A New Fuzzy Time Series Model Based on Cluster Analysis Problem

Tai Vovan & Nghiep Ledai

International Journal of Fuzzy Systems

ISSN 1562-2479
Volume 21
Number 3

Int. J. Fuzzy Syst. (2019) 21:852-864
DOI 10.1007/s40815-018-0589-x

Int. J. Fuzzy Syst. (2019) 21(3):852–864
https://doi.org/10.1007/s40815-018-0589-x

A New Fuzzy Time Series Model Based on Cluster Analysis Problem

Tai Vovan[1] · Nghiep Ledai[2]

Received: 12 July 2018 / Revised: 28 September 2018 / Accepted: 20 November 2018 / Published online: 5 January 2019

Abstract This article proposes a new fuzzy time series (NFTS) model that can interpolate historical data to forecast effectively for the future. In this model, after normalizing original data, we establish the automatic algorithm to determine the suitable number of clusters and to find the fuzzy relationships of each element in series to the established clusters. A principle for forecasting is also proposed from these established fuzzy relationships. The convergence of the proposed algorithm is proven by theory and shown by the numerical examples. The calculation of the proposed model can be performed conveniently and efficiently by a complete Matlab procedure. Comparing with many existing models from a lot of well-known data sets with various scales and characteristics, NFTS model has shown prominent advantages.

Keywords Algorithm · Cluster · Forecast · Fuzzy time series · Interpolate

1 Introduction

Forecasting is the process of making predictions based on historical data, knowledge and experience of the related problems. Because of its important role in many fields, forecasting has been paying much attention by scientists. Although there are many discussions in the literature, it has not yet been completely solved. With time series, a common data type in reality, two major models used for forecasting are regression and time series. When constructing a regression model, we must constrain on the data conditions that are difficult to satisfy in reality. Therefore, this model has a number of limitations in many applications. Time series model was evaluated to be more advantageous in reality, so it is used very commonly today [10, 21, 27, 28]. Many researchers have used the time series models such as autoregressive, moving average and autoregressive integrated moving average (ARIMA) for applications in economics, environment, hydrology. However, when building these models, we also have to accept some conditions where the actual data are not satisfactory. As a result, they have shown disadvantage in many cases. Although many authors in [2, 3, 16, 22, 35] have tried to improve original model, they still have many drawbacks in forecasting for the real problems. This model is evaluated better than others based on the specific data only that not for all of the cases. The traditional time series models cannot deal with forecasting problems in which the historical data are presented by linguistic values. Fuzzy time series (FTS) model has been proposed to solve this drawback. FTS model is developed in two main directions. The first one is to build the FTS model from the original data and directly use this model to forecast. Abbasov and Manedova [1] had important contributions to this direction. The second one is to interpolate data in order get the relation between elements in series and then to use this fuzzy data to forecast by the known forecasting models. This research has been of great interest by many statisticians. Song and Chissom [28] were the pioneer in this direction with data on enrollment of the University of Alabama (EnrollmentUA). Quang [25] used the triangular fuzzy relation for performing. Ming et al. [9, 23] improved the model of Qiang and Brad [25]

✉ Tai Vovan
vvtai@ctu.edu.vn

[1] College of Natural Science, Can Tho University, Can Tho, Vietnam

[2] College of Basis Science, Nam Can Tho University, Can Tho, Vietnam

when taking notice of fuzzy level. Huarng et al. [18, 24] presented a heuristic model for FTS using heuristic knowledge to improve the forecast for EnrollmentUA. Based on neural network, the model of Alpaslan et al. [5] gave the interesting results in some cases. From the fuzzy model in accordance with different linguistic levels, many scientists such as [17, 21, 27, 31] have proposed the new models.

A FTS model usually consists of three stages: (i) determining universal set, dividing intervals for universal set and finding the number of elements for each interval, (ii) building the fuzzy relationships, and (iii) defuzzification for data. For (i), many authors used the values min and max of original data to divide the interval for a universal set [9, 10]. In addition, Huarng et al. [18, 19] proposed two new techniques for finding intervals based on the mean of the distributions. Abbasov and Manedova [1] have built the universal set based on the change of data between consecutive periods of time or their percentage change. Determining the number of fuzzy sets and elements in each fuzzy set is very important for establishing a model. Many authors divided the number of fuzzy set based on testing for many cases to find the suitable number for each case. This means that it is not a common rule for all. The number of fuzzy sets and their elements was also determined by the k-mean [36] and the genetic algorithm [14]. According to our knowledge, although there are a lot of discussions about this problem, the optimal choice has not been still found yet so far. For (ii), several important studies have been performed. For instance, Song and Chissom [28] used matrix operations, and Chen [10] took the fuzzy logic relations. Moreover, many authors in [4, 11–13, 19] used artificial neural networks to determine fuzzy relations. In addition, the fuzzy relationship based on the triangle and trapezoid fuzzy number was also considered in [17]. For (iii), many studies had used either the centroid method [10, 18, 19] or the adaptive expectation method [3, 9] to perform.

This article contributes to three stages: (i), (ii), and (iii) for FTS model. For (i), after normalizing data, we proposed an automatic algorithm to determine the suitable numbers of fuzzy set for each series. The number of fuzzy sets depends the similar levels of elements in series. This method is more suitable than existing ones that were presented as linguistic values with levels are constant. (It is usually five or seven in applications.) This algorithm also gives specific clusters of series. For (ii), we also build an automatic algorithm to find the fuzzy relationships between each element in series with the established clusters from (i). For (iii), based on the principle for normalizing series and the fuzzy relations found from (ii), a new defuzzification method is also proposed. Incorporating all these improvements, we have a new fuzzy time series (NFTS) model better than the existing ones through many well-known data sets. The convergence of the proposed algorithms is considered by theory and illustrated by the numerical examples. We also establish the Matlab procedure for the proposed model. This procedure can perform effectively the NFTS model for numerical examples. In addition, we also apply the proposed model to forecast flood peak for the main river in Vietnam.

The remainder of this article is organized as follows. Section 2 reviews some basic concepts of FTS model, proposes a new FTS model and considers the convergence of the proposed model. Section 3 presents numerical examples to illustrate for the present theories. This section also compares the proposed model with some existing models. A real application that it is very urgent in Vietnam is present in Sect. 4. The final section is destined for the conclusion.

2 Some Definitions and the Proposed Algorithm

2.1 Definitions

Definition 1 Let U be universe of discourse, $U = \{u_2, u_2, \ldots, u_n\}$. A fuzzy set A of U is defined as follows:

$$A = \{\mu_A(u_1)/u_1, \mu_A(u_2)/u_2, \ldots, \mu_A(u_n)/u_n\},$$

where $\mu_A(u_i)$ is the membership function of A, $\mu_A(u_i) : U \to [0, 1]$, $\mu_A(u_i)$ indicates the grade of membership of u_i in $\mu_A(u_i) \in [0, 1], 1 \le i \le n$.

Definition 2 Let $X(t), (t = 1, 2, \ldots)$, a subset of real numbers be the universe of discourse by which the fuzzy sets $f_i(t)$ are defined. If $F(t)$ is a collection of $f_1(t), f_2(t), \ldots$, then, $F(t)$ is called a FTS defined on $X(t)$.

Definition 3 Given a chain of historical data $\{X_i\}$ and predictive value $\{\widehat{X}_i\}$, $i = 1, 2, \ldots, n$, respectively, then we have the popular parameters to evaluate built FTS models as follows:

Mean squared error:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{X}_i - X_i\right)^2. \quad (1)$$

Mean absolute error:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{X}_i - X_i\right|. \quad (2)$$

Mean absolute percentage error:

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left|\hat{X}_i - X_i\right|}{X_i}\cdot 100\right). \quad (3)$$

Symmetric mean absolute percentage error:

$$\text{SMAPE} = \sum_{i=1}^{n} \left(\frac{|\hat{X}_i - X_i|}{(X_i + \hat{X}_i)/2} 100 \right). \quad (4)$$

Mean absolute scaled errors:

$$\text{MASE} = \frac{\sum_{i=1}^{n} |\hat{X}_i - X_i|}{\frac{n}{n-1} \sum_{i=2}^{n} |X_i - X_{i-1}|}. \quad (5)$$

2.2 The Proposed Model

Assume that the data set X_i corresponds to time $t_i, i = 1, 2, \ldots, n$. A new fuzzy time series (NFTS) model with 5 steps is proposed as follows:

Step 1. Standardizing data on scale 10, $Y_i = 10X_i/\max\{X_i\}, i = 1, 2, \ldots, n$. Then, we have the universal set $U = \{Y_i, i = 1, 2, \ldots, n\}$.

Step 2. Determining the suitable number of clusters for the universal set U. This problem is performed by the SNC algorithm (suitable number of clusters). This algorithm has 3 steps as follows:

Step 2.1. Initialize $t = 0$ and $Z^{(0)} = \{z_1^{(0)}, z_2^{(0)}, \ldots, z_n^{(0)}\} = (Y_1, Y_2, \ldots, Y_n\}$.

Step 2.2. Every fuzzy data point is updated according to

$$z_i^{(t+1)} = \frac{\sum_{i'=1}^{n} f(z_i^{(t)}, z_{i'}^{(t)}) z_{i'}^{(t)}}{\sum_{i'=1}^{n} f(z_i^{(t)}, z_{i'}^{(t)})}, \quad (6)$$

where *f*(.) is the truncated Gauss kernel:

$$f(z_i^{(t)}, z_{i'}^{(t)}) = \begin{cases} \exp(-d/\lambda) & \text{if } d(z_i^{(t)}, z_{i'}^{(t)}) \le d_s, \\ 0 & \text{if } d(z_i^{(t)}, z_{i'}^{(t)}) > d_S, \end{cases} \quad (7)$$

with λ is constant, $d(z_i^{(t)}, z_{i'}^{(t)})$ is measure for similarity between $z_i^{(t)}$ and $z_{i'}^{(t)}$ and d_s is the mean of measures of all pair elements:

$$d_S = \frac{2}{n(n-1)} \sum_{i<i'} d(z_i^{(t)}, z_{i'}^{(t)}), \quad (8)$$

$d(.)$ is distance between the prototype elements of two clusters. The larger d is, the smaller the value of the truncated Gauss kernel is. λ measures variance of the truncated Gauss kernel. The larger λ is, the larger the standard deviation of each established clusters is taken. Then, the number of clusters for the universal set is otherwise. When $\lambda \to 0$, the data have n intervals and when $\lambda \to \infty$, the data have only one interval. In studying about cluster analysis problem, Chen and Hung [8] have taken $\lambda = 5$. We see that this value is not suitable for the considered series. To take the suitable value of λ for all series, Step 1 of the proposed algorithm has standardized data on scale 10. Performing with many time series, we choose $\lambda = 16$ in numerical examples.

Step 2.3. Repeat *Step 2.2* until the following condition is satisfied:

$$\max_i\{d(z_i^{(t)}, z_i^{(t+1)})\} < \varepsilon.$$

In the SNC algorithm, after an iteration has finished, each element in data set will converge to the representative element $z_i^{(t)}, i = 1, 2, \ldots, c$. When the algorithm stops, we have sequences of c representative elements, and c is the number of clusters divided for the universal set.

Step 3. Determining the elements in each cluster w_i and the fuzzy relation μ_{ij} from each element Y_i to the cluster $w_j; i = 1, 2, \ldots, n; j = 1, 2, \ldots, c$. This problem is performed by the DFR algorithm (determining fuzzy relation) as follows:

Step 3.1. Divide U into c clusters $w_1, w_2, \ldots, w_c$ randomly. Establish the initial partition matrix $U^{(0)} = [\mu_{ij}]_{k\times n}$, with $\mu_{ij} = 1$ if the jth element belongs to the w_i and $\mu_{ij} = 0$ for otherwise.

Step 3.2. Find the representative element v_i for each cluster by (9).

$$v_i = \left(\sum_{j}^{n} \mu_{ij}{}^2 y_j \right) \Big/ \left(\sum_{j}^{n} \mu_{ij}{}^2 \right), \quad (9)$$

where $1 \le i \le c$, μ_{ij} is the probability of the jth element assigned to w_i.

Step 3.3. Update the new partition matrix $U^{(1)}$ by Formula (10):

$$\mu_{ij}{}^{(1)} = \begin{cases} \dfrac{1}{\sum_{l=1}^{c} \left(d_{ij}/d_{lj}\right)^2} & \text{if } d_{ij} > 0, \\ 0 & \text{if } d_{ij} \le 0, \end{cases} \quad (10)$$

where d_{ij} is the distance from y_j to v_i and d_{lj} is the distance from y_l to v_i.

Step 3.4. Compute the $S = \max_{ij}\left(\left|\mu_{ij}^{(1)} - \mu_{ij}^{(0)}\right|\right)$.

Repeat *Step 3.2, Step 3.3* and *Step 3.4* until $S < \varepsilon$.In this algorithm, the Euclidian distance is also used. The end of this algorithm is a matrix of size $(c \times n)$. In this matrix, the sum of each column always equals 1 $(\sum_{j=1}^{c} \mu_{ij} = 1)$. If $\max\{\mu_{ij}\} = \mu_{im}$, $1 \le m \le c$ then the element y_i, $1 \le i \le n$ is assigned to w_m.

Step 4. Calculating the center m_i of each cluster, $i = 1, 2, \ldots, c$ and forecast Y_i according to the following rule:

$$Y_i = \sum_{j=1}^{c} \mu_{ij} c_j, \; i = 1, 2, \ldots, n. \quad (11)$$

Step 5. Forecasting X_i from the results of Y_i by (12):

$$X_i = Y_i \cdot \max\{X_i\}/10. \quad (12)$$

The proposed algorithm is illustrated in Fig. 1.

We have established a completely Matlab procedure to perform the proposed (IFTS) model. The calculation of the IFTS model can be performed conveniently and efficiently by this procedure. It is applied for numerical examples in Sects. 2.3 and 3.

2.3 The Convergence of the Proposed Algorithm

The convergence of the proposed algorithm is shown by the SNC algorithm (Step 2) and the DFR algorithm (Step 3). The DFR algorithm is improved from the fuzzy c-means clustering of time series data that its convergence was presented by [8, 26]. Therefore, to evaluate the convergence of the proposed algorithm, we consider the convergence of the SNC algorithm. It is presented by Theorem 1.

Theorem 1 If the function $f(u, v)$ in (7) satisfies:

(i) $f(u, v)$ depends only on $d(u, v)$, the distance from u to v.

(ii) $0 \le f(u, v) \le 1$ and $f(u, v) = 1$ only when $u = v$,

(iii) $f(u, v)$ is decreasing with respect to $d(u, v)$, then there exists t so that $z_i^{(t+1)}$ satisfies: $\max_i\{d(z_i^{(t)}, z_i^{(t+1)})\} < \varepsilon$.

Proof Let $C_1^{(t)}$ be the convex hull of $z^{(t)} = \left\{z_1^{(t)}, z_2^{(t)}, \ldots, z_n^{(t)}\right\}$, we have $z_j^{(t+1)}$ determined by (6) is a

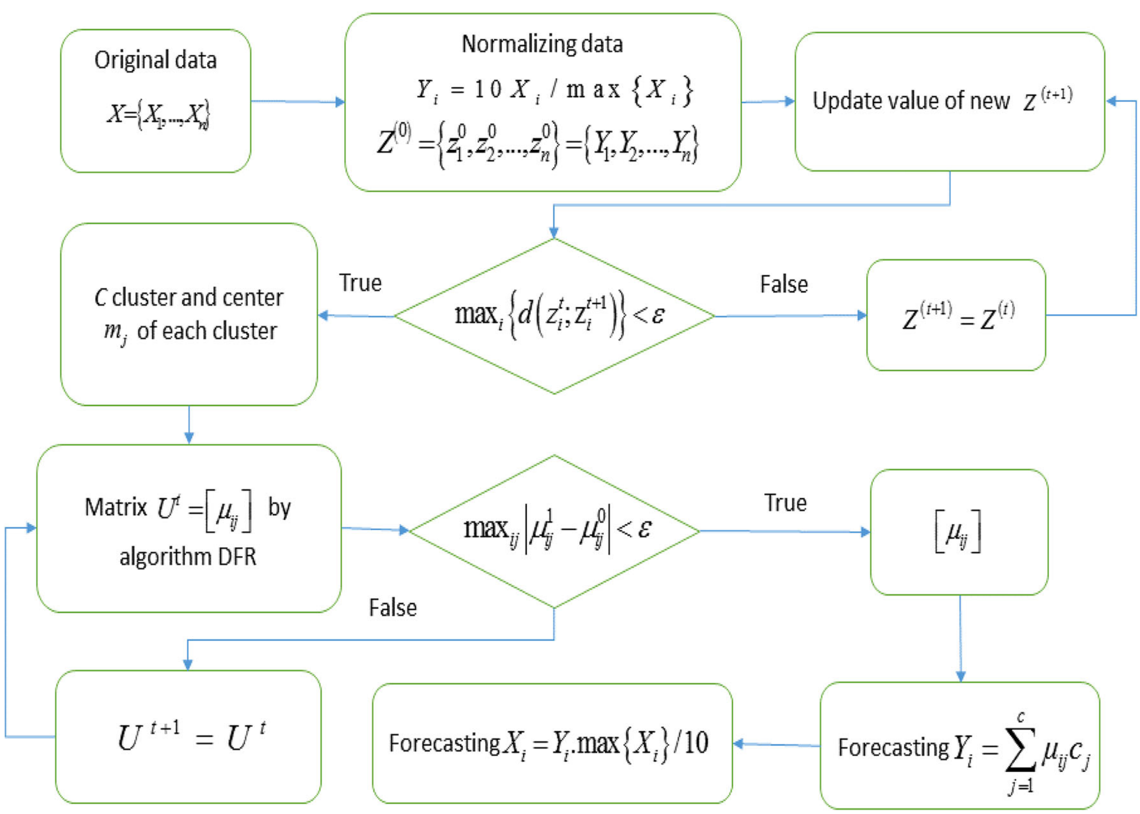

Fig. 1 Diagram for the proposed algorithm

weighted average of $z_j^{(t)}, j = 1, 2, \ldots, n$. Therefore, $z_j^{(t+1)} \in C_1^{(t)}$, that means:

$$C_1^{(0)} \supseteq C_1^{(1)} \supseteq \cdots \supseteq C_1^{(t)} \cdots$$

Let C_1 be the limit of $C_1^{(t)}, C_1 = \lim_{t \to \infty} C_1^{(t)}$. For each vertex $u_{1,i}$ of C_1, we prove that there exists at least one j, such that

$$\lim_{t \to \infty} z_j^{(t)} = u_{1,i}. \quad (13)$$

Since $\forall t$, $u_{1,i}^{(t)} = z_k^{(t)}$ for at least one k, there exists j, such that for infinite many t, $z_j^{(t)} = u_{1,i}^{(t)}$. Therefore, there exists $t_n \to \infty$, such that $z_j^{(t_n)} = u_{1,i}^{(t_n)}$ which leads to $\lim_{n \to \infty} z_j^{(t_n)} = u_{1,i}$. If $z_j^{(t_n)} = u_{1,i}$ except for any finite t, Eq. (13) is established. Otherwise, there exists $j' \neq j$ and $s_n \to \infty$, such that $z_{j'}^{(s_n)} = u_{1,i}^{(s_n)}$. Without loss of generosity, assume that $u_{1,i}^{(t)} = z_j^{(t)}$ or $z_{j'}^{(t)}, \forall t > T$. From Eq. (10), if $z_j^{(s)} = z_{j'}^{(s)}$ for some s, $z_j^{(t)} = z_{j'}^{(t)}, \forall t > s$. Therefore, for any $s > 0$ there exists $t > s$, such that $u_{1,i}^{(t)} = z_j^{(t)}$ and $u_{1,i}^{(t+1)} = z_{j'}^{(t+1)}$. Furthermore, we can choose s large enough, so that $C_1^{(s)}$ is close enough to C_1. Precisely, for any ε, there exists s, such that

$$\left| u_{1,k}^{(s)} - u_{1,k} \right| < \varepsilon, \forall k.$$

From the definition of f in (7), f is smaller than 1 unless the subjects are the same, which means each subject is most similar to itself. Since $z_{j'}^{(t+1)}$ is the weighted average of $z_k^{(t)}, z_{j'}^{(t)}$ cannot be too far from $u_{1,i}$, otherwise, $z_{j'}^{(t+1)}$ will not be at $u_{1,k}^{(t+1)}$, which is not inside the C_1. $u_{1,k}^{(t)}$ is also not inside the C_1, and is within ε to $u_{1,i}$. Therefore, $X_{j'}^{(t)}$ has to be within ε to $u_{1,i}$ that $z_{j'}^{(t+1)}$ can be at $u_{1,k}^{(t+1)}$. Since ε can be chosen arbitrary small, now we let ε small enough that all the projections, except $k = j, j'$, from $z_k^{(t)}$ to $\overrightarrow{z_{j'}^{(t)} z_j^{(t)}}$ fall into the negative side. This means that all other subjects are closer to $z_{j'}^{(t)}$ than $z_j^{(t)}$, and they have effects to pull both toward the convex hull. Since $z_{j'}^{(t)}$ is closer to other subjects, the values of f s are larger. Recall that

$$z_i^{(t+1)} = \frac{\sum_{j=1}^{n} f\left(z_i^{(t)}, z_j^{(t)}\right) \cdot z_j^{(t)}}{\sum_{j=1}^{n} f\left(z_i^{(t)}, z_j^{(t)}\right)},$$

$f(z_j^{(t)}, z_k^{(t)}) < f(z_{j'}^{(t)}, z_k^{(t)})$, for $k = j, j'$. Since $f(z_{j'}^{(t)}, z_j^{(t)}) < 1$, the effect from itself is larger than that from the other subject. This means that $z_{j'}^{(t+1)}$ is closer to $z_{j'}^{(t)}$ and that $z_{j'}^{(t+1)}$ is closer to $z_j^{(t)}$ if ignoring the effects from other subjects. Combining the fact that the effects from other subjects to pull $z_{j'}^{(t+1)}$ toward the convex hull are larger, $z_{j'}^{(t+1)}$ cannot replace $z_j^{(t+1)}$ as a new vertex. This contradicts to the assumption. Therefore, $u_{1,i}^{(t)} = z_j^{(t)}$ for some j and for all t large enough. Then,

$$\lim_{t \to \infty} z_j^{(t)} = \lim_{t \to \infty} u_{1,i}^{(t)} = u_{1,i}.$$

Let C_2 be the limit of $C_2^{(t)}$, $C_2 = \lim_{t \to \infty} C_2^{(t)}$, apply similar as $C_1^{(t)}$, we have at least one subject convergence to each vertex of C_2. Then, we can run similar steps again for $C_3, C_4, \ldots$ until all subjects convergence. It can be tested that the proposed function $f(u, v)$ in (7) satisfies: $f(u, v)$ depends only on $d(u, v), 0 \leq f(u, v) \leq 1$, $f(u, v) = 1$ only when $u = v$, and $f(u, v)$ is decreasing with respect to $d(u, v)$. Therefore, after the algorithm finishes, we have m elements $z_i^{(t)}, i = 1, 2, \ldots, m$ so that $\max_i \{d(z_i^{(t)}, z_i^{(t+1)})\} < \varepsilon$. □

In sum, the proposed algorithm converges for all time series. It means that this algorithm is controlled by the finite time. We know that the finite time control is more meaningful than infinite time control for nonlinear systems [32]. In our knowledge, this problem is not almost considered in the researches about the FTS models. Considering about time control is necessary to evaluate the effectivity of a FTS model, so we will further research about it in the next time.

3 Numerical Examples

3.1 Illustration for the Proposed Algorithm

In this section, we use the EnrollmentAU data presented in many studies such as [9, 10] to illustrate the steps of the proposed algorithm. This data set is often used to compare the effects of FTS models.

Step 1. From the given data set $\{X_i\}, i = 1, 2, \ldots, 22$, standardizing the data on the scale 10, we obtain the values Y_i in Table 1.

Step 2. Apply the SNC algorithm with different values of λ,, and we always obtain the convergence. Some cases for convergence of the SNC algorithm are shown in Fig. 2.

As presented in Step 2 of the proposed algorithm, performing with many time series, we choose $\lambda = 16$ for all numerical examples in this article. Then, after 18 iterations,

Table 1 EnrollmentAU data and its standardized data

Year	X_i	Y_i	Year	X_i	Y_i
1971	13,055	6.751	1982	15,433	7.981
1972	13,563	7.014	1983	15,497	8.014
1973	13,867	7.171	1984	15,145	7.832
1974	14,696	7.600	1985	15,163	7.841
1975	15,460	7.995	1986	15,984	8.266
1976	15,311	7.918	1987	16,859	8.719
1977	15,603	8.069	1988	18,150	9.386
1978	15,861	8.202	1989	18,970	9.810
1979	16,807	8.692	1990	19,328	9.995
1980	16,919	8.750	1991	19,337	10.00
1981	16,388	8.475	1992	18,876	9.762

the algorithm will converge to the following values: 7.0112 7.0113 7.0113 7.9804 7.9804 7.9804 7.9804 7.9804 8.6780 8.6780 8.6780 7.9804 7.9804 7.9804 7.9804 7.9804 8.6780 9.4139 9.8920 9.8920 9.8920 9.8920.

The result gives six representative elements, so we divide this series into 6 clusters.

Step 3. Using the DFR algorithm with 6 clusters, we have the specific clusters:

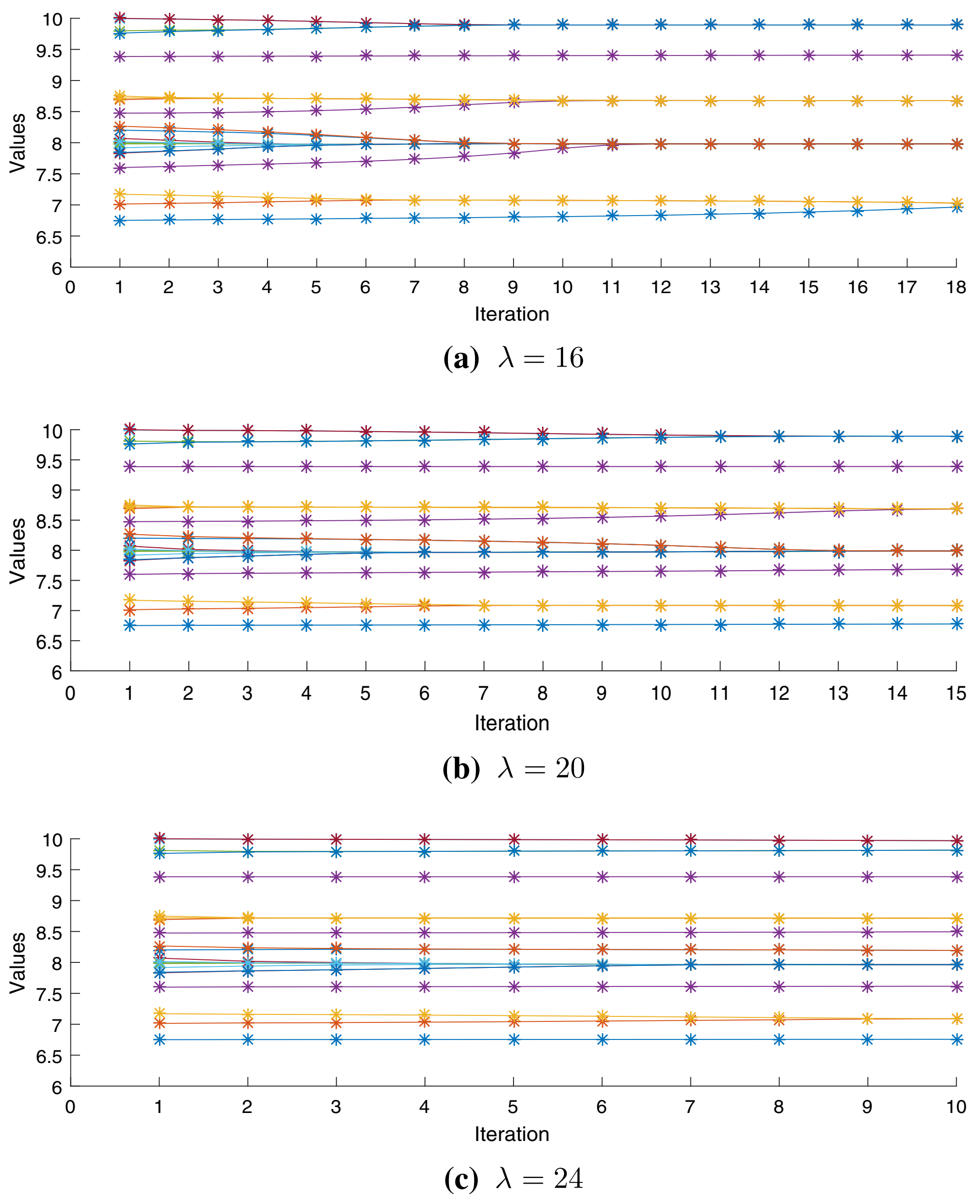

Fig. 2 The convergence of the SNC for some cases: **a** $\lambda = 16$, **b** $\lambda = 20$, and **c** $\lambda = 24$

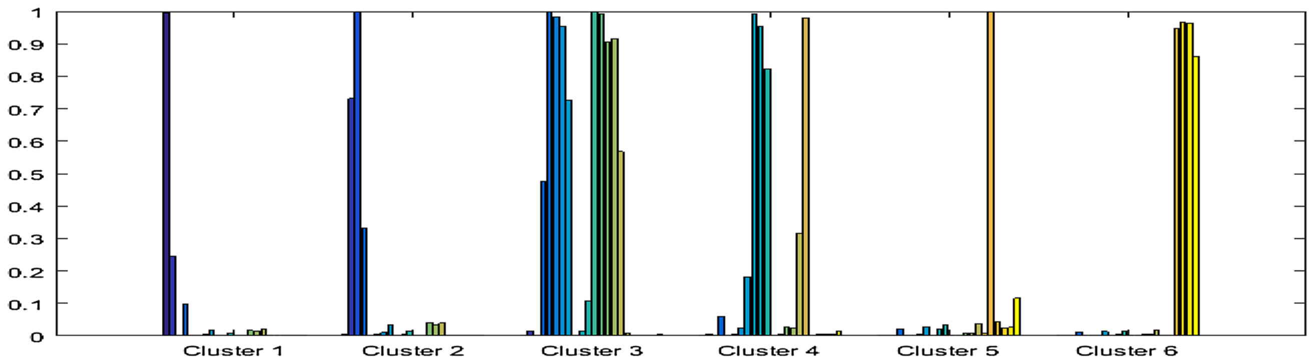

Fig. 3 The graph shows the relation between each element with 6 clusters

Table 2 The forecasted values for the EnrollmentAU data

Year	$\widehat{Y}_i$	$\widehat{X}_i$	Year	$\widehat{Y}_i$	$\widehat{X}_i$
1971	6.7535	13,059	1982	7.9718	15,415
1972	7.0407	13,615	1983	7.9733	15,418
1973	7.0928	13,715	1984	7.9542	15,381
1974	7.6540	14,801	1985	7.9579	15,388
1975	7.9721	15,416	1986	8.2170	15,889
1976	7.9717	15,415	1987	8.6595	16,745
1977	7.9853	15,441	1988	9.3860	18,150
1978	8.1069	15,676	1989	9.8536	19,054
1979	8.6588	16,743	1990	9.8635	19,073
1980	8.6630	16,752	1991	9.8608	19,068
1981	8.5829	16,597	1992	9.7949	18,940

$$w_1 = \{Y_1\},\ w_2 = \{Y_2;\ Y_3\} w_3 = Y_4;\ Y_5;\ Y_6;\ Y_7; Y_8;\ Y_{12}; Y_{13};\ Y_{14}; Y_{15};\ Y_{16}\}.$$
$$w_4 = \{Y_9;\ Y_{10}; Y_{11};\ Y_{17}\},\ w_5 = \{Y_{18}\},\ w_6 = \{Y_{19};\ Y_{20}; Y_{21};\ Y_{22}\}.$$

Calculating the center of each cluster, we obtain the results: 6.7510, 7.0925, 7.9718, 8.6590, 9.3860 and 9.8918.

The DFR algorithm also gives the relationships μ_{ij} from each element y_i to the cluster $w_j; i = 1, 2, \ldots, 22, j = 1, 2, \ldots, 6$ by the partition matrix:

$$[\mu_{ij}]_{6\times 22} = \begin{bmatrix} 0.9955 & 0.2446 & 0.0023 & \ldots & 0.0008 & 0.0018 \\ 0.0037 & 0.7310 & 0.9968 & \ldots & 0.0012 & 0.0024 \\ 0.0004 & 0.0151 & 0.0006 & \ldots & 0.0024 & 0.0051 \\ 0.0002 & 0.0052 & 0.0002 & \ldots & 0.0054 & 0.0132 \\ 0.0001 & 0.0024 & 0.0001 & \ldots & 0.0262 & 0.1156 \\ 0.0001 & 0.0017 & 0.0000 & \ldots & 0.9638 & 0.8619. \end{bmatrix}$$

These probabilities are shown in Fig. 3.

Step 4. Forecast for Y_i according to (11), we obtain $\widehat{Y}_i$ in Table 2.

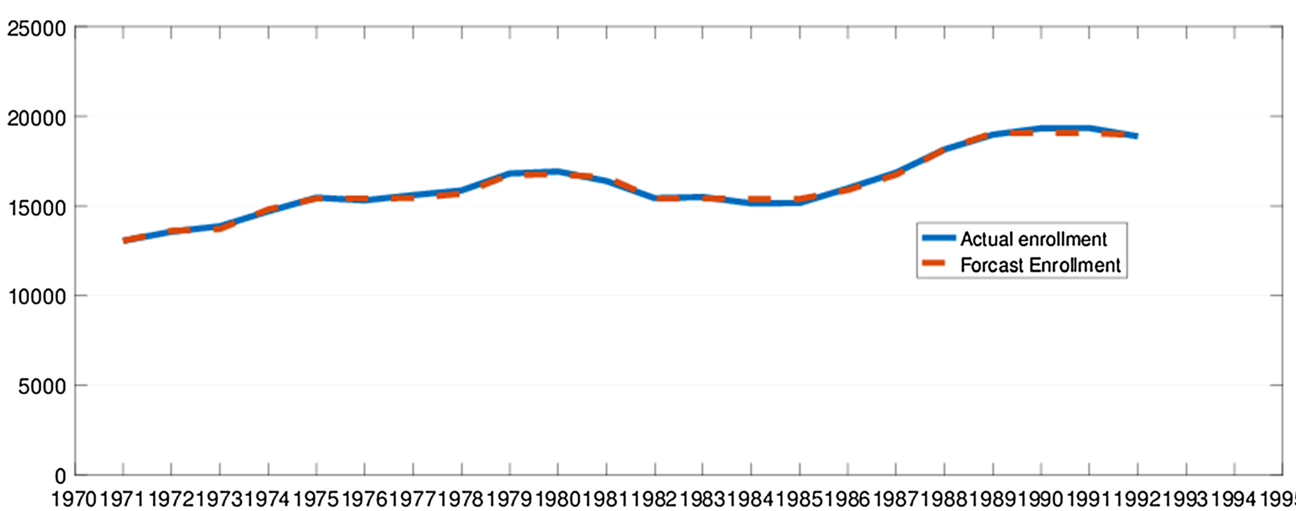

Fig. 4 The graph of actual and forecasted values of the EnrollmentAU data

Table 3 The parameters of the proposed algorithm and others

Data	Criteria	L-C	Hua	AM	Si	Gh
EnrollmentUA	MAE	296.15	299.15	479.57	254.16	298.68
	MAPE	2.69	2.45	2.87	1.53	1.82
	MSE	255,227	226,611	342,326	95,305	186,421
Taifex	MAE	38.27	96.71	89.30	46.01	71.10
	MAPE	0.89	1.39	1.32	0.70	1.03
	MSE	918.16	14,391	14,136	2968	937
Outpatient	MAE	76.23	96	181	119.03	56.18
	MAPE	11.54	13.75	22.50	2.12	1.98
	MSE	12,703	14,706	42,767	17,995.74	16,754.35
Foodgrain	MAE	47.76	58.64	89.60	8.69	8.17
	MAPE	6.47	4.53	5.81	5.43	4.98
	MSE	175.43	4772	10,672	104.25	123.45
Data	Criteria	C-H	Y-H	Tai	C-K	B-R
EnrollmentUA	MAE	293.45	216.50	168.84	314.34	285.28
	MAPE	1.76	2.15	1.02	2.17	1.65
	MSE	138,366.80	47,231.03	28,525.00	41,235	174,390.90
Taifex	MAE	11.36	21.32	11.40	25.71	9.27
	MAPE	0.17	1.42	0.17	1.03	0.16
	MSE	230.76	22,801	527.81	7679.0	94.65
Outpatient	MAE	107.40	138.38	159.80	167.15	249.17
	MAPE	1.89	2.17	24.45	2.74	3.06
	MSE	16,255.32	156.39	37,551.87	3890.76	165,755.00
Foodgrain	MAE	107.71	67.23	60.35	7.45	7.95
	MAPE	7.01	5.96	4.55	5.21	6.62
	MSE	183.56	2987.15	6460	2345.21	124.07
Data	Criteria	Chen	Yus	Egr	Kha	Proposed
EnrollmentUA	MAE	502.38	182.51	192.15	211.12	121.96
	MAPE	3.08	1.62	1.83	2.12	0.75
	MSE	413,980.98	31,752	34,280	31,021	21,292
Taifex	MAE	45.24	19.32	21.15	17.18	7.30
	MAPE	0.66	0.78	0.98	0.85	0.11
	MSE	4225.29	824.00	1012	921.15	85.68
Outpatient	MAE	325.96	96.34	86.28	49.98	43.74
	MAPE	5.82	1.34	1.45	1.09	0.76
	MSE	181,554.56	3421.24	3017.36	2908.48	2578.60
Foodgrain	MAE	16.18	109.15	6.98	5.98	4.99
	MAPE	10.13	7.57	5.09	4.87	3.93
	MSE	440.26	256.57	123.08	98.28	60.10

Step 5. From the results of $\widehat{Y}_i$ according to (12), the forecasted values for X_i are $\widehat{X}_i$ given in Table 2 and shown in Fig. 4. We also obtain the parameters MSE $=$ 21,292, MAE $=$ 121.96 and MAPE $=$ 0.75.

Figure 4 shows that the actual and forecasted values are almost identical.

3.2 Comparing Some Benchmark Data Sets

In this section, we use many series with different characteristics and numbers to compare the results of the proposed model with those of the models in [1] (AM), [21] (L-C), [18] (Hua), [6] (B-R), [27] (Si), [33] (Y-H), [17] (Gh), [10] (Chen), [7] (C-K), [9] (C-H), [20] (Kha), [34] (Yus), [15] (Egr), and Tai [30]. These are typical models, in which there are current works. The considered data sets are EnrollmentUA [10], Taifex (Taiwan Stock Exchange) [7],

Table 4 MAE, MAPE and MSE of four training sets

Model	Error	EnrollmentAU	Taifex	Outpatient	Foodgrain
AMR	MAE	482.28	90.41	463.12	9.46
	MAPE	3.04	1.31	7.86	6.68
	MSE	329,266.60	12,389.36	293,362.08	135.47
AMP	MAE	442.07	71.95	459.08	9.54
	MAPE	2.80	1.04	7.84	6.70
	MSE	391,803.40	9676.98	289,234.21	190.94
ARIMAR	MAE	423.21	39.20	385.49	7.06
	MAPE	2.68	0.57	6.42	5.31
	MSE	283,110.36	3373.71	218,896.50	70.67
ARIMAP	MAE	388.51	39.06	373.84	6.11
	MAPE	2.49	0.57	6.21	4.29
	MSE	226,972.97	3181.28	233,783.75	52.37

Table 5 MAE, MAPE and MSE of four test sets

Data	Model	MAE	MAPE	MSE
EnrollmentUA	ARIMAR	742.27	3.93	901,655.37
	AM	1785.28	9.39	3,326,909.30
	AMP	1089.69	5.74	1,376,307.00
	ARIMAP	739.16	3.92	731,600.93
Taifex	ARIMAR	79.61	1.17	7740.10
	AM	79.00	1.16	7117.50
	AMP	64.04	0.94	4581.28
	ARIMAP	67.85	1.00	5882.97
Outpatient	ARIMAR	335.57	7.09	195,066.15
	AM	930.75	19.52	1,303,655.09
	AMP	790.76	16.40	823,289.30
	ARIMAP	232.14	3.74	68,996.02
Foodgrain	ARIMAR	12.91	6.33	281.57
	AM	15.77	7.91	404.69
	AMP	13.16	6.64	299.98
	ARIMAP	12.28	5.97	251.23

Table 6 MAPE, MASE and E(SMAPE) for the M_3-competition data

Methods	MAPE	MASE	E(SMAPE)
ForecastPro	18.00	1.47	13.19
ForecastX	17.35	1.42	13.49
BJ automatic	19.13	0.54	14.01
Autobox1	18.23	1.51	14.41
Autobox2	20.36	1.69	15.23
Autobox3	19.31	1.57	15.33
ETS	17.38	1.43	13.13
AutoARIMA	18.92	1.46	13.59
Hybrid	17.59	1.40	12.82
Proposed model	6.77	1.00	10.76

Outpatient [6] and Foodgrain [17]. These well-known data are widely studied in the context of FTS models. If there is a new method that relates to FTS, then these data sets are often used to compare the performances between the new method and existing methods. In each data set, we will perform for two cases:

(i) All of the data are used to build the models and evaluate them according to the parameters MAE, MAPE and MSE.

(ii) Each data set is divided into two parts: Eighty percent of them are used as the training set to build the models, and about twenty percent of the remaining data is used as the validation set. For the training set, the ARIMA and Abbasov–Manedova (AM) models with original data (ARIMAR and AMR), ARIMA and AM models with fuzzy data of the proposed method (ARIMAP, AMP) are established. Using the established models from training set to forecast for the validation set.

- For (i): The results are presented in Table 3. Table 3 shows that the MAE, MAPE, and MSE of the proposed model are always smaller than the compared existing models for all data sets. This finding shows the stability and the advantages of the proposed model.
- For (ii):
  - With each training set, we perform the models AMR, AMP, ARIMAR and ARIMAP. Their results are given in Table 4.
  - Using the established models from training set (AMR, AMP, ARIMAR and ARIMAP) to forecast for the test set, we have Table 5.

Table 7 Flood peak of Tien River from 1990 to 2017

Year	Flood peak	Year	Flood peak	Year	Flood peak	Year	Flood peak
1990	418	1997	418	2004	440	2011	486
1991	463	1998	281	2005	436	2012	432
1992	343	1999	420	2006	417	2013	435
1993	344	2000	506	2007	408	2014	396
1994	453	2001	479	2008	377	2015	251
1995	430	2002	482	2009	412	2016	307
1996	486	2003	406	2010	320	2017	343

Table 8 The forecasted results for the flood peak of training set

Year	Actual	NFTS	Error (%)	Year	Actual	NFTS	Error (%)
1990	418	420.90	0.69	2001	479	483.01	0.83
1991	463	466.25	0.70	2002	482	483.25	0.26
1992	343	343.50	0.15	2003	406	417.65	2.86
1993	344	343.50	0.15	2004	440	457.98	3.86
1994	453	456.27	0.72	2005	436	450.78	3.39
1995	430	443.61	0.03	2006	417	420.65	0.88
1996	486	483.39	0.53	2007	408	419.22	0.87
1997	418	420.90	0.69	2008	377	377.00	0.00
1998	281	281.00	0.00	2009	412	420.47	2.05
1999	420	422.01	0.48	2010	320	320.00	0.00
2000	506	505.98	0.03	2011	486	483.39	0.53
MAE = 5.10; MAPE = 1.19; MSE = 59.99							

Table 9 Comparing the models of the test set for the flood peak

Year	Actual	ARIMAR	ARIMAP	AM	AMP
2012	432	520.51	463.69	500.00	494.63
2013	435	373.01	399.83	514.04	505.88
2014	396	423.48	440.51	528.05	517.12
2015	251	423.48	441.60	541.97	528.36
2016	307	423.48	431.10	556.05	539.60
2017	343	423.48	420.59	569.78	550.85
MAE		91.23	79.44	174.33	162.07
MAPE		28.63	25.81	55.08	51.37
MSE		10,370.37	8735.04	37,750.02	32,975.05

Tables 4 and 5 show that the proposed model has the best result in both interpolating and forecasting for all considered data sets. With a lot of considered models, this comparison is very meaningful to evaluate the advantages of the NFTS model.

3.3 Comparing M_3-Competition Data

To increase convincement about the effectivity of the proposed model, we use the M_3-competition data, which is a well-known benchmark data pool in the forecasting literature to perform. This data set was organized by Spyros and Michle [29]. Entrants had to forecast 3003 time series and the results were compared to a test set that was withheld from the participants. The 3003 series of the M_3-Competition were selected on a quota basis to include various types of time series data (micro, industry, macro, etc.) and different time intervals between successive observations (yearly, quarterly, etc.). In order to ensure that enough data were available to develop an adequate forecasting model, it was decided to have a minimum number of observations for each type of data. This minimum was set as 14 observations for yearly series (the median length of the 645 year series is 19 observations), 16 for quarterly (the median length of the 756 quarterly series is 44 observations), 48 for a monthly (the median length of the 1428 monthly series is 115 observations) and 60 for another series (the median length of the 174 other series is 63 observations). All the data (both training and test sets) and the forecasts of the original participants are publicly available in the Mcomp package for R software. The considered important models are ForecastPro, ForecastX, BJ automatic, Autobox1, Autobox2, Autobox3, Hybrid, ETS and AutoARIMA (see https://robjhyndman.com/m3comparisons.R). Using the proposed model and the above existing models, we perform for each series of the M_3-forecasting competition data. In the proposed model,

Table 10 Interpolating for all flood peak data

Year	Actual	NFTS	Error (%)	Year	Actual	NFTS	Error (%)
1990	418	413.68	1.03	2004	440	438.06	0.44
1991	463	463.00	0.00	2005	436	434.78	0.28
1992	343	343.34	0.10	2006	417	412.81	1.00
1993	344	343.34	0.19	2007	408	412.04	0.99
1994	453	453.01	0.01	2008	377	377.26	0.07
1995	430	433.84	0.89	2009	412	412.02	0.01
1996	486	482.80	0.66	2010	320	316.74	1.02
1997	418	413.68	1.03	2011	486	482.79	0.66
1998	281	281.00	0.00	2012	432	434.53	0.59
1999	420	416.42	0.85	2013	435	434.63	0.08
2000	506	506.00	0.00	2014	396	396.42	0.11
2001	479	480.44	0.30	2015	251	251.00	0.00
2002	482	483.01	0.21	2016	307	313.69	2.18
2003	406	411.26	1.30	2017	343	343.34	0.10
MAE = 2.02; MAPE = 0.50; MSE = 7.78							

Table 11 The forecasted flood peak through the year 2025

Year	ARIMAP	Year	ARIMAP
2018	403.88	2022	406.01
2019	407.70	2023	406.46
2020	405.47	2024	406.19
2021	406.77	2025	406.35

we use the established Matlab procedure to run and for existing models, and we use the results about MAPE, MASE and E(SMAPE) present in *Robs* blog (https://robjhyndman.com/hyndsight/show-me-the-evidence). This comparison is shown in Table 6.

Table 6 shows that the proposed model is advantageous than compared existing models. With the large numbers of the considered time series and the different features of the M_3-competition data set, this comparison is very meaningful to evaluate the advantages of the proposed model.

4 A Real Application in Vietnam

Mekong delta in Vietnam is strongly influenced by the Tien River. This river has brought the fertility of the soil, the abundance of fresh water and fisheries for this land. However, the complicated hydrological regime, especially floods, causes much damage to the resident every year. Flood forecasting for this river is an important issue of the region. In this section, we use the proposed model to forecast the flood peak at the main station located on Tien River.

In this application, based on the data presented in Table 7, we also consider two cases:

(i) *Case 1: Evaluating the Model* Divide the data into two parts: 80% for the training set (22 years) and

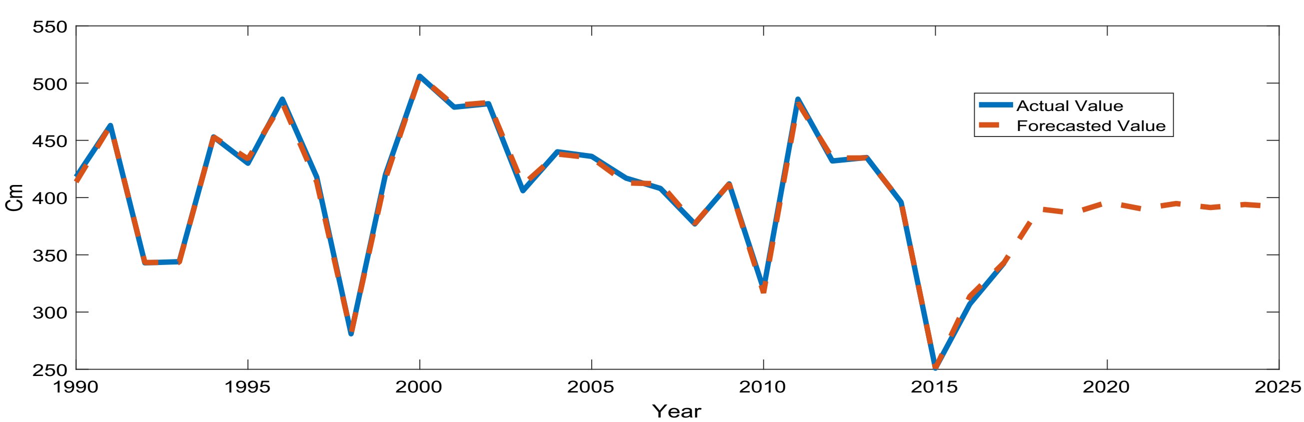

Fig. 5 Graph for actual and forecast flood peak

20% for the test set (6 years). Interpolate the training set by the proposed model and forecast for years of the test set by ARIMA and AM models and compare those with original data by the MAE, MAPE and MSE parameters. The results of performing are summarized in Table 8. Table 8 shows that the errors between the actual and the interpolated flood peak in training set are very low (0.00–3.86%) and the MAE, MAPE, and MSE parameters are small. Using this fuzzy data and original data to forecast by ARIMA and AM models for years of test set, we obtain Table 9. For test data, Table 9 also shows that the parameters MAE, MAPE and MSE of the proposed model are smallest.

(ii) *Case 2: Forecasting for Future* Interpolating all data by the proposed model, we have Table 10. Using the data from Table 10, forecasting for the next several years by ARIMA and AM methods, we obtain Table 11.
The results of interpolating and forecasting for the flood peak are shown in Fig. 5.

It is seen that the forecasted and the actual data are almost identical. In the future, the flood peak of the Tien River is slow.

5 Conclusion

This study has set up a new fuzzy time series model. This model is based on the two important algorithms: determining the suitable number of clusters for universe set and finding the fuzzy relationships between an element with clusters in series. These improvements make the proposed model more advantages than the existing models. The numerical examples from different data sets with various scales and characteristics show this problem. The proposed model is solved effectively by the established Matlab procedure. The practical application shows the logicality and potential to many different applications. Our further studies will focus on forecasting of many problems in reality.

References

1. Abbasov, A., Manedova, M.: Application of fuzzy time series to population forecasting. Vienna Univ. Technol. **12**, 545–552 (2003)
2. Abreu, P.H., Silva, D.C., Mendes-Moreira, J., Reis, L.P., Garganta, J.: Using multivariate adaptive regression splines in the construction of simulated soccer team's behavior models. Int. J. Comput. Intell. Syst. **6**(5), 893–910 (2013)
3. Aladag, S., Aladag, C.H., Mentes, T., Egrioglu, E.: A new seasonal fuzzy time series method based on the multiplicative neuron model and sarima. Hacet. J. Math. Stat. **41**(3), 337–345 (2012)
4. Aladag, C.H., Basaran, M.A., Egrioglu, E., Yolcu, U., Uslu, V.R.: Forecasting in high order fuzzy times series by using neural networks to define fuzzy relations. Expert Syst. Appl. **36**(3), 4228–4231 (2009)
5. Alpaslan, F., Cagcag, O., Aladag, C., Yolcu, U., Egrioglu, E.: A novel seasonal fuzzy time series method. Hacet. J. Math. Stat. **41**(3), 375–385 (2012)
6. Bindu, G., Rohit, G.: Enhanced accuracy of fuzzy time series model using ordered weighted aggregation. Appl. Soft Comput. **48**, 265–280 (2016)
7. Chen, S.M., Kao, P.Y.: TAIEX forecasting based on fuzzy time series, particle swarm optimization techniques and support vector machines. Inf. Sci. **247**, 62–71 (2013)
8. Chen, J.H., Hung, W.L.: An automatic clustering algorithm for probability density functions. J. Stat. Comput. Simul. **85**(15), 3047–3063 (2015)
9. Chen, S.M., Hsu, C.: A new method to forecast enrollments using fuzzy time series. Int. J. Appl. Sci. Eng. **2**, 3234–3244 (2004)
10. Chen, S.M.: Forecasting enrollments based on fuzzy time series. Fuzzy Sets Syst. **81**(3), 311–319 (1996)
11. Egrioglu, E., Aladag, C., Yolcu, U., Basaran, M., Uslu, V.: A new hybrid approach based on SARIMA and partial high order bivariate fuzzy time series forecasting model. Expert Syst. Appl. **36**(4), 7424–7434 (2009)
12. Egrioglu, E., Aladag, C., Yolcu, U., Uslu, V., Basaran, M.A.: A new approach based on artificial neural networks for high order multivariate fuzzy time series. Expert Syst. Appl. **36**(7), 10589–10594 (2009)
13. Egrioglu, E., Uslu, V., Yolcu, U., Basaran, M., Aladag, C.: A new approach based on artificial neural networks for high order bivariate fuzzy time series. Appl. Soft Comput. **36**(7), 265–273 (2009)
14. Eren, B., Vedide, R., Erol, E.: A modified genetic algorithm for forecasting fuzzy time series. Appl. Intell. **41**(2), 453–463 (2014)
15. Egrioglu, S., Bas, E., Aladag, C.H., Yolcu, U.: Probabilistic fuzzy time series method based on artificial neural network. Am. J. Intell. Syst. **62**, 42–47 (2016)
16. Friedman, J.H.: Multivariate adaptive regression splines. Ann. Stat. **19**(1), 1–67 (1991)
17. Ghosh, H., Chowdhury, S., Prajneshu, S.: An improved fuzzy time-series method of forecasting based on L-R fuzzy sets and its application. J. Appl. Stat. **43**(6), 1128–1139 (2015)
18. Huarng, K.: Heuristic models of fuzzy time series for forecasting. Fuzzy Sets Syst. **123**(3), 369–386 (2001)
19. Huarng, K., Yu, T.H.K.: Ratio-based lengths of intervals to improve fuzzy time series forecasting. IEEE Trans. Syst. Man Cybern. B (Cybernetics) **36**(2), 328–340 (2006)
20. Khashei, M., Bijari, M., Hejazi, C.S.R.: An extended fuzzy artificial neural networks model for time series forecasting. Iran. J. Fuzzy Syst. **3**, 45–66 (2011)
21. Lee, H.S., Chou, M.T.: Fuzzy forecasting based on fuzzy time series. Int. J. Comput. Math. **81**(7), 781–789 (2004)
22. Lewis, P.A., Stevens, J.G.: Nonlinear modeling of time series using multivariate adaptive regression splines (mars). J. Am. Stat. Assoc. **86**(416), 864–877 (1991)
23. Ming, C.S.: Forecasting enrollments based on high-order fuzzy time series. Fuzzy Sets Syst. **33**(1), 1–16 (2002)
24. Own, C.M., Yu, P.T.: Forecasting fuzzy time series on a heuristic high-order model. Cybern. Syst. Int. J. **62**(1), 1–8 (2005)
25. Qiang, S., Brad, C.: Forecasting enrollments with fuzzy time series—part II. Fuzzy Sets Syst. **62**(1), 1–8 (1994)

26. Richard, J.H., James, C.B.: Recent convergence results for the fuzzy c-means clustering algorithms. J. Classif. **5**, 237–247 (1998)
27. Singh, S.: A simple method of forecasting based on fuzzy time series. Appl. Math. Comput. **186**(1), 330–339 (2007)
28. Song, Q., Chissom, B.S.: Forecasting enrollments with fuzzy time series—part I. Fuzzy Sets Syst. **54**(3), 269–277 (1993)
29. Spyros, M., Michle, H.: The M_3—competition: results, conclusions and implications. Int. J. Forecast. **16**(4), 451–476 (2000)
30. Tai, V.V.: An improved fuzzy time series forecasting model using variations of data. Fuzzy Optim. Decis. Making (2018). https://doi.org/10.1007/s10700-018-9290-7
31. Teoh, H.J., Cheng, C.H., Chu, H.H., Chen, J.S.: Fuzzy time series model based on probabilistic approach and rough set rule induction for empirical research in stock markets. Data Knowl. Eng. **67**(1), 103–117 (2008)
32. Wang, F., Chen, B., Lin, C., Zhang, J., Meng, X.: Adaptive neural network finite-time output feedback control of quantized nonlinear systems. IEEE Trans. Cybern. **48**(6), 1839–1848 (2018)
33. Yu, H.K., Huarng, K.: A neural network- based fuzzy time series model to improve forecasting. Expert Syst. Appl. **37**, 3366–3372 (2010)
34. Yusuf, S.M., Mohammad, A., Hamisu, A.A.: A novel two-factor high order fuzzy time series with applications to temperature and futures exchange forecasting. Niger. J. Technol. **36**(4), 1124–1134 (2017)
35. Zhang, G.P.: Time series forecasting using a hybrid arima and neural network model. Neurocomputing **50**, 159–175 (2003)
36. Zhiqiang, Z., Qiong, Z.: Fuzzy time series forecasting based on k-means clustering. Open J. Appl. Sci. **25**(1), 100–105 (2012)

Tai Vovan received the bachelor's degree in education of mathematics and the master's degree in theory of probability and statistical mathematics from the Can Tho University in 1996 and 2003, respectively, and the Ph.D. degree in theory of probability and statistical mathematics in 2011 from the University of Science (Vietnam National University Ho Chi Minh City). He is currently a senior lecturer and vice dean of College of Natural Science, Can Tho University. His research interests include statistical pattern recognition (classification problem and cluster analysis) and fuzzy time series and their applications in data mining. He has authored many journal papers ISI about above subjects.

Nghiep Ledai received the B.Sc. degree in education of mathematics, the M.Sc. degree in applied mathematics from Can Tho University, Can Tho, Vietnam, and the M.Sc. degree in theory of probability and statistical mathematics from the Can Tho University in 2004 and 2016, respectively. He is currently a lecturer in Nam Can Tho University, Can Tho City, Vietnam. His research interests include unsupervised recognition and supervised recognition and fuzzy time series model.