Iranian Journal of Fuzzy Systems

Volume 19, Number 4, (2022), pp. 89-106

Original paper

University of Sistan and Baluchestan

# Building the forecasting model for time series based on the improvement of fuzzy relationships

T. Vo-Van[1], L. Nguyen-Huynh[2] and K. Nguyen-Huu[3]

[1,3] *College of Natural Science, Can Tho University, Can Tho City, Vietnam.*
[2] *Faculty of Mechanical - Electrical and Computer Engineering, School of Engineering and Technology, Van Lang University, Ho Chi Minh City, Vietnam*

vvtai@ctu.edu.vn, nguyenhuynhluan15@gmail.com, nhkhanh@ctu.edu.vn

**Abstract**

This study builds a new forecasting model for time series based on some important improvements. First, we choose the universal set to be the percentage variation of the series. This universal set is divided to clusters by the automatic algorithm. The suitable number of cluster depends on the similar level of elements in the universal set. Second, a principle to find the relationship of each element in the series to the found clusters is established. Finally, we propose the forecasting rule from the established fuzzy relationships. The proposed model is illustrated in detail by the numerical examples, and can be quickly applied to real data by the established Matlab procedure. Comparing many series with the differences about the number of elements, fields, and characteristics, the proposed model has shown the outstanding advantages. Using the proposed model, we forecast the salty peak for a coastal province in Vietnam to illustrate for application of this study.

*Keywords:* Cluster analysis, forecasting model, fuzzy relation, time series.

## 1 Introduction

Forecasting is the prediction of the results for the future based on historical data, knowledge, and experience of the related problems. It is the scientific basis for plans and development strategies in all areas. Therefore, the forecast always receives the attention of managers, mathematicians, and statisticians. However, it is still a problem that has not been solved fully yet [27, 28, 34]. In statistics, using time series and regression models are the common forecasting methods. When constructing a regression model, we must constrain the conditions for data that are difficult to satisfy in reality. As a result, the regression model does not often give the good forecasting result in reality [2, 3, 10, 17, 29].

Time series is a popular data type stored in many fields. As a result, it becomes an important research direction of statistics, and attracts many scientists. Non-fuzzy time series (NFTS) is a good method to forecast, and used very commonly today. In of them, ARIMA is the most popular model at present [27]. To have an effective NFTS model, the time series must be a stationary sequence, and the error must be white noise. Many series do not satisfy these conditions, so NTFS models are limited in many cases [5, 14, 28]. We know that NFTS models were built based on the association of data by mathematical expressions that are not language level. This is the main disadvantage of NFTS. Fuzzy time series (FTS) model has been proposed to solve this drawback.

The fuzzy time series model is developed in two directions: (i) Interpolating the original data to create the alignment of elements, then applying the known model to forecast for the future, and (ii) building the direct forecasting model for the future. For (i), there was a lot of great interest in many statisticians. Song and Chissom [22] were the pioneers with enrollment data of the University of Alabama (EnrollmentUA). Song and Chissom [23] used the triangular fuzzy relation for performing. Ming [6], Chen and Hsu [7] improved the model of [23] when building the new fuzzy relationship.

Corresponding Author: T. Vo-Van
Received: June 2021; Revised: May 2022; Accepted: June 2022.
https://doi.org/10.22111/IJFS.2022.7089

Huarng [14] and Own and Yu [19] presented a model for FTS using the heuristic knowledge to improve and to apply for EnrollmentUA data. Based on the neural network, the model of Aladag et al [4] gave the interesting results in some cases. From the fuzzy model by different linguistic levels, many scientists such as Ghosh et al [12], Lee and Chou [16], Singh [21] and Teoh et al [25] have proposed the new models. Tai and Nghiep [29] proposed a model based on the cluster analysis problem. This model continued to be improved in recent years by these two authors. Recently, combining the cluster analysis problem and the improved triangle fuzzy number, Tai and Thuy [28] proposed the new model for FTS. Winita et al. [24] proposed the model for two hybrid methods that can be used for forecasting time series. The first combines the singular spectrum analysis with the linear recurrent formula (SSA-LRF) and neural networks, while the second combines the SSA-LRF and weighted fuzzy time series. This model has shown the advantage in comparing the existing models in some cases. However, we see that the steps of this model are difficult to perform. When series have complex changes, this model does not often forecast well. Yanpeng et al. [33] proposed a novel forecasting model based on multiple linear regression and clustering algorithm for forecasting market prices. The model of [33] employed a reprocessing to transform the set of fuzzy high-order time series into a set of high-order time series, with synthetic minority oversampling technique. However, this model only optimizes for TAIEX series in comparing to other models. Tinh [26] presented a hybrid forecasting model combined particle swarm optimization and fuzzy C-means clustering. This model also shown the advantage in comparing to others for three data sets: Enrolment data of the University of Alabama, the Taiwan futures exchange, and the yearly deaths in car road accidents in Belgium. It is similar the model of [33], the model of [26] does not get good predictive results for the many data sets considered in reality. Recently, Dinh and Tai [20] have proposed the interpolating model for time series using the genetic algorithm. This model has shown the outstanding advantages in comparison with the existing models throughout many benchmark data sets. However, it is very complicated in terms of calculations, so it takes a lot of time to apply in practice.

It can be affirmed that the first research direction of FTS is of great interest to statisticians with many proposed models. In our opinion, there are two common weaknesses of the above models. The first weakness is the problem of dividing the universal set into intervals. The existing models often divide the universal set into intervals by specific numbers based on the Liker scale. It means that the number of divisions are chosen subjectively. A series with a simple or complex transformation can also be divided into intervals of the same number. This is not good for establishing the fuzzy relationships to build the model. The second disadvantage is that they can only interpolate historical data without forecasting for the future. Because the ultimate goal is to forecast, they have to resort to other models to perform. When building models to interpolate, we must have certain errors, and the forecasting models also have many limitations, so the practical application of the first research direction still faces a lot of challenges. Because information can be lost during the interpolating stage, the first research direction usually only gives good results if the series only has an increase or decrease in the future. When future data have a large fluctuation, the forecasting result is very limited.

According to our research, comparing with the first direction, FTS models developed in the second direction have not had much attention. In this direction, two typical models are AM [1] and IFTS [27]. Abbasov and Manedova [1] (AM) proposed a fuzzy time series model to forecast population for a country. This model is based on the new fuzzy relationship between each element of the series and the divided groups. In this model, many parameters have not been investigated to obtain the optimization in specific cases. Therefore, it was only suitable for the proposed data set but not for others. To improve the AM model, Tai [27] proposed the IFTS model. In this model, the parameters of the AM model have been investigated and given by the specific algorithms which were suitable for each set of data. This model also improved the universal set, and the fuzzy relationship to forecast. IFTS has shown effectiveness when it was compared to many benchmark data sets. However, as the conclusion of this article, IFTS still had limitations when data had complex changes. According to our knowledge, the development of FTS in the second direction is still a challenging problem at present.

This article studies for the second direction. That is, it builds a forecasting model directly from the data with some improvements to overcome the drawbacks from the above models. For example, it has the contributions as follows:

(i) Propose the universal set to be the percentage change of two consecutive time points. This is more suitable than original data in determining the similarity of elements in series. The study also considers dividing the elements of series into intervals with the appropriate number depending on the similarity level of elements in series. The specific elements in each interval are determined by an improved non-fuzzy clustering technique.

(ii) Based on the established clusters from (i), we develop the expression for the fuzzy relationship of elements. This expression is established from the improvement of the known triangle relationship.

(iii) Develop the new rule for forecasting. The result of forecasting for the future is based on previous actual data, and the changes determined throughout the established fuzzy relationships from (ii).

The proposed model is shown in detail by the numerical examples, and performed rapidly by the established Matlab procedure. Matlab procedure automatically performs the steps of the proposed algorithm to give the final results. To

evaluate the effectiveness of the proposed model, we compared it to the existing models. Comparisons are made with many well-known data sets, including M3-competition data with 3003 series of many fields. All of them show that the proposed model has outstanding advantages. In addition, the study has applied effectively to real data sets.

The next sections of the paper are structured as follows. Section 2 presents some concepts related to fuzzy time series, and the proposed algorithm. An example illustrating step by step for the proposed algorithm is presented in Section 3. This section also gives two real applications. Section 4 gives the comparison of the proposed algorithm with the existing algorithms and some discussions. The last section is the conclusion.

# 2 Definitions and the proposed model

This section gives the basis concepts about fuzzy time series introduced by Song and Chissom [22]. We also present the proposed model in detail in this section.

## 2.1 Definitions

**Definition 2.1.** *Let $U$ be universe of discourse, $U = \{u_1, u_2, ..., u_n\}$. A fuzzy set $A$ of $U$ is defined as follows:*

$$A = f_A(u_1)/u_1 + f_A(u_2)/u_2 + ... + f_A(u_b)/u_b,$$

*where $f_A$ is the membership function of the fuzzy set $A$, $f_A : U \to [0, 1]$. If $u_a$ is a generic element of fuzzy set $A$, then $f_A(u_a)$ is the degree of belongingness of $u_a$ to $A$; $f_A(u_i) \in [0, 1]$, and $1 \le a \le b$.*

**Definition 2.2.** *A fuzzy number $A = (a, b, c)$ is called triangular fuzzy number if its membership function is given by*

$$f(x) = \begin{cases} 0 & \text{if} \quad x < a, \\ \frac{x-a}{b-a} & \text{if} \quad a \le x \le b, \\ \frac{c-x}{c-b} & \text{if} \quad b \le x \le c, \\ 0 & \text{if} \quad x > c. \end{cases}$$

**Definition 2.3.** *Let $X(t)(t = 0, 1, 2, ...)$ be a subset of real numbers, be the universe of discourse by which fuzzy sets $f_j(t)$ are defined. If $F(t)$ is a collection of $f_1(t), f_2(t), ...$, then $F(t)$ is called a fuzzy time series defined on $X(t)$.*

**Definition 2.4.** *Let $F(t), t = 1, 2, ...$ be a fuzzy time series. Assume that there is a relationship $R(t-1, t)$ between $F(t)$ and $F(t-1)$ that satisfies $F(t) = F(t-1) * R(t-1, t)$ where $F(t)$ and $F(t-1)$ are fuzzy sets and $*$ is max-min composition operator; then $R(t-1, t)$ is a fuzzy logic relationship. To sum up, let $F(t-1) = A_i$ and $F(t) = A_j$. The fuzzy logical relationship between $F(t)$ and $F(t-1)$ can be denoted as $A_i \to A_j$, where $A_i$ refers to the left-hand side, and $A_j$ refers to the right-hand side of the fuzzy logical relationship. Furthermore, these fuzzy logical relationships can be grouped to establish different fuzzy relationship.*

## 2.2 The proposed model

Given a time series $X = \left\{X_t; t = \overline{1, n}\right\}$, then the proposed model has the following steps:

**Step 1.** Calculate the percentage variation for the consecutive periods of time to get the new series $Y = \left\{Y_i; i = \overline{1, n-1}\right\}$, with

$$Y_i = \frac{X_{i+1} - X_i}{X_i} \cdot 100, i = \overline{1, n-1}.$$

**Step 2.** Divide series $Y$ into the clusters with the appropriate number depending on the proximity of the elements in this series by the DSNC algorithm (the determining the suitable number of clusters) as follows:

Step 2.1. Initialize $t = 0$, $V^{(0)} = \left\{v_1^{(0)}, v_2^{(0)}, ..., v_{n-1}^{(0)}\right\} = \{Y_1, Y_2, ..., Y_{n-1}\}$ and $\varepsilon = 0.0001$.

Step 2.2. Update the sequence $V^{(t)}$ according to following rule:

$$v_i^{(t+1)} = \frac{\sum\limits_{j=1}^{n-1} f\left(v_i^{(t)}, v_j^{(t)}\right).v_j^{(t)}}{\sum\limits_{j=1}^{n-1} f\left(v_i^{(t)}, v_j^{(t)}\right)}, i = \overline{1, n-1}, \tag{1}$$

where

$$f\left(v_i^{(t)}, v_j^{(t)}\right) = \begin{cases} e^{-|v_i^{(t)} - v_j^{(t)}|/\lambda} & \text{if} \quad |v_i^{(t)} - v_j^{(t)}| \le d_s, \\ 0 & \text{if} \quad |v_i^{(t)} - v_j^{(t)}| > d_s, \end{cases} \tag{2}$$

with $d_s = \frac{1}{n(n-1)} \sum\limits_{i<j} |v_i^{(t)} - v_j^{(t)}|, \lambda = d_s/m$, and $m$ is constant.

Step 2.3. Repeat Step 2.2 until $\max\limits_i |v_i^{(t)} - v_i^{(t+1)}| < \varepsilon$.

The value of $\lambda$ in (2) measures the variance of the elements in clusters. The larger $\lambda$ is the larger the standard deviation of each established cluster is taken. Then, the number of clusters for the universal set is otherwise. Because $d_s$ is constant, $\lambda$ depends on $m$. Performing for a lot of series, we choose $m = 48$ in this study.

$v_i^{(t+1)}$ calculated by (1) is expansion or narrowing of $Y$ so that the elements of $Y$ will be changed to become the centroid clusters. When Step 2 ends, if there are $k$ different elements in $V^{(t)}$ then we divide $Y$ into $k$ clusters $(w_1, w_2, ..., w_k)$, and continue Step 3.

**Step 3.** Find the elements for clusters by the FEC algorithm. This algorithm has the following steps:

Step 3.1. Initialize the elements for $k$ clusters according to the following rule: The elements $Y_i$ corresponding to the same $v_i^{(t)}$ in the sequence $V^{(t)}$ are arranged in a cluster. Then, calculate the initial centers of the clusters by (3):

$$c_j = \frac{1}{n_j} \sum_{h=1}^{n_j} Y_h, \tag{3}$$

where $n_j$ is the number of elements in the $j^{th}$ cluster, $j = \overline{1,k}$.

Step 3.2. Calculate the distance from each element of the universal set $Y_i$ to the center of clusters $c_j$ :

$$d_{ij} = |Y_i - c_j|, i = \overline{1, n-1}, j = \overline{1,k}.$$

If there exists $Y_h \in w_l$ ($w_l$ is the $l^{th}$ cluster) so that $d_{hl} > d_{hm}$ (where $m = \overline{1,k}$ and $m \neq l$), we assign $Y_h$ to $w_m$. Then, we recalculate the centers of the two new clusters.

Step 3.3. Repeat Step 3.2 until it no longer exists $Y_h$ and $w_l$ so that $d_{hl} > d_{hm}$ with $m = \overline{1,k}$ and $m \neq l$.

We find that it is quite simple to identify the elements in each cluster of Step 3. An element $Y_i$ is classified into clusters $w_j$ if $d_{ij}$, the distance from $Y_i$ to the center of clusters $c_j$ is the minimum value.

**Step 4.** Determine universe of discourse for $Y$ by $U$:

$$U = \left[c_1 - \frac{c_2 - c_1}{2}; c_k + \frac{c_k - c_{k-1}}{2}\right].$$

Divide $U$ into $k$ intervals as follows:

$$\begin{aligned} U_1 &= \left[c_1 - \tfrac{c_2 - c_1}{2}; \tfrac{c_1 + c_2}{2}\right), \\ U_j &= \left[\tfrac{c_{j-1} + c_j}{2}; \tfrac{c_j + c_{j+1}}{2}\right), 2 \le j \le k-1, \\ U_k &= \left[\tfrac{c_{k-1} + c_k}{2}; c_k + \tfrac{c_k - c_{k-1}}{2}\right]. \end{aligned}$$

**Step 5.** Find the fuzzy relationship for each interval $U_j \left(j = \overline{1,k}\right)$ by the following rule:

$$T_j = \begin{cases} \frac{1.5}{\frac{1}{|a_1|} + \frac{0.5}{|a_2|}} \cdot \Delta_1 & \text{if} \quad j = 1, \\ \frac{2}{\frac{0.5}{|a_{j-1}|} + \frac{1}{|a_j|} + \frac{0.5}{|a_{j+1}|}} \cdot \Delta_j & \text{if} \quad j = \overline{2, k-1}, \\ \frac{1.5}{\frac{0.5}{|a_{k-1}|} + \frac{1}{|a_k|}} \cdot \Delta_k & \text{if} \quad j = k, \end{cases}$$

where $a_j$ is the middle point of $U_j$ and

$$\Delta_j = \begin{cases} 1 & \text{if} \quad a_j > 0. \\ 0 & \text{if} \quad a_j = 0. \\ -1 & \text{if} \quad a_j < 0. \end{cases}$$

From the triangular fuzzy number defined in Section 2.1, the authors in [34] have proposed a relationship to build the time series model. The relationship proposed by us in this step is improvement from the studying of [34].

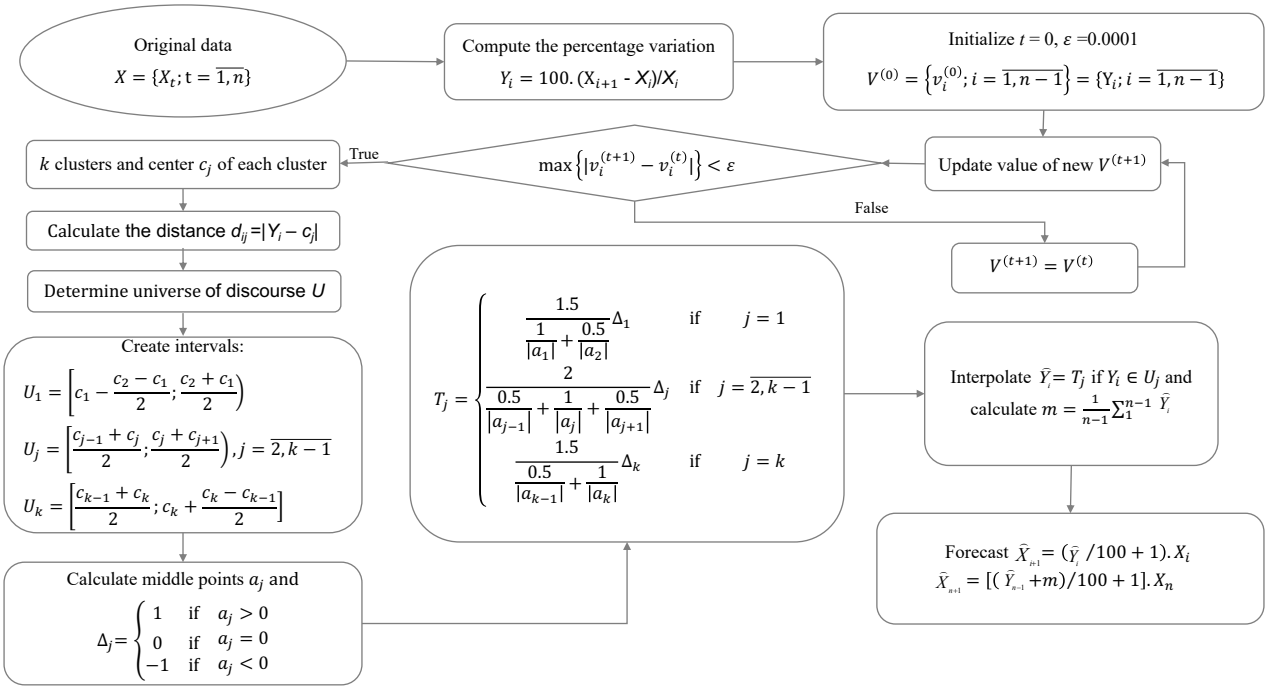



Figure 1: Diagram of the proposed algorithm.

**Step 6.** Interpolate the elements of $Y_i$ by the rule: If $Y_i \in U_j$, then $\hat{Y}_i = T_j$. After that, calculate the average $m = \frac{1}{n-1}\sum_{i=1}^{n-1} \hat{Y}_i$.

**Step 7.** Interpolate for $X$ and forecast for the next year respectively according to the following rules::

$$\hat{X}_{i+1} = \left(\frac{\hat{Y}_i}{100} + 1\right) \cdot X_i, i = \overline{1,n-1}.$$
$$\hat{X}_{n+1} = \left(\frac{\hat{Y}_{n-1} + m}{100} + 1\right) \cdot X_n.$$

The proposed model is illustrated in Figure 1.

## 2.3 Convergence of the proposed algorithm

The convergence of the proposed algorithm is shown by Step 2 (DSNC) and Step 3 (FEC). FEC is improved from the k-means clustering algorithm of time series that its convergence was present in [13]. Therefore, to evaluate the convergence of the proposed algorithm, we only consider the convergence of the DSNC algorithm. It is shown by the following theorem: If the function $f(u,v)$ in (2) satisfies:

(i) $f(u,v)$ depends only on $|u-v|$.

(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 $|u-v|$,

then there exists $t$ so that $v_i^{(t+1)}$ determined by (1) satisfies:

$$\max_i\{|v_i^{(t+1)} - v_i^{(t)}|\} < \varepsilon.$$

First of all, given the convex hull $C(V)$ for a set of points $V$ in a vector space $X$ is the minimal convex set containing $V$. Let $C_1^{(t)}$ be the convex hull of $\left\{v_1^{(t)}, ..., v_n^{(t)}\right\}$. Then $v_i^{(t+1)} \in C_1^{(t)}$ is a weighted average of $v_j^{(t)}, j = 1, ..., n$. Therefore, $C_1^{(t)} \supseteq C\left(\left\{v_1^{(t+1)}, ..., v_n^{(t+1)}\right\}\right) = C_1^{(t+1)}$. Since

$$C_1 = \lim_{t\to\infty} C_1^{(t)},$$

there exists $i$ such that $\lim_{t\to\infty} u_i^{(t)} = u_i$, where $u_i^{(t)} = \left\{u_1^{(t)}, ..., u_n^{(t)}\right\}$ is a vertex of $C_1^{(t)}$. For each $t$ and $i$, $u_i^{(t)} = v_k^{(t)}$ for at least one $k$; there exists $j$ such that $v_j^{(t)} = u_i^{(t)}$ for infinite many ts. Therefore, there exists $t \to \infty$ such that $v_j^{(t_n)} = u_i^{(t_n)}$, which leads to $\lim_{n\to\infty} v_j^{(t_n)} = u_i$. We consider two possible cases as follows:

Case 1: If $v_j^{(t_n)} = u_i$ except for any finite $t$ then $\lim_{t\to\infty} v_j^{(t)} = u_i$.

Case 2: If there exists $j' \neq j$ and $s_n \to \infty$ such that $\forall n$, $v_{j'}^{(s_n)} = u_j^{(s_n)}$. Assume that $u_i^{(t)} = v_j^{(t)}$ or $v_{j'}^{(t)}$ for all $t > T$. From (2), if $v_j^{(s)} = v_{j'}^{(s)}$ for some $s$, $v_j^{(t)} = v_{j'}^{(t)}$ for all $t > s$. Therefore, for any $s > 0$, there exists $t > s$ such that $u_i^{(t)} = v_j^{(t)}$ and $u_i^{(t+1)} = v_{j'}^{(t+1)}$. We claim that this case; however; can never happen with $t$ is large enough.

Without loss of generality, it is assumed that $u_i = 0, v_j^{(t)} \leq 0$, and $v_k^{(t)} > 0$ for $k \neq j$ or $k \neq j'$. If $v_{j'}^{(t+1)}$ later becomes the new vertex. Then,

$$v_{j'}^{(t+1)} < v_j^{(t+1)}. \tag{4}$$

Moreover, since $v_{j'}^{(t+1)}$ is the new vertex, we have

$$v_{j'}^{(t+1)} \leq 0 \Rightarrow \sum_{k=1}^{N} f(v_{j'}^{(t)}, v_k^{(t)}) v_k^{(t)} \leq 0.$$

Since $v_j^{(t)}$ is the current vertex; $|v_j^{(t)} - v_k^{(t)}| > |v_{j'}^{(t)} - v_k^{(t)}|$ for all $k$. Then,

$$\sum_{k=1}^{N} f(v_j^{(t)}, v_k^{(t)}) v_k^{(t)} \leq \sum_{k=1}^{N} f(v_{j'}^{(t)}, v_k^{(t)}) v_k^{(t)} < 0,$$

and

$$0 < \sum_{k=1}^{N} f(v_j^{(t)}, v_k^{(t)}) \leq \sum_{k=1}^{N} f(v_{j'}^{(t)}, v_k^{(t)}),$$

we have

$$v_{j'}^{(t+1)} < v_j^{(t+1)},$$

which is a contradiction to the assumption. Therefore, $u_i^{(t)} = v_j^{(t)}$ for some $j$ and for all $t$ large enough. Then,

$$\lim_{t\to\infty} v_j^{(t)} = u_i.$$

We can apply a similar result for $C_2$ as $C_1^{(t)}$; at least one subject convergence to each vertex of $C_2$. Then, we can run similar steps again for $C_3, C_4, ...$ until all subjects convergence. This completes the proof of theorem. Because the function $f(.)$ defined by (2) is satisfied all conditions of Theorem, the algorithm DSNC convergences.

# 3 Numerical examples and real applications

For detail about computing the steps of the proposed model, we take two data sets to illustrate. We also present the proposed forecasting model for two real series in this section.

## 3.1 Numerical examples

In this section, we use the NYSE and Lahi data sets [29] to illustrate the steps of the proposed algorithm.

(*a*) *NYSE data set:*

The dataset is given by Column $X$ of Table 1.

**Step 1.** Calculating the percentage variation for the first year to the second year, we have:

$$Y_1 = \frac{X_2 - X_1}{X_1} \cdot 100 = \frac{10582 - 10649}{10649} \cdot 100 = -0.6261.$$

Performing the next years similarly, we obtain Column $Y$ of Table 1.

**Step 2.** Applying the DSNC algorithm for $Y$, after 23 iterations we obtain the following results:

Table 1: Values of series $X$ and $Y$

| No | $X$ | $Y$ | No | $X$ | $Y$ |
|---|---|---|---|---|---|
| 1 | 10649 | | 19 | 10989 | −0.1750 |
| 2 | 10582 | −0.6261 | 20 | 10948 | −0.3730 |
| 3 | 10609 | 0.2572 | 21 | 10878 | −0.6404 |
| 4 | 10640 | 0.2882 | 22 | 10998 | 1.1085 |
| 5 | 10704 | 0.6094 | 23 | 11039 | 0.3698 |
| 6 | 10719 | 0.1342 | 24 | 11286 | 2.2386 |
| 7 | 10617 | −0.9472 | 25 | 11225 | −0.5420 |
| 8 | 10598 | −0.1844 | 26 | 11107 | −1.0544 |
| 9 | 10595 | −0.0230 | 27 | 11068 | −0.3469 |
| 10 | 10769 | 1.6424 | 28 | 11135 | 0.6024 |
| 11 | 10761 | −0.0796 | 29 | 11225 | 0.8075 |
| 12 | 10932 | 1.5945 | 30 | 11193 | −0.2861 |
| 13 | 10883 | −0.4504 | 31 | 11160 | −0.2897 |
| 14 | 11036 | 1.4023 | 32 | 11225 | 0.5838 |
| 15 | 11004 | −0.2882 | 33 | 11245 | 0.1743 |
| 16 | 11090 | 0.7782 | 34 | 11257 | 0.1115 |
| 17 | 10996 | −0.8467 | 35 | 11144 | −1.0052 |
| 18 | 11008 | 0.1140 | | | |

| | | | | | |
|---|---|---|---|---|---|
| −0.6106 | 0.2924 | 0.2924 | 0.5993 | 0.1291 | −1.0060 |
| −0.1857 | −0.0552 | 1.6184 | −0.0552 | 1.6184 | −0.4409 |
| 1.4024 | −0.3096 | 0.7926 | −0.8514 | 0.1291 | −0.1857 |
| −0.3096 | −0.6106 | 1.1085 | 0.2924 | 2.2386 | −0.6106 |
| −1.0060 | −0.3096 | 0.5993 | 0.7926 | −0.3096 | −0.3096 |
| 0.5993 | 0.1291 | 0.1291 | −1.0060 | | |

The convergence of Step 2 is illustrated by Figure 2.

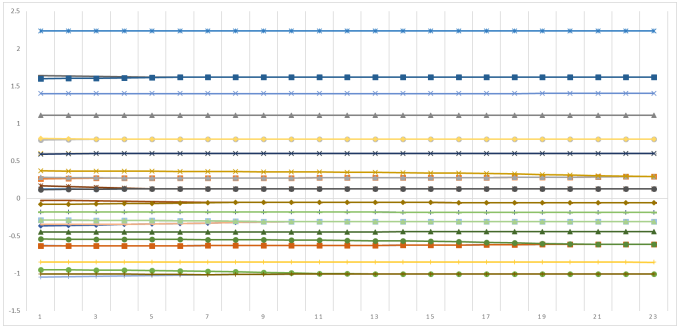


Figure 2: Illustration of convergence of DNCS algorithm

Since at the end of Step 2, we get 17 different elements, the series $Y$ is divided into 17 clusters.

**Step 3.** The result of FEC is given in Table 2. Based on the obtained result, we can determine specific elements in 17 clusters and their centers.

**Step 4.** Based on the center of the clusters, we determine the universe of discourse $U = [-1.0801; 2.5487]$ and divide it into 17 intervals as follows:

$U_1 = [-1.0801; -0.9245)$ ; $U_2 = [-0.9245; -0.7400)$ ;$U_3 = [-0.74; -0.5876)$ ; $U_4 = [-0.5876; -0.4962)$ ;
$U_5 = [-0.4962; -0.3836)$ ; $U_6 = [-0.3836; -0.2482)$ ;$U_7 = [-0.2482; -0.1155)$ ; $U_8 = [-0.1155; 0.0411)$ ;
$U_9 = [0.0411; 0.2031)$ ; $U_{10} = [0.2031; 0.3213)$ ;$U_{11} = [0.3213; 0.4841)$ ; $U_{12} = [0.4841; 0.6957)$ ;
$U_{13} = [0.6957; 0.9507)$ ; $U_{14} = [0.9507; 1.2554)$ ;$U_{15} = [1.2554; 1.5104)$ ; $U_{16} = [1.5104; 1.9285)$ ;
$U_{17} = [1.9285; 2.5487]$ .

Table 2: Elements and centers of each cluster of the NYSE dataset

| Cluster | Elements | Center |
|---|---|---|
| 1 | $\{Y_1; Y_{25}; Y_{34}\}$ | $-1.0023$ |
| 2 | $\{Y_{16}\}$ | $-0.8467$ |
| 3 | $\{Y_1; Y_{20}\}$ | $-0.6332$ |
| 4 | $\{Y_{24}\}$ | $-0.5420$ |
| 5 | $\{Y_{12}\}$ | $-0.4504$ |
| 6 | $\{Y_{14}; Y_{19}; Y_{26}; Y_{29}; Y_{30}\}$ | $-0.3168$ |
| 7 | $\{Y_7; Y_{18}\}$ | $-0.1797$ |
| 8 | $\{Y_8; Y_{10}\}$ | $-0.0513$ |
| 9 | $\{Y_5; Y_{17}; Y_{32}; Y_{33}\}$ | 0.1335 |
| 10 | $\{Y_2; Y_3\}$ | 0.2727 |
| 11 | $\{Y_{22}\}$ | 0.3698 |
| 12 | $\{Y_4; Y_{27}; Y_{31}\}$ | 0.5985 |
| 13 | $\{Y_{15}; Y_{28}\}$ | 0.7928 |
| 14 | $\{Y_{21}\}$ | 1.1085 |
| 15 | $\{Y_{13}\}$ | 1.4023 |
| 16 | $\{Y_9; Y_{11}\}$ | 1.6185 |
| 17 | $\{Y_{23}\}$ | 2.2386 |

**Steps 5.** Calculating the middle points $a_i$ of intervals $U_i$ corresponding to the coefficients $\Delta_i$, and the fuzzy relationships $T_i$.

For example, $a_1, a_2, \Delta_1$ and $\Delta_2$ are determined as follows:

$a_1 = (-1.0801 + (-0.9245))/2 = -1.0023$, $a_2 = (-0.9245 + (-0.74))/2 = -0.8322$, $\Delta_1 = -1$, $\Delta_2 = -1$ because $a_1, a_2 < 0$.

Then,

$$T_1 = \frac{1.5}{\frac{1}{|-0.1.0023|} + \frac{0.5}{|-0.8322|}} \cdot (-1) = -0.9384.$$

Calculating the next values similarly, we obtain Table 3.

Table 3: The values $a_i$, $\Delta_i$ and $T_i$

| Interval | $a_j$ | $\Delta_j$ | $T_j$ | Interval | $a_j$ | $\Delta_j$ | $T_j$ |
|---|---|---|---|---|---|---|---|
| 1 | $-1.0023$ | $-1$ | $-0.9384$ | 10 | 0.2622 | 1 | 0.2186 |
| 2 | $-0.8322$ | $-1$ | $-0.8151$ | 11 | 0.4027 | 1 | 0.3818 |
| 3 | $-0.6638$ | $-1$ | $-0.6601$ | 12 | 0.5899 | 1 | 0.5643 |
| 4 | $-0.5419$ | $-1$ | $-0.5354$ | 13 | 0.8232 | 1 | 0.795 |
| 5 | $-0.4399$ | $-1$ | $-0.4185$ | 14 | 1.1030 | 1 | 1.0663 |
| 6 | $-0.3159$ | $-1$ | $-0.2836$ | 15 | 1.3829 | 1 | 1.3631 |
| 7 | $-0.1819$ | $-1$ | $-0.0974$ | 16 | 1.7194 | 1 | 1.7145 |
| 8 | $-0.0372$ | $-1$ | $-0.0593$ | 17 | 2.2386 | 1 | 2.0339 |
| 9 | 0.1221 | 1 | 0.0850 | | | | |

**Step 6.** Interpolate the values of $Y$.

For instance, the elements $Y_1, Y_{25}$ and $Y_{34}$ are in the interval $U_1$, $\widehat{Y}_1 = \widehat{Y}_{25} = \widehat{Y}_{34} = T_1 = -1.0023$.

In a similar manner, we obtain Table 4 (Column $\widehat{Y}$). Then, we calculate the average $m = \frac{1}{34}\sum_{i=1}^{34} \widehat{Y}_i = 0.2597$.

**Step 7.** Interpolate the values of $X$, we obtain column $\widehat{X}$ of Table 4.

The performance from Table 4 is illustrated by Figure 3.

(*b*) *Lahi data set*

The data is given by Column $X$ of Table 5.

Table 4: Original and interpolated values of NYSE data set

| No | $X$ | $\widehat{Y}$ | $\widehat{X}$ | No | $X$ | $\widehat{Y}$ | $\widehat{X}$ |
|---|---|---|---|---|---|---|---|
| 1 | 10649 | – | – | 18 | 11008 | 0.0850 | 1.1005 |
| 2 | 10582 | −0.6601 | 1.0578 | 19 | 10989 | −0.0974 | 1.0997 |
| 3 | 10609 | 0.2186 | 1.0605 | 20 | 10948 | −0.2836 | 1.0958 |
| 4 | 10640 | 0.2186 | 1.0632 | 21 | 10878 | −0.6601 | 1.0876 |
| 5 | 10704 | 0.5643 | 1.0700 | 22 | 10998 | 1.0663 | 1.0994 |
| 6 | 10719 | 0.0850 | 1.0714 | 23 | 11039 | 0.3818 | 1.1040 |
| 7 | 10617 | −0.9384 | 1.0618 | 24 | 11286 | 2.0339 | 1.1264 |
| 8 | 10598 | −0.0974 | 1.0607 | 25 | 11225 | −0.5354 | 1.1226 |
| 9 | 10595 | −0.0593 | 1.0591 | 26 | 11107 | −0.9384 | 1.1120 |
| 10 | 10769 | 1.7145 | 1.0777 | 27 | 11068 | −0.2836 | 1.1075 |
| 11 | 10761 | −0.0593 | 1.0763 | 28 | 11135 | 0.5643 | 1.1131 |
| 12 | 10932 | 1.7145 | 1.0945 | 29 | 11225 | 0.7950 | 1.1223 |
| 13 | 10883 | −0.4185 | 1.0887 | 30 | 11193 | −0.2836 | 1.1193 |
| 14 | 11036 | 1.3631 | 1.1031 | 31 | 11160 | −0.2836 | 1.1161 |
| 15 | 11004 | −0.2836 | 1.1004 | 32 | 11225 | 0.5643 | 1.1223 |
| 16 | 11090 | 0.7950 | 1.1091 | 33 | 11245 | 0.0850 | 1.1235 |
| 17 | 10996 | −0.8151 | 1.0999 | 34 | 11257 | 0.0850 | 1.1254 |

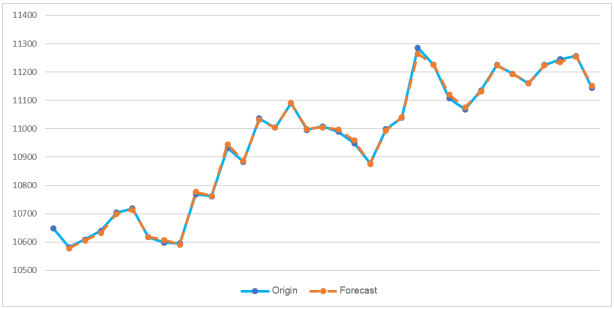



Figure 3: Actual and interpolated values from the proposed model for NYSE data set

**Step 1.** Calculating the percentage variation for the first year to the second, we have:

$$Y_1 = \frac{X_2 - X_1}{X_1} \cdot 100 = \frac{512 - 1025}{1025} \cdot 100 = -50.05.$$

Performing the next years similarly, we obtain Column $Y$ of Table 5.

**Step 2.** Applying the DSNC algorithm for $Y$, after 659 iterations we obtain following results:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| −49.1056 | 96.2891 | −15.4409 | −49.1056 | 17.0736 | 54.9454 | −36.0254 | 70.966 |
| 17.0736 | −15.4409 | −15.4409 | 54.9454 | −15.4409 | 17.0736 | −36.0254 | −15.4409 |
| 17.0736 | 54.9454 | −15.4409 | −15.4409 | 33.2054 | −15.4409 | | |

The result of 659 iterations is given by Figure 4.

The convergence of Step 2 is illustrated by Figure 4. Since at the end of Step 2, we get 9 different elements, the series $Y$ is divided into 9 clusters.

**Step 3.** The result algorithm FEC is given in Table 6. Based on the obtained result, we can determine specific elements in 9 clusters and their centers.

Table 5: Values of series $X$ and $Y$ for Lahi data set

| No | $X$ | $Y$ | No | $X$ | $Y$ |
|---|---|---|---|---|---|
| 1 | 1025 | – | 13 | 994 | 56.54 |
| 2 | 512 | −50.05 | 14 | 759 | −23.64 |
| 3 | 1005 | 96.29 | 15 | 883 | 16.34 |
| 4 | 852 | −15.22 | 16 | 599 | −32.16 |
| 5 | 440 | −48.36 | 17 | 499 | −16.69 |
| 6 | 502 | 14.09 | 18 | 590 | 18.24 |
| 7 | 775 | 54.38 | 19 | 911 | 54.41 |
| 8 | 465 | −40.00 | 20 | 862 | −5.38 |
| 9 | 795 | 70.97 | 21 | 801 | −7.08 |
| 10 | 970 | 22.01 | 22 | 1067 | 33.21 |
| 11 | 742 | −23.51 | 23 | 917 | −14.06 |
| 12 | 635 | −14.42 | 24 | | |

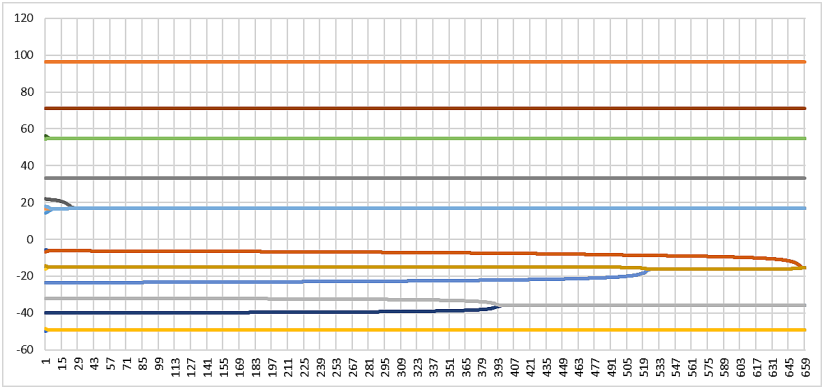



Figure 4: Illustration of convergence of DNCS algorithm for Lahi data set

Table 6: Elements and centers of each cluster of the Lahi data set

| Cluster | Elements | Center |
|---|---|---|
| 1 | $\{Y_1; Y_4\}$ | -49.2028 |
| 2 | $\{Y_7; Y_{15}\}$ | -36.0815 |
| 3 | $\{Y_3; Y_{10}; Y_{11}; Y_{13}; Y_{16}; Y_{22}\}$ | -17.924 |
| 4 | $\{Y_{19}; Y_{20}\}$ | -6.2276 |
| 5 | $\{Y_5; Y_9; Y_{14}; Y_{17}\}$ | 17.6693 |
| 6 | $\{Y_{21}\}$ | 33.2085 |
| 7 | $\{Y_6; Y_{12}; Y_{18}\}$ | 55.1082 |
| 8 | $\{Y_8\}$ | 70.9677 |
| 9 | $\{Y_2\}$ | 96.2891 |

**Step 4.** Based on the center of the clusters, we determine the universe of discourse $U = [-55.76; 108.95]$, and divide it into 9 intervals as follows:

$$U_1 = [-55.7634; -42.6422)\,; \quad U_2 = [-42.6422; -27.0028)\,; \quad U_3 = [-27.0028; -12.0758)\,;$$
$$U_4 = [-12.0758; 5.7208)\,; \quad U_5 = [5.7208; 25.4389)\,; \quad U_6 = [25.4389; 44.1584)\,;$$
$$U_7 = [44.1584; 63.0380)\,; \quad U_8 = [63.0380; 83.6284)\,; \quad U_9 = [83.6284; 108.9497]\,.$$

**Step 5.** Calculating the middle points $a_i$ of intervals $U_i$ corresponding to the coefficients $\Delta_i$, and the fuzzy relationships $T_i$.

For example, $a_1, a_2, \Delta_1$ and $\Delta_2$ are determined as follows:

$a_1 = (-55.7634 + (-42.6422))/2 = -49.2028$, $a_2 = (-42.6422 + (-27.0028))/2 = -34.8225$, $\Delta_1 = -1$, $\Delta_2 = -1$ because $a_1, a_2 < 0$.

Then,

$$T_1 = \frac{1.5}{\frac{1}{|-49.2028|} + \frac{0.5}{|-34.8225|}} \cdot (-1) \approx -43.25.$$

Calculating the next values similarly, we obtain Table 7.

Table 7: The values $a_i$, $\Delta_i$ and $T_i$ for Lahi data set

| Interval | $a_j$ | $\Delta_j$ | $T_j$ |
|---|---|---|---|
| 1 | −49.2028 | −1 | −43.2494 |
| 2 | −34.8225 | −1 | −31.0229 |
| 3 | −19.5393 | −1 | −8.9729 |
| 4 | −3.1775 | −1 | −5.3706 |
| 5 | 15.5799 | 1 | 8.4778 |
| 6 | 34.7986 | 1 | 28.5070 |
| 7 | 53.5982 | 1 | 50.1959 |
| 8 | 73.3332 | 1 | 71.0284 |
| 9 | 96.2891 | 1 | 87.1911 |

**Step 6.** Interpolate the values of $Y$.
For instance, the elements $Y_1$ and $Y_4$ are in the interval $U_1$, $\widehat{Y}_1 = \widehat{Y}_4 = T_1 = -43.25$.

In a similar manner, we obtain Table 8 (Column $\widehat{Y}$). Then, we calculate the average $m = \frac{1}{21}\sum_{i=1}^{21} \widehat{Y}_i \approx 17.42$.

**Step 7.** Interpolate the values of $X$, we obtain column $\widehat{X}$ of Table 8.

Table 8: Original and interpolated values of Lahi data set

| No | $X$ | $\widehat{Y}$ | $\widehat{X}$ | No | $X$ | $\widehat{Y}$ | $\widehat{X}$ |
|---|---|---|---|---|---|---|---|
| 1 | 1025 | – | – | 1 | 994 | 50.20 | 954 |
| 2 | 512 | −43.25 | 582 | 2 | 759 | −8.97 | 905 |
| 3 | 1005 | 87.19 | 958 | 3 | 883 | 8.48 | 823 |
| 4 | 852 | −8.97 | 915 | 4 | 599 | −31.02 | 609 |
| 5 | 440 | −43.25 | 484 | 5 | 499 | −8.97 | 545 |
| 6 | 502 | 8.48 | 477 | 6 | 590 | 8.48 | 541 |
| 7 | 775 | 50.20 | 754 | 7 | 911 | 50.20 | 886 |
| 8 | 465 | −31.02 | 535 | 8 | 862 | −5.37 | 862 |
| 9 | 795 | 71.03 | 795 | 9 | 801 | −5.37 | 816 |
| 10 | 970 | 8.48 | 862 | 10 | 1067 | 28.51 | 1029 |
| 11 | 742 | −8.97 | 883 | 11 | 917 | −8.97 | 971 |
| 12 | 635 | −8.97 | 675 | 12 | | | |

The performance from Table 8 is illustrated by Figure 5.

## 3.2 Real applications

Vietnam is one of the countries heavily affected by climate change. In the impacts of climate change in Vietnam, salty intrusion in the coastal provinces is considered the severest. It has caused much damage to the farming and livestock in recent years. To deal with this problem, an urgent requirement from the managers for scientists are to predict the level of salty intrusion in the future. The forecasting result will be an important scientific basis to take the suitable countermeasures, and to reduce the impacts to a minimum. Although many efforts have been made, this problem is still unresolved so far.

In this section, we apply the proposed model to forecast the salty peak at two stations on two main rivers in Tra Vinh province, a coastal province in Vietnam. There are Tra Vinh 1 and Tra Vinh 2 stations. The data sets are presented in Column "Actual" in Table 9.

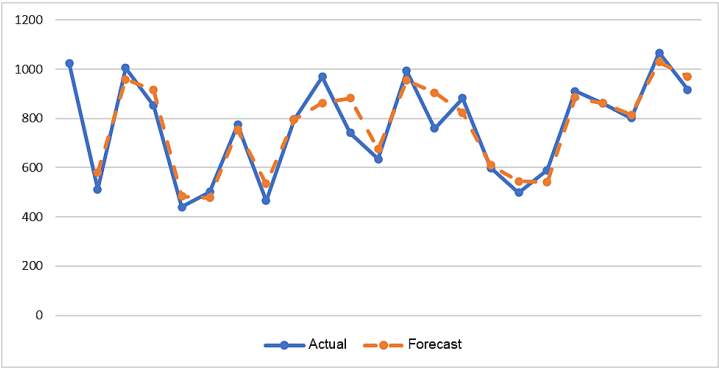



Figure 5: Actual and interpolated values from the proposed model for Lahi data set

Table 9: Salty peak at two stations and interpolated values

| Year | Tra Vinh 1 | | Tra Vinh 2 | |
|---|---|---|---|---|
| | Actual | Forecast | Actual | Forecast |
| 2002 | 6.3 | - | 7.9 | - |
| 2003 | 7.9 | 7.6 | 11.3 | 11.2 |
| 2004 | 10.6 | 10.5 | 8.3 | 8.3 |
| 2005 | 10.9 | 10.9 | 10.7 | 10.5 |
| 2006 | 9.7 | 10.3 | 9.0 | 9.7 |
| 2007 | 8.9 | 9.2 | 9.5 | 9.4 |
| 2008 | 10.0 | 9.4 | 9.9 | 9.9 |
| 2009 | 6.3 | 7.6 | 9.9 | 9.4 |
| 2010 | 11.8 | 11.1 | 10.8 | 10.7 |
| 2011 | 8.3 | 8.9 | 11.1 | 11.3 |
| 2012 | 4.4 | 5.0 | 9.1 | 10.1 |
| 2013 | 9.2 | 8.7 | 12.4 | 12.3 |
| 2014 | 5.9 | 7.0 | 6.0 | 7.2 |
| 2015 | 8.5 | 8.8 | 8.9 | 8.8 |
| 2016 | 10.4 | 10.2 | 10.7 | 10.4 |
| 2017 | 11.5 | 11.0 | 11.7 | 11.6 |
| 2018 | 10.5 | 11.2 | 11.4 | 11.1 |
| 2019 | 11.5 | 12.0 | 11.9 | 11.9 |

Before forecasting by the proposed model, we also compare the interpolating and forecasting performance of the models. The considered models are ARIMA, AM and IFTS (the others are not used for forecasting purpose) throughout the following steps:

(i) Divide the original data into two parts, training set and test set with ratio be 80% (2002 - 2015) and 20% (2016 - 2019), respectively.

(ii) For the training set, the interpolating performance of the comparative methods is considered and presented in the first half of Table 10.

(iii) For the test set, the forecasting performance of the comparative methods is considered and presented in the second half of Table 10.

Table 10 shows that the proposed model gives the good result for both interpolating and forecasting process. It is the best result in comparison with the considered models.

(iv) Using the proposed model with all data, we forecast for the next 4 years. The result of this implementation is presented in Table 11.

The original values and the forecasted values for the salt peak is shown by Figure 6.

Figure 6 shows a significant increase of salty peaks in the near future. In general, the salty peak of two stations

Table 10: The results of the training set and test set for salty peak

| Data set | Model | Tra Vinh 1 | | | Tra Vinh 2 | | |
|---|---|---|---|---|---|---|---|
| | | MAE | MAPE | MSE | MAE | MAPE | MSE |
| Training set | Proposed | 0.70 | 8.97 | 0.69 | 0.50 | 5.58 | 0.41 |
| | ARIMAR | 2.91 | 38.54 | 14.78 | 1.02 | 10.75 | 1.55 |
| | AM | 4.36 | 62.29 | 19.10 | 3.13 | 35.62 | 12.61 |
| | IFTS | 4.36 | 62.29 | 19.10 | 3.13 | 35.62 | 12.61 |
| Test set | Proposed | 1.80 | 17.91 | 3.30 | 0.70 | 6.22 | 0.74 |
| | ARIMAR | 2.83 | 26.31 | 10.74 | 1.57 | 13.60 | 2.69 |
| | AM | 2.63 | 25.63 | 7.30 | 6.27 | 54.36 | 43.63 |
| | IFTS | 2.63 | 25.63 | 7.30 | 6.27 | 54.36 | 43.63 |

Table 11: The results of forecasting salty peak at Cau Quan station

| Year | 2020 | 2021 | 2022 | 2023 |
|---|---|---|---|---|
| Tra Vinh 1 | 10.8 | 12.1 | 13.5 | 14.2 |
| Tra Vinh 2 | 13.0 | 14.1 | 15.2 | 15.6 |

are increase in the next four years. According to the actual data, we obtain in 2020 and 2021, these forecasts are quite appropriate when the difference of actual and forecasting values of the years are below 5%.

# 4 Comparison and discussion

## 4.1 Comparisons

Given a series of real data $\{X_i\}$ and forecasted values $\{\widehat{X}_i\}$, $\left(i = \overline{1,n}\right)$, respectively. Then, we have the popular parameters to evaluate the built FTS models as follows:

*Mean squared error:*

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{X}_i - X_i\right)^2. \tag{5}$$

*Mean absolute error:*

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\widehat{X}_i - X_i\right|. \tag{6}$$

*Mean absolute percentage error:*

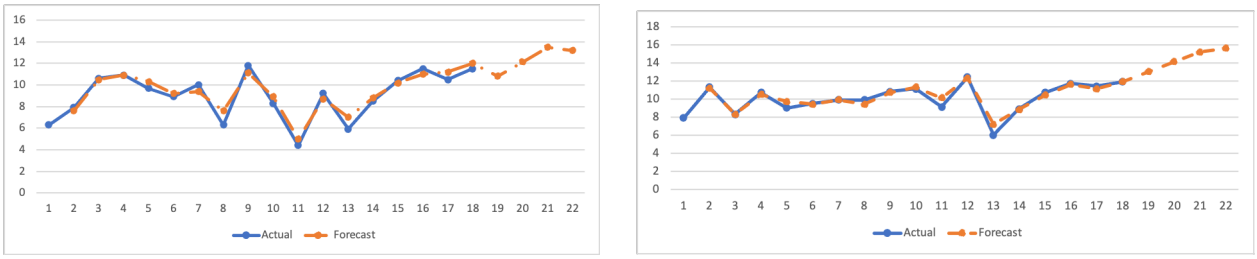
$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left|\widehat{X}_i - X_i\right|}{X_i}\cdot 100\right). \tag{7}$$

*Symmetric mean absolute percentage error:*

$$SMAPE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left|X_i - \widehat{X_i}\right|}{\left(X_i + \widehat{X_i}\right)\Big/2}\cdot 100\right). \tag{8}$$

*Mean absolute scaled error:*

$$MASE = \frac{\sum_{i=1}^{n}\left|X_i - \widehat{X_i}\right|}{\frac{n}{n-1}\sum_{i=2}^{n}|X_i - X_{i-1}|}. \tag{9}$$




(a) Tra Vinh 1 (b) Tra Vinh 2

Figure 6: Actual and forecasting values from the proposed model for salty peak data sets

For the built models, the smaller these parameters are, the better the models are.

To evaluate the effectiveness of the proposed model, we compare it with many existing forecasting models through a lot of series. The first comparison is made for a very famous dataset. It is the M3-Competition data set with 3003 series. In this comparison, all data of each series is taken to perform. The second comparison was done on the Lahi, NYSE and Taifex data sets [27]. These are also very common data sets in studying time series. When a new time series model is proposed, these data sets are often used as benchmark data for comparison.

(i) The third competition data set called the M3-Competition is intended to both replicate, and extend the features of the M1-Competition and M2- Competition, through the inclusion of more methods and researchers and more series [18]. There are 3003 series in M3-Competition set including yearly series, quarterly series, monthly series, daily series, and others. It has the following domains: micro, industry, macro, finance, demographics, and others. Table 12 introduces the general characters about M3-Competition data set.

Table 12: Introduce about M3-Competition data

| Interval | Micro | Industry | Macro | Finance | Demog | Other | Total |
|---|---|---|---|---|---|---|---|
| Yearly | 146 | 102 | 83 | 58 | 245 | 11 | 645 |
| Quarterly | 204 | 83 | 336 | 76 | 57 | 0 | 756 |
| Monthly | 474 | 334 | 312 | 145 | 111 | 52 | 1428 |
| Others | 4 | 0 | 0 | 29 | 0 | 141 | 174 |
| Total | 828 | 519 | 731 | 308 | 413 | 204 | 3003 |

For the M3-Competition data, according to [18], the most suitable models are ForecastPro, ForecastX, Bj automatic, Autobox1, Autobox2, Autobox3, Hybrid, ETS and AutoARIMA. Therefore, we will compare the proposed model with these models. In addition, we also compare the proposed model with models published in recent years such as AM [1], IFTS [27], and Tai and Nghiep [29], Tai and Nghiep [30], Tai and Thuy [28], Yanpeng et al. [33], Dinh and Tai [20]. In this case, all series are taken to compare the efficiency of the interpolation process from models, and the result is present in Table 13.

Table 13 shows the outstanding advantages of the proposed model in comparison to others with all parameters.

(ii) For these data, we divide each of the Lahi, NYSE and Taifex data sets into training and test set with rate 80% and 20%, respectively. The training set is used to build models and compare their interpolating efficiency, the test set is used to compare the forecasting effect of the built models from training set. In this case, because the aim is to compare the efficiency of both the interpolation and prediction processes, only direct predictive models are considered. The result on training set is presented in Table 14.

Table 14 shows that the parameters MAE, MAPE and MSE of the proposed model are smaller than the considered forecasting models.

Using the models established from the training set to forecast for the test set, and comparing them with real data, we obtain Table 15.

The obtained result in Table 15 is similar to Table 14, when the proposed model again achieves the best result.

## 4.2 Discussion

The M3 data set has 3003 series with different characteristics, so whichever model has advantages over others will be appreciated. The proposed model has overcome the popular models with the above data set, so it is really meaningful in our opinion. Beside the popular models such as Theta, ForecastPro, ForecastX, Bj automatic, Autobox1, Autobox2,

Table 13: Comparison of models on the M3-Competition data set

| Models | E(SMAPE) | MAPE | MASE |
|---|---|---|---|
| Theta | 13.01 | 17.42 | 1.39 |
| ForecastPro | 13.19 | 18.00 | 1.47 |
| ForecastX | 13.49 | 17.35 | 1.42 |
| BJ automatic | 14.01 | 19.13 | 1.54 |
| Autobox2 | 14.41 | 18.23 | 1.51 |
| Autobox1 | 15.23 | 20.36 | 1.69 |
| Autobox3 | 15.33 | 19.31 | 1.57 |
| ETS | 13.13 | 17.38 | 1.43 |
| AutoARIMA | 13.59 | 18.92 | 1.46 |
| Combined ETS | 12.82 | 17.59 | 1.40 |
| AM [1] | 11.77 | 15.76 | 1.32 |
| IFTS [27] | 12.76 | 17.31 | 1.36 |
| Tai and Nghiep [29] | 10.76 | 6.77 | 1.00 |
| Tai and Nghiep [30] | 9.89 | 6.53 | 1.12 |
| Tai and Thuy [28] | 11.37 | 10.98 | 1.09 |
| Yanpeng et al. [33] | 11.15 | 10.67 | 1.02 |
| Dinh and Tai [20] | 9.20 | 6.34 | 0.74 |
| Proposed model | 9.07 | 6.08 | 0.70 |

Table 14: The result of the training set for 3 data sets

| Data | Paramater | ARIMA | AM | IFTS | Proposed |
|---|---|---|---|---|---|
| Lahi | MAE | 295.28 | 203.39 | 203.3 | 38.43 |
| | MAPE | 44.19 | 26.94 | 26.93 | 5.26 |
| | MSE | 141168.3 | 50562.9 | 50487.7 | 2661.6 |
| NYSE | MAE | 61.41 | 94.29 | 94.28 | 6.46 |
| | MAPE | 0.56 | 0.86 | 0.86 | 0.06 |
| | MSE | 5638.0 | 11224.6 | 11224.4 | 65.1 |
| Taifex | MAE | 39.2 | 76.72 | 76.7 | 18.36 |
| | MAPE | 0.57 | 1.12 | 1.12 | 0.27 |
| | MSE | 3373.7 | 9371.5 | 9366.7 | 683.9 |

Autobox3, ETS, AutoARIMA, Combined ETS, we also compare to the models given in the recent years such as Tai and Nghiep (2019) [29], Tai and Thuy (2020) [28], Yangpeng et al. [33], Dinh and Tai [20]. The authors concluded that their models are better than models from [1] (AM), [16] (LC), [14] (Hua), [11] (BR), [21] (Si), [31] (YH), [12] (Gh), [5] (Chen), [8] (CK), [7] (CH), [15] (Kha), [32] (Yus), [9] (Egr), Tai and Nghiep [29, 30], Tai and Thuy [28]. Because the proposed model is more favorable than the above two models, we can conclude that it has an advantage over many of the models listed above with these considered data sets. In our opinion, there are two main reasons for the advantage of the proposed model:

(i) The universal set is chosen in the proposed model is suitable. This is more suitable than original data in determining the similarity of elements in series. This study has divided the universal set into intervals with the numbers found by the automatic algorithm while others are chosen subjectively. If the elements in the series have many variables, the number of intervals to be divided in the universal set will be more, and vice versa.

(ii) The fuzzy relationship between the elements in series and the divided intervals of Step 5 is improved. This relationship is used to build the new forecasting rule. This rule is considered for two cases: interpolating the historical data and forecast for the future. The forecasting value for future equals its predecessor to plus the mean variation of

Table 15: The result of the test set for 3 data sets

| Data set | Parameter | ARIMA | AM | IFTS | Proposed |
|---|---|---|---|---|---|
| Lahi | MAE | 292.15 | 351.43 | 351.6 | 127.36 |
| | MAPE | 31.55 | 37.95 | 37.97 | 13.56 |
| | MSE | 92160.2 | 132347.6 | 132464.8 | 24376 |
| NYSE | MAE | 207.02 | 192.55 | 192.55 | 61.57 |
| | MAPE | 1.85 | 1.72 | 1.72 | 0.55 |
| | MSE | 49810.0 | 52960.4 | 52960.2 | 7182.5 |
| Taifex | MAE | 79.61 | 78.99 | 79.00 | 73.45 |
| | MAPE | 1.17 | 1.16 | 1.16 | 0.87 |
| | MSE | 7740.1 | 7116.5 | 7117.5 | 6124.7 |

series. This principal makes forecasting results stable for many data sets.

However, the proposed model does not consider the seasonality of the series. This will be our further direction in the near future.

# 5 Conclusion

This study has presented a new forecasting model for time series. The proposed model is the combination of many important improvements from the traditional models. There are the determination of the universe of discourse, the divided interval, and the principle of building the fuzzy relationship to forecast. It is a good combination between cluster analysis and traditional fuzzy time series models. The proposed model has surpassed the existing models for a lot of the benchmark data sets. With many considered time series, including the M3-Competition data set, this comparison is very meaningful. A very important contribution to this research is the establishment of the Matlab procedure for the proposed model. The proposed algorithm can be quickly and effectively applied for real data. With this procedure, we have applied the proposed model to many practical problems and recorded the appropriate forecasting results that the application presented in this study is an example.

# Acknowledgments

This research is funded by Ministry of Education and Training in Viet Nam in the period of 2023-2024 for the project with title "Building the forecasting model for time series based on the improvement of the cluster analysis problem and fuzzy relationships."

# References

[1] A. Abbasov, M. Mamedova, *Application of fuzzy time series to population forecasting*, Vienna University of Technology, **12** (2003), 545-552.

[2] P. H. Abreu, D. C. Silva, J. Mendes-Moreira, L. P. Reis, J. Garganta, *Using multivariate adaptive regression splines in the construction of simulated soccer team's behavior models*, International Journal of Computational Intelligence Systems, **6**(5) (2013), 893-910.

[3] S. Aladag, C. H. Aladag, T. Mentes, E. Egrioglu, *A new seasonal fuzzy time series method based on the multiplicative neuron model and sarima*, Hacettepe Journal of Mathematics and Statistics, **41**(3) (2012), 337-345.

[4] C. H. Aladag, M. A. Basaran, E. Egrioglu, U. Yolcu, V. R. Uslu, *Forecasting in high order fuzzy times series by using neural networks to define fuzzy relations*, Expert Systems with Applications, **36**(3) (2009), 4228-4231.

[5] S. M. Chen, *Forecasting enrollments based on fuzzy time series*, Fuzzy Sets and Systems, **81**(3) (1996), 311-319.

[6] S. M. Chen, *Forecasting enrollments based on high-order fuzzy time series*, Cybernetics and Systems, **33**(1) (2002), 1-16.

[7] S. M. Chen, C. C. Hsu, *A new method to forecast enrollments using fuzzy time series*, International Journal of Applied Science and Engineering, **2**(3) (2004), 234-244.

[8] S. M. Chen, P. Y. Kao, *Taiex forecasting based on fuzzy time series, particle swarm optimization techniques and support vector machines*, Information Sciences, **247** (2013), 62-71.

[9] S. Egrioglu, E. Bas, C. H. Aladag, U. Yolcu, *Probabilistic fuzzy time series method based on artificial neural network*, American Journal of Intelligent Systems, **62**(2) (2016), 42-47.

[10] J. H. Friedman, *Multivariate adaptive regression splines*, The Annals of Statistics, **19**(1) (1991), 1-67.

[11] B. Garg, R. Garg, *Enhanced accuracy of fuzzy time series model using ordered weighted aggregation*, Applied Soft Computing, **48** (2016), 265-280.

[12] H. Ghosh, S. Chowdhury, Prajneshu, *An improved fuzzy time-series method of forecasting based on L–R fuzzy sets and its application*, Journal of Applied Statistics, **43**(6) (2016), 1128-1139.

[13] R. J. Hathaway, J. C. Bezdek, *Recent convergence results for the fuzzy c-means clustering algorithms*, Journal of Classification, **5**(2) (1988), 237-247.

[14] K. Huarng, *Heuristic models of fuzzy time series for forecasting*, Fuzzy Sets and Systems, **123**(3) (2001), 369-386.

[15] M. Khashei, M. Bijari, S. R. Hejazi, *An extended fuzzy artificial neural networks model for time series forecasting*, Iranian Journal of Fuzzy Systems, **8**(3) (2011), 45-66.

[16] H. S. Lee, M. T. Chou, *Fuzzy forecasting based on fuzzy time series*, International Journal of Computer Mathematics, **81**(7) (2004), 781-789.

[17] P. A. Lewis, J. G. Stevens, *Nonlinear modeling of time series using multivariate adaptive regression splines (mars)*, Journal of the American Statistical Association, **86**(416) (1991), 864-877.

[18] S. Makridakis, M. Hibon, *The M3-competition: Results, conclusions and implications*, International Journal of Forecasting, **16**(4) (2016), 451-476.

[19] C. M. Own, P. T. Yu, *Forecasting fuzzy time series on a heuristic high-order model*, Cybernetics and Systems: An International Journal, **36**(7) (2005), 705-717.

[20] D. Phamtoan, T. Vovan, *Building fuzzy time series model from unsupervised learning technique and genetic algorithm*, Neural Computing and Applications, (2021). DOI:10.1007/s00521-021-06485-7.

[21] S. Singh, *A simple method of forecasting based on fuzzy time series*, Applied Mathematics and Computation, **186**(1) (2007), 330-339.

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

[23] Q. Song, B. S. Chissom, *Forecasting enrollments with fuzzy time series–Part II*, Fuzzy Sets and Systems, **62**(1) (1994), 1-8.

[24] W. Sulandari, S. Subanarb, M. Hisyam Lee, P. Canas Rodriguesd, *Time series forecasting using singular spectrum analysis, fuzzy systems and neural networks*, MethodsX, **7** (2020), 1-12.

[25] H. J. Teoh, C. H. Cheng, H. H. Chu, J. S. Chen, *Fuzzy time series model based on probabilistic approach and rough set rule induction for empirical research in stock markets*, Data and Knowledge Engineering, **67**(1) (2008), 103-117.

[26] N. V. Tinh, *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) (2020), 1-26.

[27] T. Vovan, *An improved fuzzy time series forecasting model using variations of data*, Fuzzy Optimization and Decision Making, **18**(2) (2019), 151-173.

[28] T. Vovan, T. Lethithu, *A fuzzy time series model based on improved fuzzy function and cluster analysis problem*, Communications in Mathematics and Statistics, **6** (2022), 51-66.

[29] T. Vovan, L. D. Nghiep, *A new fuzzy time series model based on cluster analysis problem*, International Journal of Fuzzy Systems, **21**(3) (2019), 852-864.

[30] T. Vovan, L. D. Nghiep, *Interpolating time series based on fuzzy cluster analysis problem*, Iranian Journal of Fuzzy Systems, **17**(3) (2020), 151-161.

[31] T. H. K. Yu, K. H. Huarng, *A neural network-based fuzzy time series model to improve forecasting*, Expert Systems with Applications, **37**(4) (2010), 3366-3372.

[32] S. Yusuf, A. Mohammad, A. Hamisu, *A novel two–factor high order fuzzy time series with applications to temperature and futures exchange forecasting*, Nigerian Journal of Technology, **36**(4) (2017), 1124-1134.

[33] Y. Zhang, H. Qu, W. Wang, J. Zhao, *A novel fuzzy time series forecasting model based on multiple linear regression and time series clustering*, Mathematical Problems in Engineering, ID 9546792, (2020), 1-17.

[34] Z. Zhang, Q. Zhu, *Fuzzy time series forecasting based on K-means clustering*, Open Journal of Applied Sciences, **2**(4) (2012), 100-103.