Tạp chí Khoa học Trường Đại học Cần Thơ
Phần A: Khoa học Tự nhiên, Công nghệ và Môi trường
website: sj.ctu.edu.vn

DOI:10.22144/ctu.jvn.2020.009

MỘT MÔ HÌNH DỰ BÁO CHUỖI THỜI GIAN MỜ CẢI TIẾN

Võ Văn Tài[1*], Trang Thị Mỹ Kim[1], Nguyễn Thị Hồng Dân[1], Nguyễn Văn Quang[2], Lê Đại Nghiệp[2] và Huỳnh Văn Hiếu[3]

[1]*Khoa Khoa học Tự nhiên, Trường Đại học Cần Thơ*

[2]*Khoa Cơ bản, Trường Đại học Nam Cần Thơ*

[3]*Khoa Cơ bản, Trường Đại học Công nghiệp Thành phố Hồ Chí Minh*

**Người chịu trách nhiệm về bài viết: Võ Văn Tài (email: vvtai@ctu.edu.vn)*

Thông tin chung:
Ngày nhận bài: 18/10/2019
Ngày nhận bài sửa: 07/12/2019
Ngày duyệt đăng: 28/02/2020

Title:
An improved fuzzy time series forecasting model

Từ khóa:
Chuỗi thời gian mờ, dữ liệu gốc, dự báo, tương lai

Keywords:
Forecasting, fuzzy time series, future, original data

ABSTRACT

The paper proposes a forecasting model for time series based on improvements in determining the universe set and establishment the fuzzy relation. The proposed model is specifically illustrated the steps by the enumerical example and performed effectively by an established R procedure. It has shown more advantages than the popular models such as ARIMA and Abbasov-Manedova (2003) with a lot of the considered bench mark data. The proposed model is also applied in forecasting salty peak for a coastal province in Viet Nam. The examples and applications have shown potential in reality of the researched problem.

TÓM TẮT

Nghiên cứu đề xuất một mô hình dự báo cho chuỗi thời gian dựa trên những cải tiến trong việc xác định tập nền và việc thiết lập các mối quan hệ mờ. Mô hình đề nghị được minh họa cụ thể các bước thực hiện bởi ví dụ số và được thực hiện một cách hiệu quả bằng một chương trình được thiết lập trên phần mềm thống kê R. Nó có ưu điểm hơn các mô hình dự báo phổ biến hiện tại như ARIMA và Abbasov-Manedova (2003) qua nhiều bộ số liệu đối chứng quan trọng. Mô hình đề nghị cũng được áp dụng trong dự báo đỉnh mặn cho một tỉnh ven biển Đồng bằng sông Cửu Long. Các ví dụ và áp dụng đã cho thấy tiềm năng trong thực tế của vấn đề được nghiên cứu.

Trích dẫn: Võ Văn Tài, Trang Thị Mỹ Kim, Nguyễn Thị Hồng Dân, Nguyễn Văn Quang, Lê Đại Nghiệp và Huỳnh Văn Hiếu, 2020. Một mô hình dự báo chuỗi thời gian mờ cải tiến. Tạp chí Khoa học Trường Đại học Cần Thơ. 56(1A): 86-94.

1 GIỚI THIỆU

Dự báo là việc tiên đoán những kết quả sẽ xảy ra trong tương lai dựa vào những nguyên tắc suy luận nào đó. Kết quả dự báo luôn là cơ sở khoa học quan trọng để lập những kế hoạch, những định hướng, những chiến lược phù hợp mang lại hiệu quả cao nhất. Dự báo luôn có một vai trò rất quan trọng trong các lĩnh vực, vì vậy nó luôn nhận được sự quan tâm của các nhà khoa học và quản lý. Mặc dù đã có rất nhiều công trình được công bố, những đầu tư, những cải tiến sâu rộng về lý thuyết và kỹ thuật thực hiện, nhưng cho đến nay nó vẫn là bài toán chưa có lời giải cuối cùng. Để tiến hành dự báo, các nghiên cứu phải dựa vào nhiều yếu tố trong đó dữ liệu quá khứ là vấn đề quan trọng. Trong các loại dữ liệu, chuỗi thời gian được lưu trữ phổ biến và có nhu cầu rất lớn trong thực tế cho việc dự báo. Với dữ liệu chuỗi, hồi qui và chuỗi thời gian là hai mô hình chính được được sử dụng phổ biến trong thống kê.

Khi sử dụng mô hình hồi qui để dự báo, những áp dụng phải giả sử các điều kiện mà trong thực tế dữ liệu rất khó đáp ứng, chính vì vậy mô hình này thường chỉ áp dụng tốt cho những trường hợp đặc thù (Aladag *et al.*, 2012; Abreu *et al.*, 2013). Các mô hình chuỗi thời gian không mờ như tự hồi quy, trung bình di động, trung bình di động tự hồi qui (ARIMA) đã mang lại nhiều kết quả tốt hơn trong dự báo so với các mô hình hồi quy trong nhiều trường hợp. Trong các mô hình này, mô hình ARIMA với phương pháp BoxJenkins (Box and Jenkins, 1973) được sử dụng rất rộng rãi trong nhiều áp dụng thực tế ngày nay. Tuy nhiên các mô hình chuỗi thời gian không mờ chỉ thực sự tốt khi dữ liệu phải có tính dừng và sai số nhiễu của nó phải là một ồn trắng. Vì sự biến đổi phức tạp trong các dữ liệu thực tế, nên chuỗi thời gian không mờ cũng chưa đáp ứng được các yêu cầu khi dự báo. Nhiều trường hợp dự báo rất kém độ chính xác (Chen 1996).

Dựa trên lý thuyết mờ, chuỗi thời gian mờ (FTS) được đề xuất để giải quyết các yếu điểm của chuỗi thời gian không mờ. Song and Chissom (1993) đã đi tiên phong trong nghiên cứu mô hình FTS với dữ liệu tuyển sinh từ Đại học Alabama. Kế thừa nghiên cứu này, hàng loạt các công trình liên quan về FTS được công bố (Huarng, 2001; Chen, 2004; Eren *et al.*, 2014; Chen and Hung, 2016). Những mô hình này được thiết lập dựa trên các cấp độ mờ hóa theo ngôn ngữ với 3, 5 hoặc 7 cấp độ. Hạn chế của các mô hình này là chỉ mờ hóa dữ liệu mờ lịch sử mà không thực hiện được dự báo cho tương lai. Trong số các mô hình này, mô hình của Singh (2007) được đánh giá rất cao. Mặc dù được đề xuất khá lâu nhưng khi đề xuất một mô hình mới, mô hình của Singh (2007) thông thường được sử dụng để so sánh hiệu quả. Tuy nhiên mô hình này cũng chỉ để mờ hóa dữ liệu mà không trực tiếp để dự báo. Sau khi có dữ liệu từ mờ hóa, muốn dự báo chúng ta phải sử dụng một mô hình không mờ nào đó để thực hiện. Việc mờ hóa trước khi thực hiện dự báo có thể làm mất đi một số qui luật của dữ liệu, nên khi dự báo nhiều trường hợp không nhận được kết quả tốt. Một hướng phát triển khác của FTS là trực tiếp dự báo cho tương lai từ nguyên tắc mờ hóa đã thiết lập. Các tài liệu cho thấy, hướng nghiên cứu này chưa có nhiều kết quả được công bố. Abbasov and Mamedova (2003) đã đề xuất hướng nghiên cứu này khi dự báo dân số nước Áo. Mặc dù mô hình đã nhận được kết quả tốt trong nghiên cứu này, nhưng nó lại không thích hợp cho nhiều tập dữ liệu khác. Những tham số trong mô hình này cũng là một thách thức trong áp dụng thực tế. Dựa trên mô hình của Singh (2007) và ý tưởng của Abasov and Mamedova (2003), bài viết này đề xuất một mô hình chuỗi thời gian mờ để dự báo. Mô hình này có thể mờ hóa dữ liệu và dự báo cho tương lai. Mô hình đề nghị nhận được kết quả tốt hơn các mô hình dự báo trực tiếp không mờ ARIMA và mô hình chuỗi thời gian mờ của Abasov and Mamedova (2003) trong các bộ số liệu đối chứng được so sánh.

2 MÔ HÌNH ĐỀ NGHỊ

2.1 Các định nghĩa

Định nghĩa 1. Cho U là không gian nền, $U = \{u_1, u_2, ...u_n\}$. Một tập mờ A của U được xác định như sau:

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

trong đó đó μ_A là hàm thuộc của A, $\mu_A : U \to [0,1]$. $\mu_A(u_i)$ chỉ mức độ thuộc của u_i vào A, $\mu_A(u_i) \in [0,1]$, $1 \le i \le n$.

Định nghĩa 2. Giả sử $F(t)$ được suy ra từ $F(t-1)$, khi đó quan hệ logic mờ giữa $F(t)$ và $F(t-1)$ được biểu diễn bởi phương trình mờ:

$$F(t) = F(t-1) * R(t, t-1),$$

trong đó * là toán tử hợp, $R(t, t-1)$ là quan hệ logic mờ. Nếu ta đặt $F(t-1) = A_{t-1}$ và $F(t) = A_t$ thì quan hệ logic mờ được viết bởi: $A_{t-1} \to A_t$.

Định nghĩa 3. Cho một chuỗi dữ liệu thực tế $\{X_i\}$ và giá trị dự đoán tương ứng $\{\hat{X}_i\}$, i = 1, 2, …, n, khi đó ta có các tiêu chuẩn sau để đánh giá các mô hình FTS:

Bình phương sai số trung bình:

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

Sai số tuyệt đối trung bình:

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

Sai số phần trăm tuyệt đối trung bình:

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

Khi thực hiện dự báo, mô hình có các tiêu chuẩn trên càng nhỏ thì càng tốt.

2.2 Thuật toán đề nghị

Cho X_t, $t = 1, 2, \ldots, n$ là một chuỗi thời gian. Mô hình dự báo chuỗi thời gian mờ (FTSF) đề nghị gồm 8 bước sau:

Bước 1: Chuẩn hóa dữ liệu chuỗi về thang đo 100:

$$K_i = 100X_i / \max\{X_i\},\ i = 1, 2, \ldots, n.$$

Tính các biến đổi của dữ liệu giữa hai khoảng thời gian liên tiếp đã được chuẩn hóa:

$$E_t = K_{t+1} - K_t, t = 1, 2, ..., n-1.$$

Xác định tập nền $U = [D_{\min}, D_{\max}]$, trong đó $D_{\min} = \min\{E_t\}, D_{\max} = \max\{E_t\}, t = 1, 2, ..., n-1.$

Bước 2: Chia tập U thành *m* khoảng bằng nhau và gán tập mờ A_i, $i = \overline{1, m}$ với các hàm thuộc tam giác như sau:

$$A_1 = \{1/u_1, 0.5/u_2, 0/u_3, ..., 0/u_{m-1}, 0/u_m\},$$

$$A_2 = \{0.5/u_1, 1/u_2, 0.5/u_3, ..., 0/u_{m-1}, 0/u_m\},$$

……

$$A_m = \{0/u_1, 0/u_2, 0/u_3, ..., 0.5/u_{m-1}, 1/u_m\}.$$

Trong ứng dụng của bài viết này, nếu một khoảng khi chia chứa nhiều hơn 4 phần tử, tiếp tục chia khoảng này thành 2 khoảng nhỏ đều nhau.

Bước 3: Thiết lập các mối quan hệ logic mờ:

Nếu A_i là giá trị mờ hóa tại thời điểm *i* và A_j là giá trị mờ tại thời điểm *j* thì quan hệ mờ được biểu thị là $A_i \to A_j$. Ở đây A_i được gọi là trạng thái hiện tại và A_j là trạng thái tiếp theo.

Bước 4: Thiết lập nguyên tắc dự báo:

* Trong trường hợp hàm thuộc A_j đạt giá trị lớn nhất, ta kí hiệu

$\left[A_j\right]$ là khoảng tương ứng của u_j,

$L\left[A_j\right]$ và $U\left[A_j\right]$ lần lượt là cận dưới và cận trên của u_j,

$l\left[A_j\right]$ là độ dài của khoảng A_j,

$M\left[A_j\right]$ là giá trị trung bình của khoảng u_j,

* Tính các giá trị:

$$D_i = \left|\left|E_i - E_{i-1}\right| - \left|E_{i-1} - E_{i-2}\right|\right|,$$

$$Z_i = E_i + \frac{D_i}{2},\ ZZ_i = E_i - \frac{D_i}{2},$$

$$Y_i = E_i + D_i,\ YY_i = E_i - D_i,$$

$$P_i = E_i + \frac{D_i}{4},\ PP_i = E_i - \frac{D_i}{4},\ Q_i = E_i + D_i \times 2,$$

$$QQ_i = E_i - D_i \times 2,\ G_i = E_i + \frac{D_i}{6},\ GG_i = E_i - \frac{D_i}{6},$$

$$H_i = E_i + D_i \times 3,\ HH_i = E_i - D_i \times 3.$$

* Với các giá trị ban đầu $R = 0$, $S = 0$, ta có các nguyên tắc tính *R* và *S* như sau:

Nếu $L\left[A_j\right] \le Z_i \le U\left[A_j\right]$ thì $R = R + Z_i$ và $S = S + 1$, ngược lại $R = 0$, $S = 0$.

Nếu $L\left[A_j\right] \le ZZ_i \le U\left[A_j\right]$ thì $R = R + ZZ_i$ và $S = S + 1$, ngược lại $R = 0$, $S = 0$.

Nếu $L\left[A_j\right] \le Y_i \le U\left[A_j\right]$ thì $R = R + Y_i$ và $S = S + 1$, ngược lại $R = 0$, $S = 0$.

Nếu $L\left[A_j\right] \le YY_i \le U\left[A_j\right]$ thì $R = R + YY_i$ và $S = S + 1$, ngược lại $R = 0$, $S = 0$.

Nếu $L\left[A_j\right] \le P_i \le U\left[A_j\right]$ thì $R = R + P_i$ và $S = S + 1$, ngược lại $R = 0$ và $S = 0$.

Nếu $L\left[A_j\right] \le PP_i \le U\left[A_j\right]$ thì $R = R + PP_i$ và $S = S + 1$, ngược lại $R = 0$ và $S = 0$.

Nếu $L\left[A_j\right] \le Q_i \le U\left[A_j\right]$ thì $R = R + Q_i$ và $S = S + 1$, ngược lại $R = 0$ và $S = 0$.

Nếu $L\left[A_j\right] \le QQ_i \le U\left[A_j\right]$ thì $R = R + QQ_i$ và $S = S + 1$, ngược lại $R = 0$ và $S = 0$.

Nếu $L\left[A_j\right] \le G_i \le U\left[A_j\right]$ thì $R = R + G_i$ và $S = S + 1$, ngược lại $R = 0$ và $S = 0$.

Nếu $L\left[A_j\right] \le GG_i \le U\left[A_j\right]$ thì $R = R + GG_i$ và $S = S + 1$, ngược lại $R = 0$ và $S = 0$.

Nếu $L\left[A_j\right] \le H_i \le U\left[A_j\right]$ thì $R = R + H_i$ và $S = S + 1$, ngược lại $R = 0$ và $S = 0$.

Nếu $L\left[A_j\right] \le HH_i \le U\left[A_j\right]$ thì $R = R + HH_i$ và $S = S + 1$, ngược lại $R = 0$ và $S = 0$.

Bước 5: Tính sự biến đổi tại thời điểm t bằng công thức:

$$\hat{V}(t) = \frac{\sum P_j \times V_j}{\sum P_j}, \quad (4)$$

trong đó

P_j là tỷ số giữa số lần của mối quan hệ logic mờ $A_i \to A_j$ xảy ra và tổng số lần tất cả các mối quan hệ logic mờ,

$$V_j = \frac{R + M[A_j]}{S + 1}. \quad (5)$$

Bước 6: Tính giá trị dự báo $\hat{K}(t)$ tại thời điểm *t* bởi công thức sau:

$$\hat{K}(t) = K(t-1) + \hat{V}(t), \quad (6)$$

trong đó

$K(t-1)$ là giá trị của chuỗi tại thời điểm t-1

$\hat{K}(t)$ là biến đổi dự báo tại thời điểm *t*.

3 VÍ DỤ MINH HỌA VÀ KIỂM CHỨNG

3.1 Ví dụ minh họa

Trong phần này chúng tôi lấy số liệu về tuyển sinh của Trường Đại học Alabama (ACD) để minh họa cho thuật toán đề nghị. Đây là số liệu đối chứng được sử dụng trong nhiều nghiên cứu về FTS khi so sánh các mô hình với nhau. Số liệu được cho bởi cột thứ 2 của Bảng 1.

Bảng 1: Số liệu gốc, chuẩn hóa, biến đổi và tập mờ tương ứng của dữ liệu ACD

Năm	X_i	K_i	E_i	A_i
1971	13055	67,513	-	-
1972	13563	70,140	2,627	A_6
1973	13867	71,712	1,572	A_5
1974	14696	75,999	4,287	A_8
1975	15460	79,950	3,951	A_7
1976	15311	79,180	−0,771	A_3
1977	15603	80,690	1,510	A_5
1978	15861	82,024	1,334	A_5
1979	16807	86,916	4,892	A_8
1980	16919	87,496	0,579	A_4
1981	16388	84,749	−2,746	A_2
1982	15433	79,811	−4,939	A_1
1983	15497	80,142	0,331	A_4
1984	15145	78,321	−1,820	A_2
1985	15163	78,414	0,093	A_4
1986	15984	82,660	4,246	A_8
1987	16859	87,185	4,525	A_8
1988	18150	93,862	6,676	A_9
1989	18970	98,102	4,241	A_8
1990	19328	99,954	1,851	A_6
1991	19337	100,00	0,047	A_4
1992	18876	97,616	−2,384	A_2

Theo thuật toán đề nghị, chúng ta có những bước thực hiện cụ thể sau:

Bước 1: Chuẩn hóa dữ liệu về thang đo 100, ta có tập dữ liệu được cho bởi cột K_i của Bảng 1. Tính sự biến đổi số lượng sinh viên giữa hai năm liên tiếp của K_i, ta có giá trị E_i của Bảng 1. Vì $D_{\min} = -4.939$, $D_{\max} = 6.676$, do đó tập nền đoạn $U = [-4.939; 6.676]$.

Bước 2: Chia tập nền U thành 7 đoạn đều nhau:

u_1=[−4,939; −3,279], u_2 = [−3,279; −1,620], u_3 = [−1,620;0,309], u_4 = [0,309;1,699],

u_5 = [1,699;3,358]; u_6 = [3,358;5,017]; u_7 = [5,017;6,676].

Vì u_4 và u_6 chứa lần lượt 7 và 6 giá trị của E_i nên ta chia mỗi khoảng này thành 2 khoảng nhỏ, các khoảng khác giữ nguyên. Như vậy ta có 9 khoảng u_i và các điểm giữa của nó u_m^i được cho bởi Bảng 2.

Bảng 2: Các đoạn chia của tập nền *U*

u_i	u_m^i
[−4,939; −3,279]	−4,109
[−3,279; −1,620]	−2,450
[−1,620; 0,039]	−0,791
[0,039; 0,869]	0,454
[0,869; 1,699]	1,284
[1,699; 3,358]	2,528
[3,358; 4,187]	3,773
[4,187; 5,017]	4,602
[5,017; 6,676]	5,847

Bước 3: Các tập mờ A_i của Bảng 1 tương ứng với từng đoạn u_i được cho bởi Bảng 2 và được cụ thể như sau:

$$A_1 = \{1/u_1, 0,5/u_2, 0/u_3, 0/u_4, 0/u_5, 0/u_6, 0/u_7, 0/u_8, 0/u_9\},$$
$$A_2 = \{0,5/u_1, 1/u_2, 0,5/u_3, 0/u_4, 0/u_5, 0/u_6, 0/u_7, 0/u_8, 0/u_9\},$$
$$A_3 = \{0/u_1, 0,5/u_2, 1/u_3, 0,5/u_4, 0/u_5, 0/u_6, 0/u_7, 0/u_8, 0/u_9\},$$
$$A_4 = \{0/u_1, 0/u_2, 0,5/u_3, 1/u_4, 0,5/u_5, 0/u_6, 0/u_7, 0/u_8, 0/u_9\},$$
$$A_5 = \{0/u_1, 0/u_2, 0/u_3, 0,5/u_4, 1/u_5, 0.5/u_6, 0/u_7, 0/u_8, 0/u_9\},$$
$$A_6 = \{0/u_1, 0/u_2, 0/u_3, 0/u_4, 0,5/u_5, 1/u_6, 0,5/u_7, 0/u_8, 0/u_9\},$$
$$A_7 = \{0/u_1, 0/u_2, 0/u_3, 0/u_4, 0/u_5, 0.5/u_6, 1/u_7, 0.5/u_8, 0/u_9\},$$
$$A_8 = \{0/u_1, 0/u_2, 0/u_3, 0/u_4, 0/u_5, 0/u_6, 0,5/u_7, 1/u_8, 0,5/u_9\},$$
$$A_9 = \{0/u_1, 0/u_2, 0/u_3, 0/u_4, 0/u_5, 0/u_6, 0/u_7, 0,5/u_8, 1/u_9\}.$$

Bước 4: Mối quan hệ mờ giữa các A_i được cho bởi Bảng 3:

Bảng 3: Mối quan hệ mờ của các A_i

$A_1 \to A_4 : 1$
$A_2 \to A_1 : 1 \; A_2 \to A_4 : 1$
$A_3 \to A_5 : 1$
$A_4 \to A_2 : 1 \; A_4 \to A_8 : 1$
$A_5 \to A_5 : 1 \; A_5 \to A_8 : 2$
$A_6 \to A_4 : 1 \; A_6 \to A_5 : 1$
$A_7 \to A_3 : 1$
$A_8 \to A_4 : 1 \; A_8 \to A_6 : 1 \; A_8 \to A_7 : 1 \; A_8 \to A_8 : 1 \; A_8 \to A_9 : 1$
$A_9 \to A_8 : 1$

Bước 5: Tính sự biến đổi

Giả sử chúng ta cần phải dự báo giá trị chuỗi năm 1976. Dựa vào Bảng 1, Bảng 2 và Bảng 3, có thể thấy rằng sự khác biệt năm 1975 rơi vào tập mờ A_7, và mối quan hệ logic mờ được thiết lập $A_7 \to A_3$ với tần suất là 1.

Ta có

$\left[A_3\right] = \left[-1,620; 0,039\right]$, $U\left[A_3\right] = 0,039$, $L\left[A_3\right] = -1,620$,

$M\left[A_3\right] = -0,791$, $E_4 = 4,287$, $E_3 = 1,572$, $E_2 = 2,627$.

Khi đó chúng ta nhận được

$$D_5 = \left\| E_5 - E_4 \right| - \left| E_4 - E_3 \right\| = 1,660,$$

$$Z_5 = E_5 + \frac{D_5}{2} = 5,140,$$

$$ZZ_5 = E_5 - \frac{D_5}{2} = 2,762,$$

$$Y_5 = E_5 + D_5 = 6,330,$$

$$YY_5 = E_5 - D_5 = 1,572,$$

$$P_5 = E_5 + \frac{D_5}{4} = 4,546,$$

$$PP_5 = E_5 - \frac{D_5}{4} = 3,356,$$

$$Q_5 = E_5 + D_5 \times 2 = 8,709,$$

$$QQ_5 = E_5 - D_5 \times 2 = -0,807,$$

$$G_5 = E_5 + \frac{D_5}{6} = 4,348,$$

$$GG_5 = E_5 - \frac{D_5}{6} = 3,555,$$

$$H_5 = E_5 + D_5 \times 3 = 11,088,$$

$HH_5 = E_5 - D_5 \times 3 = -3,186$. $Z_5 > L\left[A_3\right]$ và $Z_5 > U\left[A_3\right]$; khi đó $R = 0$ và $S = 0$.

$ZZ_5 > L\left[A_3\right]$ và $ZZ_5 > U\left[A_3\right]$; khi đó $R = 0$ và $S = 0$.

$Y_5 > L[A_3]$ và $Y_5 > U[A_3]$; khi đó $R = 0$ và $S = 0$.

$YY_5 > L[A_3]$ và $YY_5 > U[A_3]$; khi đó $R = 0$ và $S = 0$.

$P_5 > L[A_3]$ và $P_5 > U[A_3]$; khi đó $R = 0$ và $S = 0$.

$PP_5 > L[A_3]$ và $PP_5 > U[A_3]$; khi đó $R = 0$ và $S = 0$.

$Q_5 > L[A_3]$ và $Q_5 > U[A_3]$; khi đó $R = 0$ và $S = 0$.

$QQ_5 > L[A_3]$ và $QQ_5 < U[A_3]$; khi đó $R = R + QQ_5 = -0.8067$ và $S = S + 1 = 1$.

$H_5 > L[A_3]$ và $H_5 > U[A_3]$; khi đó $R = -0.8067$ và $S = 1$.

$HH_5 > L[A_3]$ và $HH_5 > U[A_3]$; khi đó $R = -0.8067$ và $S = 1$.

Dựa vào công thức (5), ta tính được sự biến đổi tương ứng với quan hệ mờ $A_7 \to A_3$ cho năm 1976 như sau:

$$\hat{V}_3 = \frac{R + M[A_3]}{S+1} = \frac{-0,807 - 0,791}{1+1} = -0,797.$$

Bước 7: Tính sự biến đổi dự báo cho năm 1976 theo công thức (4), ta nhận được

$$\hat{V}_{final} = \hat{V}_3 = -0.7986.$$

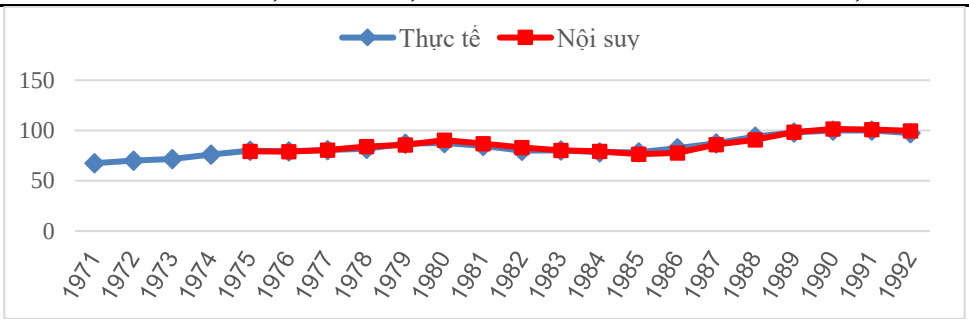
Bước 8: Tính giá trị dự báo $\hat{K}(t)$ cho năm 1976 theo công thức (6), ta có

$$\hat{K}(t) = K(t-1) + \hat{V}_{final} = 79.9504 + (-0.7986) = 79.1518.$$

Thực hiện tương tự cho các năm tiếp theo cho đến năm 1992 ta được Bảng 4.

Bảng 4: Số lượng tuyển sinh được dự báo theo mô hình đề nghị

Năm	Thực tế	Chuẩn hóa	Nội suy	Năm	Thực tế	Chuẩn hóa	Nội suy
1971	13055	67,5131	-	1982	15433	79,8107	83,0800
1972	13563	70,1401	-	1983	15497	80,1417	80,2647
1973	13867	71,7123	-	1984	15145	78,3214	79,3438
1974	14696	75,9994	-	1985	15163	78,4144	76,4770
1975	15460	79,9504	79,4133	1986	15984	82,6602	77,7277
1976	15311	79,1798	79,1518	1987	16859	87,1852	86,1320
1977	15603	80,6899	80,5327	1988	18150	93,8615	90,6237
1978	15861	82,0241	84,1233	1989	18970	98,1021	98,5648
1979	16807	86,9163	85,6697	1990	19328	99,9535	101,478
1980	16919	87,4955	90,4361	1991	19337	100,0000	100,822
1981	16388	84,7494	87,0952	1992	18876	97,6160	99,5919



Hình 1: Đồ thị số liệu thực tế và dự báo của dữ liệu Enrollment

Từ số liệu Bảng 4, áp dụng công thức (2) ta tính được các giá trị của MAE =1.6508.

Hình 1 cho thấy kết quả dự báo và thực tế khá gần nhau.

3.2 Một số so sánh

Bên cạnh dữ liệu ACD, nghiên cứu này sử dụng thêm 3 tập dữ liệu Taifex, Outpatient and Grain (Tai and Nghiep 2019) để so sánh mô hình dự báo đề nghị với mô hình ARIMA và mô hình của Abbassov and Manedova. Đây là những tập dữ liệu phổ biến

được sử dụng để đánh giá hiệu quả của các mô hình trong nhiều bài báo. Mỗi tập dữ liệu được chia thành 2 phần: Tập huấn luyện và tập kiểm tra với tỉ lệ lần lượt là 80% và 20%. Tập huấn luyện được sử dụng để xây dựng các mô hình ARIMA, Abasov-Manedova (2003) và mô hình đề nghị. Sử dụng mô hình từ tập huấn luyện, dự báo cho thời gian của tập kiểm tra để so sánh với số liệu thực tế. Các tham số đánh giá MSE, MAE và MAPE được sử dụng để so sánh hiệu quả của các mô hình.

Kết quả thực hiện của tập huấn luyện được cho bởi Bảng 5 và tập kiểm tra được cho bởi Bảng 6.

Bảng 5: So sánh mô hình đề nghị và các mô hình khác cho tập huấn luyện

Dữ liệu	Phương pháp	MAE	MAPE	MSE
Enrollment	ARIMA	2,569	2,781	8,582
	AM	2,655	2,790	11,226
	Mô hình đề nghị	**2,042**	**2,161**	**6,026**
Outpatient	ARIMA	4,938	5,582	36,266
	AM	6,480	7,276	58,956
	Mô hình đề nghị	**4,700**	**5,388**	**35,288**
Foodgrain	ARIMA	4,269	6,480	27,242
	AM	4,628	6,898	31,628
	Mô hình đề nghị	**3,369**	**5,203**	**17,692**

Bảng 6: So sánh mô hình đề nghị và các mô hình khác cho tập kiểm tra

Dữ liệu	Phương pháp	MAE	MAPE	MSE
Enrollment	ARIMA	8,074	7,181	68,725
	AM	12,333	10,974	158,777
	Mô hình đề nghị	**6,164**	**5,502**	**44,688**
Outpatient	ARIMA	12,193	16,842	189,118
	AM	14,888	20,798	322,431
	Mô hình đề nghị	**5,059**	**6,955**	**38,406**
Foodgrain	ARIMA	7,481	7,399	84,086
	AM	7,409	7,909	89,368
	Mô hình đề nghị	**5,688**	**5,942**	**52,915**

Bảng 5 và Bảng 6 cho thấy rằng mô hình đề nghị đã nhận được kết quả tốt nhất với các tập dữ liệu được xem xét.

4 ÁP DỤNG

Trong phần này chúng tôi áp dụng phương pháp đề nghị để dự báo đỉnh mặn cho ba trạm đo chính của tỉnh Cà Mau. Số liệu thực hiện được cho bởi Bảng 7.

Mục đích của nghiên cứu này là sử dụng số liệu quá khứ để dự báo đỉnh mặn tại ba trạm đo chính của tỉnh Cà Mau đến năm 2025. Toàn bộ dữ liệu quá khứ được sử dụng để kiểm tra hiệu quả của các mô hình dự báo ARIMA, Abbasov-Manedova (2003) và mô hình đề nghị bởi các tham số MAE, MAPE và MSE. Mô hình nào tốt nhất sẽ được sử dụng để dự báo cho tương lai. Kết quả so sánh các mô hình được cho bởi Bảng 8.

Bảng 7: Số liệu đỉnh mặn Cà Mau tại ba trạm đo giai đoạn 2000-2017

Năm	Gành Hào	Cửa Lớn	Ông Đốc
2000	31,5	29,6	30,8
2001	30,8	29,4	31,8
2002	30,5	34,4	34,7
2003	33,8	35,1	34,8
2004	32,6	34,3	34,1
2005	33,5	36,1	35,2
2006	32,6	31,6	31,6
2007	32,2	32,9	32,9
2008	31,4	31,5	31,5
2009	32,4	28,3	30.2
2010	33,2	37,1	39,7
2011	31,0	28,4	30,9
2012	31,9	27,3	31,7
2013	31,7	33,1	31,9
2014	30,6	313	31,8
2015	31,5	33.1	35,9
2016	32,9	35,9	37,9
2017	33,7	36,5	38,8

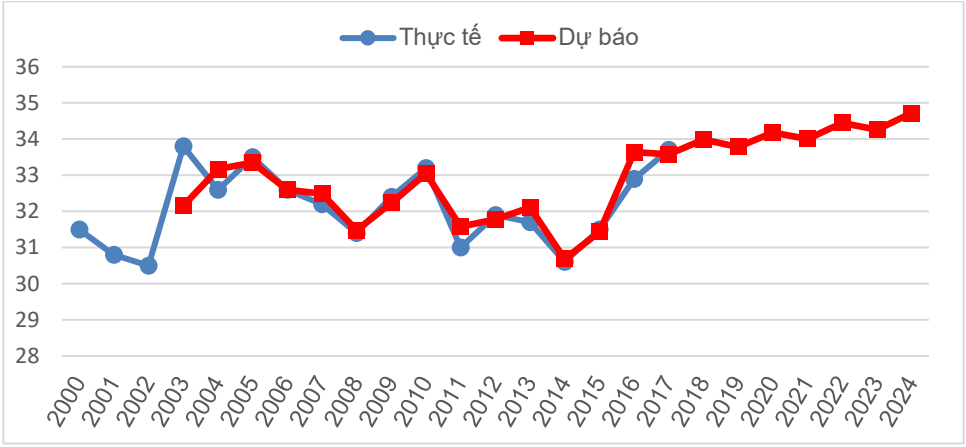
Bảng 8: So sánh hiệu quả của các mô hình dự báo đỉnh mặn tại ba trạm đo

Dữ liệu	Phương pháp	MAE	MAPE	MSE
Cà Mau	ARIMA	6,597	7,530	64,131
	AM	9,658	11,164	145,033
	Mô hình đề nghị	**4,482**	**5,167**	**38,928**
Gành Hào	ARIMA	2,409	2,511	9,230
	AM	2,674	2,849	11,508
	Mô hình đề nghị	**2,366**	**2,509**	**8,376**
Ông Đốc	ARIMA	5,888	6,681	55,354
	AM	6,655	7,747	101,425
	Mô hình đề nghị	**5,058**	**5,854**	**47,096**

Bảng 8 cho thấy mô hình đề nghị luôn nhận được kết quả tốt nhất, do đó nó được sử dụng để dự báo cho tương lai. Kết quả dự báo được cho bởi Bảng 9.

Bảng 9: Dự báo đỉnh mặn tại 3 trạm đo chính của tỉnh Cà Mau giai đoạn 2018 – 2025

Năm	Cửa Lớn	Gành Hào	Ông Đốc
2018	35,577	33,576	38,031
2019	36,781	33,984	37,224
2020	35,764	33,785	36,244
2021	36,908	34,184	39,374
2022	35,864	34,001	41,080
2023	37,008	34,448	42,774
2024	35,964	34,263	44,460
2025	37,108	34,721	45,528



Hình 2: Số liệu thực tế và dự báo đỉnh mặn trạm đo Gành Hào

Bảng 8 cho thấy đỉnh mặn tại 3 trạm đo chính của tỉnh Cà Mau trong tương lai đều có khuynh hướng tăng. Mặc dù không có sự biến đổi phức tạp như quá khứ nhưng chúng luôn ở mức cao. Hình 2 biểu thị số liệu thực tế và số liệu quá khứ đỉnh mặn tại trạm đo Gành Hào. Hình 2 cho thấy số liệu đỉnh mặn dự báo và thực tế giai đoạn 2004 – 2017 khá trùng nhau nên kết quả dự báo tương lai được đánh giá đáng tin cậy. Với hai trạm đo còn lại ta cũng có mức độ tin cậy tương tự.

5 KẾT LUẬN

Bài viết đã trình bày một mô hình mới để dự báo cho dữ liệu dạng chuỗi, một kiểu dữ liệu được lưu trữ phổ biến, có nhu cầu dự báo rất lớn ngày nay. Dựa trên nhiều bước cải tiến quan trọng của các mô hình chuỗi thời gian mờ trước đó, mô hình đề nghị nhận được kết quả tốt hơn các mô hình được sử dụng phổ biến hiện tại qua một số dữ liệu đối chứng phổ biến được so sánh. Nó cũng được áp dụng thử nghiệm trong dự báo đỉnh mặn cho một tỉnh ven biển của Đồng bằng sông Cửu Long. Với mô hình mới này, trong thời gian sắp tới, chúng tôi sẽ tiếp tục cải tiến bước chia tập nền và áp dụng cho nhiều dự báo quan trọng của thực tế.

TÀI LIỆU THAM KHẢO

Abbasov, A. M., and Mamedova, M. H., 2003. Application of fuzzy time series to population forecasting. Vienna University of Technology. 12: 545–552.

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

Aladag, S., Aladag, C. H., Mentes, T., and Egrioglu, E., 2012. A new seasonal fuzzy time series method based on the multiplicative neuron model and SARIMA. Hacettepe Journal of Mathematics and Statistics. 41(3): 145–163.

Bas, E., Vedide, Uslu, V. R., Yolcu, U., and Egrioglu, E., 2014. A modied genetic algorithm for forecasting fuzzy time series. Applied Intelligence, 41(2): 453–463.

Box, G. E. P., and Jenkins, G. M., 1970. Time series analysis: Forecasting and control. Holden-Day. San Fancisco, 546 pages.

Chen, S. M., 1996. Forecasting enrollments based on fuzzy time series. Fuzzy Sets and Systems. 81(3): 311–319.

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

Chen, J. and Hung, W., 2015. An automatic clustering algorithm for probability density functions. J. Stat. Comput. Simul. 85(1): 3047–3063.

Huarng, K., 2001. Heuristic models of fuzzy time series for forecasting. Fuzzy Sets and Systems. 123(3): 369–386.

Singh, S. R., 2007. A simple method of forecasting based on fuzzy time series. Applied Mathematics and Computation. 186(1): 330–339.

Song, Q. and Chissom, B. S., 1993. Fuzzy time series and its models. Fuzzy Sets and Systems. 54(3): 269–277.

Tai V. V., and Nghiep L. D., 2019. An improved fuzzy time series forecasting model using variations of data. 21(3): 852 – 864.