned} A_1 &= {}^{1}/_{u_1} + {}^{0.5}/_{u_2} + {}^{0}/_{u_3} + \ldots + {}^{0}/_{u_{14}} \\ A_2 &= {}^{0.5}/_{u_1} + {}^{1}/_{u_2} + {}^{0.5}/_{u_3} + \ldots + {}^{0}/_{u_{14}} \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ A_{13} &= {}^{0}/_{u_1} + \ldots + {}^{0.5}/_{u_2} + {}^{1}/_{u_{13}} + {}^{0.5}/_{u_{14}} \\ A_{14} &= {}^{0}/_{u_1} + \cdots + {}^{0}/_{u_2} + {}^{0.5}/_{u_3} + {}^{1}/_{u_{14}} \end{aligned} \quad (7)$$

Here the symbol '+' denotes the set union operator and the symbol '/' denotes the membership of u_j ($1 \le j \le 14$) which belongs to A_i, respectively.

For simplicity, the different membership values of fuzzy set A_i of {0, 0.5 and 1} are selected to indicate the grade of membership of u_j in the fuzzy set A_i. According to Eq.(7), a fuzzy set contains 14 intervals. Contrarily, an interval belongs to all fuzzy sets with different membership degrees. For example, u_1 belongs to A1 and A2 with membership degrees of 1 and 0.5 respectively, and other fuzzy sets with membership degree is 0.

Step 3: Fuzzy all historical enrollments data

In order to fuzzify all historical data, it's necessary to assign a corresponding linguistic value to each interval first. The simplest way is to assign the linguistic value with respect to the corresponding fuzzy set that each interval belongs to with the highest membership degree**.** For example, the historical enrollments of year 1972 is 13563, and it belongs to interval u_2 because 13563 is within u_2= [13385, 14358]. So, we then assign the linguistic value or the fuzzy set A_2 corresponding to interval u_2. In the same way, the results of fuzzification on enrollments of the University of Alabama are listed are listed in Table 2.

Table 2. Fuzzified historical enrollments data of the University of Alabama

Year	Actual data	Fuzzy sets	Year	Actual data	Fuzzy sets
1971	13055	A1	1982	15433	A5
1972	13563	A2	1983	15497	A5
1973	13867	A2	1984	15145	A3
1974	14696	A3	1985	15163	A4
1975	15460	A5	1986	15984	A7
1976	15311	A4	1987	16859	A10
1977	15603	A6	1988	18150	A12
1978	15861	A7	1989	18970	A13
1979	16807	A9	1990	19328	A14
1980	16919	A11	1991	19337	A14
1981	16388	A8	1992	18876	A13

Step 4: Create all λ– order fuzzy logical relationships ($\lambda \geq 1$).

Based on Definition 2 and 3, to establish a λ-order fuzzy logical relationship, we should find out any relationship which has the type $\mathrm{F(t-\lambda)}, \mathrm{F(t-\lambda+1)}, \ldots, \mathrm{F(t-1)} \rightarrow F(t)$, where $\mathrm{F(t-\lambda)}, \mathrm{F(t-\lambda+1)}, \ldots, \mathrm{F(t-1)}$ and $F(t)$ are called the current state and the next state, respectively. Then a λ - order fuzzy logical relationship in the training phase is got by replacing the corresponding linguistic values. For example, supposed $\lambda = 1$ from Table 2, a fuzzy logic relationship $A_1 \rightarrow A_2$ is got as $F(1971) \rightarrow F(1972)$. So on, all first-order fuzzy logical relationships from year 1972 to 1992 are shown in column 3 and 6 of Table 3, where there are 22 fuzzy logical relationships; the first 21 relationships are called the trained patterns, and the last one is called the untrained pattern (in the testing phase). For the untrained pattern, relation 22 has the fuzzy relation $A13 \rightarrow \#$ as it is created by the relation $F(1992) \rightarrow F(1993)$, since the linguistic value of $F(1993)$ is unknown within the historical data, and this unknown next state is denoted by the symbol '#'

Table 3. The first-order fuzzy logical relationships

No FLRs	F(t)	Fuzzy relations	No FLRs	F(t)	Fuzzy relations
1	F(1971)→F(1972)	A1 -> A2	12	F(1982) → F(1983)	A5 -> A5
2	F(1972) → F(1973)	A2 -> A2	13	F(1983) → F(1984)	A5 -> A3
3	F(1973) → F(1974)	A2 -> A3	14	F(1984) → F(1985)	A3 -> A4
4	F(1974) → F(1975)	A3 -> A5	15	F(1985) → F(1986)	A4 -> A7
5	F(1975) → F(1976)	A5 -> A4	16	F(1986) → F(1987)	A7 -> A10
6	F(1976) → F(1977)	A4 -> A6	17	F(1987) → F(1988)	A10 -> A12
7	F(1977) → F(1978)	A6 -> A7	18	F(1988) → F(1989)	A12 -> A13
8	F(1978) → F(1979)	A7 -> A9	19	F(1989) → F(1990)	A13 -> A14
9	F(1979) → F(1980)	A9 -> A11	20	F(1990) → F(1991)	A14 -> A14
10	F(1980) → F(1981)	A11 -> A8	21	F(1991) → F(1992)	A14 -> A13
11	F(1981) → F(1982)	A8 -> A5	22	F(1992) → F(1993)	A13 -> #

Step 5: Establish all time - variant fuzzy logical relationship groups

In this step, a method is different from the approach in [3] and [14] in the way where the fuzzy logical relationship groups are created. In previous approach, all the fuzzy logical relationships having the same fuzzy set on the left-hand side or the same current state can be grouped into a same fuzzy relationship group. But, according to the Definition 5 and algorithm 2.2, the appearance history of the fuzzy sets on the right-hand side of fuzzy logical relationships is need to more consider. That is, only the fuzzy set on the right - hand side appearing before forecasting time which has the same fuzzy set on the left-hand side of fuzzy logical relationship is grouped into a fuzzy logical relationship group, called time – variant FLRG. From this viewpoint and based on Table 3, we can establish all time - variant fuzzy logical relationship groups are shown in column 4 of Table 4 which consists of 21 groups in training phase and one group for testing phase.

Table 4. The completed all first-order time - variant fuzzy logical relationship groups

Years	No groups	At time	Time - variant FLRGs	Linguistic variable $F(t)$
1971				
1972	1	T=1	A1 -> A2	F(1971)→ F(1972)
1973	2	T=2	A2 -> A2	F(1972)→ F(1973)
1974	3	T=2, 3	A2 -> A2, A3	$\mathrm{F(1972)\rightarrow F(1973), F(1974)}$
1975	4	T=4	A3 -> A5	F(1974) → F(1975)
1976	5	T=5	A5 -> A4	F(1975) → F(1976)
1977	6	T=6	A4 -> A6	F(1976) → F(1977)

Years	No groups	At time	Time - variant FLRGs	Linguistic variable $F(t)$
1978	7	T=7	A6 -> A7	F(1977) → F(1978)
1979	8	T=8	A7 -> A9	F(1978) → F(1979)
1980	9	T=9	A9 -> A11	F(1979) → F(1980)
1981	10	T=10	A11 -> A8	F(1980) → F(1981)
1982	11	T=11	A8 -> A5	F(1981) → F(1982)
1983	12	T=5, 12	A5 -> A4, A5	F(1982) → F(1976), F(1983)
1984	13	T=5,12, 13	A5 -> A4, A5, A3	F(1983) → F(1976), F(1983), F(1984)
1985	14	T=4, 14	A3 -> A5, A4	F(1984) → F(1975), F(1985)
1986	15	T=6,15	A4 -> A6, A7	F(1985) → F(1977), F(1986)
1987	16	T=8, 16	A7 -> A9, A10	F(1986) → F(1979), F(1987)
1988	17	T=17	A10 -> A12	F(1987) → F(1988)
1989	18	T=18	A12 -> A13	F(1988) → F(1989)
1990	19	T=19	A13 -> A14	F(1989) → F(1990)
1991	20	T=20	A14 -> A14	F(1990) → F(1991)
1992	21	T=20, 21	A14 -> A14, A13	F(1991) → F(1991), F(1992)
1993	22	T=22	A13 -> #	F(1992) → F(1993)

Step 6: Defuzzify and calculate the forecasting values for all time –variant FLRGs

In this step, to obtain the forecasted results, a new defuzzification technique is presented to calculate the forecasted values for all time – variant FLRGs in training phase. Then we also use defuzzification rule is proposed in [14] for the time – variant FLRGs in testing phase.

For the training phase , we estimate forecast values for all time – variant FLRGs based on fuzzy sets on the right-hand within the same group. For each group in column 4 of Table 4, we divide each corresponding interval of each next state into p sub-intervals with equal length, and calculate a forecasted value for each group according to Eq. (8).

$$\text{forecasted}_{\text{output}} = \frac{1}{\text{n}}\sum_{\text{j=1}}^{\text{n}} subm_{kj} \tag{8}$$

where,($1 \leq j \leq$ n, $1 \leq k \leq$ p)

- n is the total number of next states or the total number of fuzzy sets on the right-hand side within the same group.

- $subm_{kj}$ is the midpoint of one of p sub-intervals (*means the midpoint of j-th sub-interval)* corresponding to j-th fuzzy set on the right-hand side where the highest level of A_{kj} takes place in this interval.

For instance, in column 4 of Table 4, we can see that there is a first –order time - variant FLRGs "A1 → A2" in Group 1 which has only one fuzzy set on the right - hand side as A2 ; where the highest membership level of A2 belongs to interval $u_2 = [13385, 14358)$. In this paper, we divide the interval u_2 into four sub-intervals which are $u_{2.1} = [13385, 13628.25)$, $u_{2.2} = [13628.25, 13871.5)$, $u_{2.3} = [13871.5, 14114.75)$, $u_{2.4} = [14114.75, 14358)$. In Table 4, the first - order time –variant FLRG group A1 → A2 is got as F(1971) → F(1972) ; where the historical data of year 1972 is 13563 and it is within sub-interval $u_{2.1} = [13385, 13628.25)$ and then the midpoint $subm_{2.1}$ of sub-interval $u_{2.1}$ is 13506.63. The finally, forecasted value for Group 1 according to Eq. (8) is 13506.63. Forecasted value of remaining first – order time – variant FLRGs are calculated in a similar manner

For the testing phase, we calculate a forecasted value based on master voting (MV) scheme to deal with the untrained pattern [14].

$$Forecated_{for\#} = \frac{(m_{t1} * w_h) + m_{t2} + \cdots + m_{ti} + \cdots + m_{t\lambda}}{w_h + (\lambda - 1)};\ \text{i}=\overline{1:\lambda} \tag{9}$$

Where the symbol w_h means the highest votes predefined by user, λ is the order of the fuzzy relationship, m_{ti} denote the midpoints of the corresponding intervals.

Based on the forecast rules are presented in Eq. (8) and (9), we complete forecasted results for all first-order time - variant fuzzy logical relationship groups are listed in Table 5.

Table 5. The complete forecasted values for all first-order time - variant fuzzy logical relationship groups (FLRGs)

No group	Time –variant FLRGs	Value	No group	Time –variant FLRGs	Value
1	A1 -> A2	13547	12	A5 -> A4, A5	15429
2	A2 -> A2	13872	13	A5 -> A4, A5, A3	15293
3	A2 -> A2, A3	14314	14	A3 -> A5, A4	15327
4	A3 -> A5	15460	15	A4 -> A6, A7	15765

No group	Time –variant FLRGs	Value	No group	Time –variant FLRGs	Value
5	A5 -> A4	15348	16	A7 -> A9, A10	16827
6	A4 -> A6	15571	17	A10 -> A12	18036
7	A6 -> A7	15828	18	A12 -> A13	19029
8	A7 -> A9	16794	19	A13 -> A14	19332
9	A9 -> A11	16997	20	A14 -> A14	19332
10	A11 -> A8	16376	21	A14 -> A14, A13	19082
11	A8 -> A5	15411	22	A13 -> #	18832

Based on Table 5 and the data in Table 2, we complete forecasted results for enrollments from 1971 to 1992 based on first-order fuzzy time series model with 14 intervals are listed in Table 6.

Table 6. The complete forecasted output values based on the first–order FTS under number of intervals of 14

Year	Actual data	Fuzzy set	Forecasted value	Forecasted- actual
1971	13055	A1	Not forecasted	-
1972	13563	A2	13547	-16
1973	13867	A2	13872	5
1974	14696	A3	14314	-382
-----	------	----	------	----
1991	19337	A14	19332	-5
1992	18876	A13	19082	206
1993	N/A	#	18832	-

To evaluate the performance of the proposed model, the mean square error (MSE) is employed as an evaluation criterion to represent the forecasted accuracy. The MSE value is calculated as follows:

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

where, R_i denotes actual data at year i, F_i is forecasted value at year i, n is number of the forecasted data, λ is order of the fuzzy logical relationships

IV. EXPERIMENTAL RESULTS

In this paper, the proposed method is utilized to forecast the enrolments of University of Alabama with the whole historical data [3], the period from 1971 to 1992 and handles other forecasting problems, such as the empirical data for the TAIFEX [25] from 8/3/1998 to 9/30/1998, used to perform comparative study in the training phase.

4.1. Experimental results for forecasting enrollments

Actual enrollments of the University of Alabama [3] are used to perform comparative study in the training and testing phases. In order to verify forecasting effectiveness, the proposed model is compared with those of corresponding models for various orders and different intervals. The forecasted accuracy of the proposed method is estimated according to Eq. (10).

4.1.1. Experimental results from the training phase.

In order to verify the forecasting effectiveness of our proposed model for the first – order FRLGs under different number of intervals, five FTS models are examined and compared. There are the **SCI** model [1], the **C96** model [3], the **H01** model [10], **CC06a** model [23] and **HPSO** model [14]. A comparison of the forecasting results among these models is shown in Table 7. It is obvious that the proposed model gets the smallest MSE value of **15139** among all the compared models with different number of intervals. The major difference between the CC06a model, HPSO model and our models is that at the defuzzification stage and optimization method is used. Two models in CC06a [23], HPSO [14] use the genetic algorithm and the particle swarm optimization algorithm to get the appropriate intervals, respectively, while the proposed model performs the K- mean algorithm to achieve the best interval lengths.

Table 7. A comparison of the forecasted results for the first-order FLRGs with 14 intervals

Year	Actual data	SCI	C96	H01	CC06a	HPSO	Our model
1971	13055	---	---	---	---	---	---
1972	13563	14000	14000	14000	13714	13555	13547
1973	13867	14000	14000	14000	13714	13994	13872
1974	14696	14000	14000	14000	14880	14711	14314
-----	------	------	------	-----	-----	-----	-----
1991	19337	19000	19000	19500	19149	19340	19332
1992	18876	19000	19000	19149	19014	19014	19082
1993	*N/A*						*18832*
MSE		**423027**	**407507**	**226611**	**35324**	**22965**	**15139**

The trend in forecasting of enrollments by first-order fuzzy time series model in comparison to the actual enrollments can be visualized in Fig. 1. From Fig. 1, it can be seen that the forecasted value is close to the actual enrolments each year, from 1972s to 1992s than the compared models

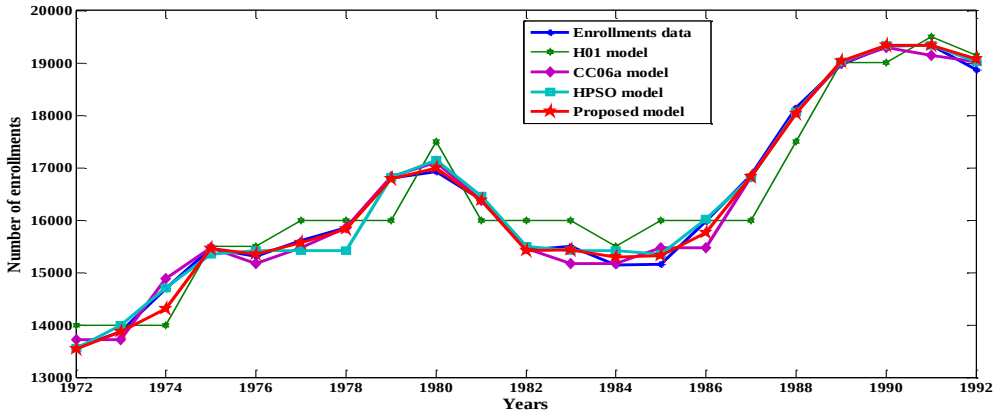

Fig. 1. The curves of the actual data and the H01, CC06a, HPSO models and our proposed model for forecasting enrollments of University of Alabama

To verify the forecasting effectiveness for high-order fuzzy time series, four existing forecasting models, the **CC06b** [11], **HPSO** [14], **AFPSO** [16] models and the **C02** model [4] are used to compare with the proposed model. A comparison of the forecasted results is listed in Table 8 where the number of intervals is seven for all forecasting models. From Table 8, it is clear that the proposed model is more precise than the four forecast models at all, since the best and the average fitted accuracies are all the best among the five models. Practically, at the same intervals, the proposed method obtains the lowest MSE values which are 8168, 6853, 6767, 6785, 3951, 3781, 6459 for 3-order, 4-order, 5-order, 6-order, 7-order, 8-order and 9-order fuzzy logical relationships, respectively. In addition, performance of proposed model is also compared with HMV-FTS algorithm [22] using enrollments dataset based on the second - order FTS with at the same interval of 7. Although the proposed model and the HMV-FTS model both use the clustering algorithm to attain the best interval lengths, but the proposed model gets lower MSE value of 16356 for the second – order fuzzy logical relationships, as the major difference between the HMV-FTS model and the proposed model is in the defuzzification rules and the establishment of fuzzy logical relationship groups used. The proposed model also gets the smallest MSE value of **3781** for the 8th-order fuzzy logical relationships among all orders of forecasting model.

Table 8. A comparison of the forecasted enrollments under various high-order FTS models with seven intervals.

Order	C02 [4]	CC06b [11]	HPSO [14]	AFPSO [16]	HMV-FTS[22]	Our model
2	N/A	N/A	N/A	N/A	22722	16356
3	86694	31123	31644	31189	N/A	8168
4	89376	32009	23271	20155	N/A	6853
5	94539	24948	23534	20366	N/A	6767
6	98215	26980	23671	22276	N/A	6785
7	104056	26969	20651	18482	N/A	3951
8	102179	22387	17106	14778	N/A	**3781**
9	102789	18734	17971	15251	N/A	6459

4.1.2. Experimental results in the testing phase.

To verify the forecasting accuracy for future enrollments, the historical enrollments are separated two parts for independent testing. The first part is used as training data set and the second part is used as the testing data set. In this paper, the historical data of enrollments from year 1971 to 1989 is used as the training data set and the historical data of enrollments from year 1990 to 1992 is used as the testing data set. For example, to forecast a new enrolment of 1990, the enrollments of 1971-1989 are used as the training data. Similarly, a new enrolment of 1991 can be forecasted based on the enrollments under years 1971-1990. After the training data have been well trained by the proposed model, future enrollments could be obtained to compare with testing data. Some experimental results of the forecasting models for the testing phase are listed in Table 9.

Table 9. A comparison of actual data and forecasted result for 14 intervals in the testing phase

Year	Actual enrollments	Forecasted value				
		1^{st} - order	2^{nd} - order	3^{rd} - order	4^{th} - order	5^{th} - order
1990	19328	18560	18560	18502	18563	18455
1991	19337	19149	19149	19087	19082	19058
1992	18876	18946	18946	18946	19030	19012

4.2. Experimental results for the TAIFEX forecasting

In this paper, we also apply the proposed method to forecast the TAIFEX index with the whole historical data [25] are used. To verify the superiority in the forecasted accuracy of the proposed model with the high-order FLRGs under numbers of intervals is 16, six FTS models **C96** [3], **H01b** [10], **L06** [24], **L08** [25], **HPSO** [14] and **MTPSO** model [26] are selected for purposes of comparison. A comparison of the forecasted results is listed in Table 10 where all forecasting models use high-order fuzzy logical relationships under different number of intervals.

Table 10. A comparison of the forecasted results of the proposed method with the existing models based on high – order of the fuzzy time series under number of intervals is 16.

Date	Actual data	C96 [3]	H01[10]	L06 [24]	L08 [25]	HPSO [14]	MTPSO [26]	Our model
8/3/1998	7552							
8/4/1998	7560	7450	7450					
8/5/1998	7487	7450	7450					
8/6/1998	7462	7500	7500	7450				
8/7/1998	7515	7500	7500	7550				
8/10/1998	7365	7450	7450	7350				
8/11/1998	7360	7300	7300	7350				
8/12/1998	7330	7300	7300	7350	7329	7289.56	7325.28	7326.69
8/13/1998	7291	7300	7300	7250	7289.5	7320.77	7287.48	7291.19
----------	-----	-----	-----	-----	------	------	-------	------
9/29/1998	6806	6850	6850	6850	6796	6800.07	6781.01	6811.38
9/30/1998	6787	6850	6750	6750	6796	7289.56	6781.01	6784.88
10/1/1998	N/A							6811.01
MSE		**9668.94**	**5437.58**	**1364.56**	**105.02**	**103.61**	**92.17**	**50.2**

In addition to, to demonstrate the effectiveness of the proposed model, two forecasting models based on high-order FTS are selected to be compared with proposed model. These two models are proposed by L08 model [25], HPSO model [14], respectively. The forecasted errors by MSE value of all models are listed in Table 11.

Table 11. A comparison of the MSE of the proposed model with that of L08 and HPSO model for the training phase based on high – order FLRGs.

Models	3^{rd}- order	4^{th}- order	5^{th}- order	6^{th}- order	7^{th}- order	8^{th}- order
L08	208.79	142.26	143.31	147.14	105.02	124.48
HPSO	152.47	148.14	112.24	122.68	103.61	108.37
Our model	70	59.4	57.4	52.2	**50.2**	57.6

From Table 11, the experimental results show that our proposed model bears all the smallest MSE in ten testing times. From these results, it is obvious that our model significantly outperforms the models proposed by L08 model [25] and HPSO model [14] and obtains the smallest MSE value of **50.2** for the 7th-order FLRGs. From Fig. 2, it can see that the forecasting values of the proposed model is close to the actual data than the compared models.

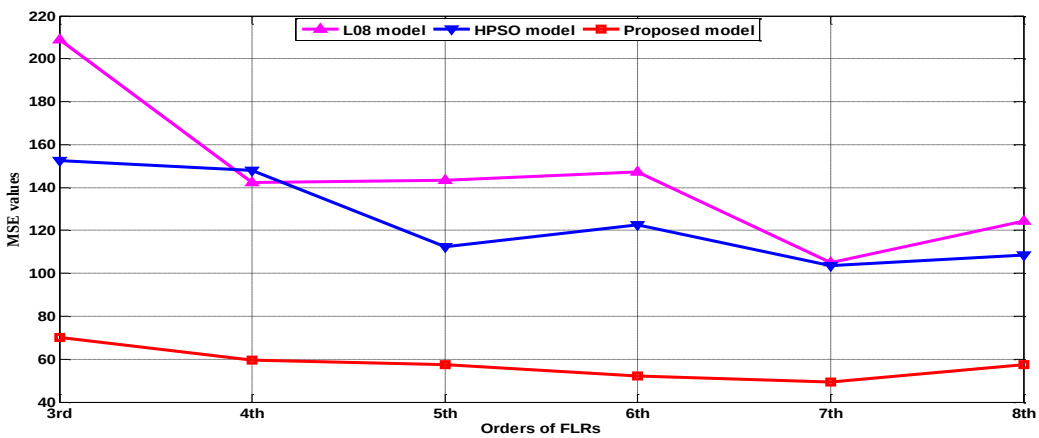

Fig. 2. A comparison of the MSE values for 16 intervals with different high-order FLRGs

V. CONCLUSION

In this paper, a hybrid forecasting model based on combining fuzzy time series model with time - variant fuzzy logical relationship groups and K-mean clustering algorithm was proposed. Using K -mean algorithm, our model found out proper lengths of intervals in the universe of discourse and establishing time – variant fuzzy logical relationship groups to obtain more exactly information served in the stage of defuzzification output. Particularly, the consideration

of more information on the right-hand side of fuzzy logical relationships in the same time –variant FLRG, the proposed model has improved the forecasting results, significantly. From the experimental study on the enrollments forecasting of the University of Alabama and TAIFEX forecasting, the results have shown that the proposed model has higher forecasting accuracy than some compared models at all. Specially, with the high-order fuzzy time series from 2^{nd} to 9^{th} for forecasting enrollments and from 3^{rd} to 8^{th} for TAIFEX prediction, our model is much more effective compared to the existing models. It also performs best for fuzzy time series with various orders of fuzzy logical relationships in the testing phases. This study has discovered the synergistic effect of K-mean clustering algorithm and time - variant fuzzy logical relationship groups in the stage of determining of fuzzy logical relationships, and also proposed a new fuzzy solving rule in the defuzzification stage to improve the forecasting accuracy. The researched results have shown the proposed model outperforms models compared for the training with various orders and different interval lengths. These results are very promising for the future work on the development of fuzzy time series and K - mean clustering algorithm in real-world forecasting applications.

REFERENCES

[1] Q. Song, B. S. Chissom. "Forecasting Enrollments with Fuzzy Time Series – Part I," Fuzzy set and system, vol. 54, pp. 1-9, 1993b.

[2] Q. Song, B. S. Chissom. "Forecasting Enrollments with Fuzzy Time Series – Part II," Fuzzy set and system, vol. 62, pp. 1-8, 1994.

[3] S. M. Chen. "Forecasting Enrollments based on Fuzzy Time Series," Fuzzy set and system, vol. 81, pp. 311-319, 1996.

[4] S. M. Chen. "Forecasting enrollments based on high-order fuzzy time series", Cybernetics and Systems: An International Journal, vol. 33, pp. 1-16, 2002.

[5] Singh, S. R. A simple method of forecasting based on fuzzytime series. Applied Mathematics and Computation, 186, 330–339, 2007a.

[6] Singh, S. R. A robust method of forecasting based on fuzzy time series. Applied Mathematics and Computation, 188, 472–484, 2007b.

[7] H. K.. Yu. "Weighted fuzzy time series models for TAIEX forecasting ", Physica A, 349 , pp. 609–624, 2005.

[8] Lee, L.-W., Wang, L.-H., & Chen, S.-M. Temperature prediction and TAIFEX forecasting based on fuzzy logical relationships and genetic algorithms. Expert Systems with Applications, 33, 539–550, 2007.

[9] Wang, N.-Y, & Chen, S.-M. Temperature prediction and TAIFEX forecasting based on automatic clustering techniques and two-factors high-order fuzzy time series. Expert Systems with Applications, 36, 2143–2154, 2009.

[10] Huarng, K, 2001b. Heuristic models of fuzzy time series for forecasting. Fuzzy Sets and Systems, 123, 369–386 .

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

[12] H. K. Yu. A refined fuzzy time-series model for forecasting, Phys. A, Stat. Mech. Appl. 346, 657–681, 2004; http://dx.doi.org/10.1016/j.physa.07.024.

[13] Huarng, K. H., Yu, T. H. K. "Ratio-Based Lengths of Intervals to Improve Fuzzy Time Series Forecasting," IEEE Transactions on SMC - Part B: Cybernetics, Vol. 36, pp. 328–340, 2006.

[14] Kuo, I. H. et al. An improved method for forecasting enrollments based on fuzzy time series and particle swarm optimization. Expert Systems with applications, 36, 6108–6117, 2009

[15] I. H. Kuo et al. Forecasting TAIFEX based on fuzzy time series and particle swarm optimization, Expert Systems with Applications. 37, 1494–1502, 2010.

[16] Huang, Y. L. et al. A hybrid forecasting model for enrollments based on aggregated fuzzy time series and particle swarm optimization. Expert Systems with Applications, 38, 8014–8023, 2011.

[17] Nguyen Cong Dieu, Nghiem Van Tinh, Fuzzy time series forecasting based on time-depending fuzzy relationship groups and particle swarm optimization, In :Proceedings of the 9th National conference on Fundamental and Applied Information Technology Research(FAIR'9), pp.125-133, 2016.

[18] Nghiem Van Tinh, Nguyen Cong Dieu, An improved method for stock market forecasting combining high-order time-variant fuzzy logical relationship groups and particle swam optimization in : Proceedings of the International Conference, Advances in Information and Communication Technology, pp.153-166, 2016.

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

[20] Zhiqiang Zhang, Qiong Zhu. "Fuzzy time series forecasting based on k-means clustering", Open Journal of Applied Sciences, 100-103, 2012.

[21] Bulut, E., Duru, O., & Yoshida, S. A fuzzy time series forecasting model formulti-variate forecasting analysis with fuzzy c-means clustering. WorldAcademy of Science, Engineering and Technology, 63, 765–771, 2012.

[22] S. Askari, N. Montazerin, A high-order multi-variable Fuzzy Time Series forecasting algorithm based on fuzzy clustering, Expert Systems with Applications ,42, 2121–2135, 2015.

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

[24] Lee, L. W. et al. Handling forecasting problems based on two-factors high-order fuzzy time series. IEEE Transactions on Fuzzy Systems, 14, 468–477, 2006.

[25] Lee, L.-W. Wang, L.-H., & Chen, S.-M. "Temperature prediction and TAIFEX forecasting based on high order fuzzy logical relationship and genetic simulated annealing techniques", Expert Systems with Applications, 34, 328–336, 2008b.

[26] Ling-Yuan Hsu et al. Temperature prediction and TAIFEX forecasting based on fuzzy relationships and MTPSO techniques, Expert Syst. Appl.37, 2756–2770, 2010.

MÔ HÌNH DỰ BÁO MỚI DỰA TRÊN SỰ KẾT HỢP GIỮA NHÓM QUAN HỆ MỜ PHỤ THUỘC THỜI GIAN VÀ KỸ THUẬT PHÂN CỤM K - MEAN

Nghiêm Văn Tính, Nguyễn Công Điều

***TÓM TẮT**: Hầu hết các phương pháp dự báo chuỗi thời gian mờ trước đây đều sử dụng độ dài khoảng giống nhau. Hạn chế khi sử dụng độ dài khoảng bằng nhau là các dữ liệu lịch sử được đưa vào các khoảng này một cách cứng nhắc, ngay cả khi dữ liệu có sự biến đổi không mạnh. Trong bài báo này, một mô hình dự báo mới dựa trên việc kết hợp giữa chuỗi thời gian mờ và thuật toán phân cụm K –mean cùng với ba cách tính như kỹ thuật phân cụm K –mean, nhóm quan hệ mờ phụ thuộc thời gian và các quy tắc giải mờ dự báo được trình bầy. Trước tiên, chúng tôi sử dụng thuật toán K – mean để chia tập dữ liệu lịch sử thành các cụm và điều chỉnh chúng thành các khoảng có độ dài không bằng nhau. Sau đó dựa vào các khoảng đạt được này, mô hình đề xuất được sử dụng bằng việc mờ hóa tất cả dữ liệu lịch và tạo các nhóm quan hệ logic mờ phụ thuộc vào từng thời điểm dựa vào khái niệm mới là nhóm quan hệ mờ phụ tuộc thời gian. Cuối cùng, tính toán kết quả đầu ra dự báo bằng quy tắc giải mờ đề xuất trong giai đoạn giải mờ. Để đánh giá hiệu quả của mô hình đề xuất, hai tập dữ liệu được sử dụng làm minh chứng và so sánh độ chính xác dự báo với các mô hình đề xuất trước đây. Kết quả dự báo đã cho thấy rằng, mô hình đề xuất đưa ra mức độ chính xác dự báo tốt hơn so với các mô hình trước đây khi thực hiện dự báo thị trường chứng khoán Đài Loan (Taiwan futures exchange –TAIFEX) và dự báo số lượng sinh viên nhập học của trường Đại học Alabama dựa trên cả hai chuỗi thời gian mờ bậc 1 và bậc cao.*

***Từ khóa**: Chuỗi thời gian mờ, nhóm quan hệ mờ phụ thuộc thời gian, phân cụm K -mean, số lượng tuyển sinh, TAIFEX.*