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.2019.099

MỘT MÔ HÌNH MỜ HÓA CHUỖI THỜI GIAN CẢI TIẾN

Võ Văn Tài[1*], Nguyễn Huỳnh Luận[2], Lê Thị Mỹ Xuân[1], La Thuận Bửu[2] và Lê Thị Thu Thùy[3]
[1]Khoa Khoa học Tự nhiên, Trường Đại học Cần Thơ
[2]Học viên cao học, Trường Đại học Cần Thơ
[3]Khoa Cơ bản, Trường Đại học Sư phạm Kỹ thuật Vĩnh Long
*Người chịu trách nhiệm về bài viết: Võ Văn Tài (email: vvtai@ctu.edu.vn)

Thông tin chung:
Ngày nhận bài: 20/01/2019
Ngày nhận bài sửa: 03/04/2019
Ngày duyệt đăng: 29/08/2019

Title:
An improved time series interpolating model

Từ khóa:
Dự báo, chuỗi thời gian mờ, mờ hóa, sự biến đổi của dữ liệu

Keywords:
Forecast, fuzzy time series, interpolate, variation of data

ABSTRACT

based on the improvements in building universe set, the relation of each element in series and the principle for defuzzification, this paper is to propose a time series interpolating model. The parameters in the proposed model are considered so that they can be applied in reality and are illustrated for specific steps by the numerical example. The proposed model has more advantages than some existing models by a lot of the considered benchmark data. It is also applied in forecasting salty peak for a coastal province in the Mekong Delta. This application also shows the potential of the proposed model in forecasting.

TÓM TẮT

Dựa trên những cải tiến trong việc xây dựng tập nền, mối quan hệ của mỗi phần tử trong chuỗi và nguyên tắc giải mờ, bài viết này đề xuất một mô hình mờ hóa dữ liệu chuỗi thời gian. Các tham số trong mô hình đề nghị được xem xét để có thể ứng dụng trong thực tế và được minh họa cụ thể qua các bước thực hiện bởi ví dụ số. Mô hình đề nghị có ưu điểm hơn một số mô hình phổ biến được sử dụng hiện tại qua nhiều tập dữ liệu chuẩn được xem xét. Nó cũng được áp dụng trong dự báo đỉnh mặn cho một tỉnh ven biển Đồng bằng sông Cửu Long. Áp dụng này cũng cho thấy tiềm năng trong dự báo của mô hình được nghiên cứu.

Trích dẫn: Võ Văn Tài, Nguyễn Huỳnh Luận, Lê Thị Mỹ Xuân, La Thuận Bửu và Lê Thị Thu Thùy, 2019. Một mô hình mờ hóa chuỗi thời gian cải tiến. Tạp chí Khoa học Trường Đại học Cần Thơ. 55(4A): 92-100.

1 GIỚI THIỆU

Dự báo là việc tiên đoán những kết quả sẽ xảy ra trong tương lai dựa vào số liệu quá khứ và một qui tắc được thiết lập. Nó là cơ sở khoa học quan trọng cho những kế hoạch, những chính sách, những chiến lược phát triển phù hợp. Chính vì vậy, dự báo luôn nhận được sự quan tâm của các nhà quản lý, các nhà khoa học. Trong thống kê để thiết lập mô hình dự báo, chúng ta phải dựa vào số liệu quá khứ. Trong các loại dữ liệu, chuỗi thời gian được lưu trữ phổ biến và có nhu cầu rất lớn trong thực tế cho dự báo. Với dữ liệu này, hai mô hình chính được sử dụng để dự báo là hồi quy và chuỗi thời gian. Mô hình hồi quy có những ràng buộc về điều kiện của dữ liệu mà trong thực tế rất khó thỏa mãn, do đó nó có hạn chế trong nhiều trường hợp (Box and Jenkins, 1970; Chen, 1996; Tai, 2018). Mô hình chuỗi thời gian (TSM) được đánh giá có nhiều ưu điểm hơn nên được sử dụng rất phổ biến ngày nay. Nhiều nhà nghiên cứu đã sử dụng các mô hình TSM như tự hồi quy (AR), mô hình tự hồi quy trung bình trượt (ARIMA) để ứng dụng trong kinh tế, môi trường và thuỷ văn (Huarng, 2001; Chen and Hsu, 2004;

Singh, 2007; Olivesia and Ludermir, 2014). Tuy nhiên, để xây dựng được mô hình TSM tốt thì dữ liệu phải dừng và sai số của nó phải là ồn trắng. Đây là những điều kiện mà thực tế rất khó để đáp ứng. Do đó, nhiều trường hợp cho kết quả dự báo kém khi sử dụng các mô hình TSM. Mặc dù nhiều tác giả như Abreue *et al*. (2013), Oliveira and Ludermir (2014) đã cố gắng cải thiện mô hình ban đầu, nhưng họ vẫn còn gặp nhiều khó khăn để thực hiện dự báo hợp lý cho nhiều dữ liệu thực tế. Cho đến nay, một mô hình có thể được đánh giá tốt hơn các mô hình khác dựa trên từng dữ liệu cụ thể mà không phải cho tất cả các trường hợp.

Một vấn đề khác là các mô hình TSM truyền thống không thể giải quyết các vấn đề dự báo, nơi các dữ liệu lịch sử được trình bày bằng các biến ngôn ngữ. Các mô hình chuỗi thời gian mờ (FTS) đã giải quyết nhược điểm này. Các mô hình FTS được phát triển theo hai hướng chính. Hướng thứ nhất là xây dựng các mô hình từ dữ liệu gốc và sử dụng nó để dự báo cho tương lai một cách trực tiếp. Abbasov and Mamedova (2003) (mô hình AM) và Tai (2018) (mô hình IFTS) đã có những đóng góp quan trọng theo hướng này. Hướng thứ hai là sự mờ hoá dữ liệu gốc để có được mối quan hệ giữa các phần tử trong chuỗi, sau đó áp dụng các mô hình dự báo đã biết cho dữ liệu đã mờ hóa này. Hướng nghiên cứu này đã và đang được rất nhiều nhà thống kê quan tâm, trong đó Song and Chissom (1993) là những người tiên phong. Bài báo này thực hiện nghiên cứu FTS theo hướng thứ hai.

Theo hướng thứ hai, mô hình FTS thông thường bao gồm ba giai đoạn: (i) xác định tập nền từ dữ liệu gốc, chia khoảng cho tập nền và tìm số lượng các phần tử cho mỗi khoảng; (ii) xây dựng các mối quan hệ mờ, và (iii) giải mờ. Đối với (i), nhiều tác giả đã sử dụng giá trị nhỏ nhất và giá trị lớn nhất của dữ liệu ban đầu để xác định tập nền (Chen, 1996; Chen and Hsu, 2004). Ngoài ra, Huarng (2001), Huarng and Yu (2006) đã đề xuất hai kỹ thuật mới để xác định các khoảng của tập nền dựa trên giá trị trung bình và phân phối của toàn chuỗi. Một cách khác để xây dựng tập nền là dựa trên sự thay đổi dữ liệu giữa các khoảng thời gian liên tiếp hoặc tỷ lệ phần trăm thay đổi (Abbasov and Mamedova, 2003). Nhiều tác giả khác đã chia số tập mờ bằng cách kiểm tra trong nhiều trường hợp để có các thông số đánh giá thích hợp cho từng dữ liệu mà không phải là một quy tắc chung cho tất cả các trường hợp (Singh, 2007; Eren *et al*. 2014). Số lượng tập mờ và các phần tử của chúng cũng được đề xuất dựa trên các thuật toán k – trung bình, thuật toán di truyền và thuật toán phân tích chùm mờ (Zhiqiang, 2012). Mặc dù đã có nhiều tác giả thảo luận về vấn đề này, nhưng cho đến nay sự lựa chọn tối ưu vẫn chưa được tìm thấy. Đối với (ii), một số nghiên cứu đã được thực hiện, Song and Chissom (1993) đã sử dụng các phép toán ma trận, Chen (1996) và một số nhà nghiên cứu khác sử dụng bảng nhóm quan hệ mờ. Trong khi đó, Aladag (2012) đã sử dụng mạng thần kinh nhân tạo để xác định mối quan hệ mờ. Đối với (iii), hầu hết các nghiên cứu sử dụng phương pháp trọng tâm để thực hiện (Chen, 1996; Huarng, 2001; Huarng and Yu, 2006).

Bài báo này đóng góp cho ba giai đoạn (i), (ii) và (iii) đối với mô hình FTS. Đối với giai đoạn (i), sau khi chuẩn hóa dữ liệu, bài viết tìm trị tuyệt đối phần trăm sự biến đổi của hai giá trị liên tiếp. Tập nền được lấy là giá trị min và max của các giá trị này. Cho giai đoạn (ii), dựa trên nghiên cứu của (Chen and Hung, 2015; Tai and Thao, 2018), chúng tôi đề xuất một nguyên tắc mới để tìm các mối quan hệ mờ. Đây cũng là đóng góp quan trọng của nghiên cứu này về mặt lý thuyết. Một quy tắc giải mờ mới cũng được thiết lập trong bài báo này. Đây là sự đóng góp cho giai đoạn (iii) của mô hình chuỗi thời gian mờ. Kết hợp tất cả các cải tiến, bài viết này đề nghị mô hình mờ hóa dữ liệu chuỗi tốt hơn so với các mô hình hiện có thông qua nhiều bộ dữ liệu khác nhau được xem xét. Một đóng góp quan trọng của mô hình này là việc ứng dụng mô hình đề nghị dự báo đỉnh mặn tại hai trạm đo chính của tỉnh Trà Vinh. Ứng dụng này có thể áp dụng tương tự cho rất nhiều vấn đề thực tế khác.

Phần tiếp theo của bài báo được cấu trúc như sau: Phần 2 xem xét một số khái niệm cơ bản về mô hình FTS, đề xuất mô hình mới và một số vấn đề liên quan đến mô hình này. Phần 3 minh họa thuật toán đề nghị và so sánh nó với các mô hình khác qua một số bộ số liệu chuẩn quan trọng. Một ứng dụng thực tế được trình bày trong Phần 4. Cuối cùng là phần kết luận của bài viết.

2 MỘT SỐ ĐỊNH NGHĨA VÀ THUẬT TOÁN ĐỀ NGHỊ

2.1 Một số định nghĩa

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

$$A=\{\mu_A(u_1)/u_1,\mu_A(u_2)/u_2,....,\mu_A(u_n)/u_n\}, \quad (1)$$

trong đó: $\mu_A(u_i)$ là hàm thuộc, $\mu_A(u_i):U\rightarrow[0,1]$, $\mu_A(u_i)$ cho biết mức độ liên hệ của u_i trong A, $\mu_A(u_i)\in[0,1]$, $1\le i\le n$.

Định nghĩa 2. Cho $X(t)$, (t = 1, 2, …) là tập nền với tập mờ $\mu_A(u_i)$, (i = 1, 2, …) và $F(t)$ là tập hợp các giá trị $\mu_A(u_i)$, (i = 1, 2, …). Khi đó $F(t)$ được gọi là chuỗi thời gian mờ (FTS) trên $X(t)$.

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

Bình phương sai số trung bình:

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

Sai số tuyệt đối trung bình:

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

Sai số phần trăm tuyệt đối trung bình:

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

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

2.2 Thuật toán đề nghị

Cho chuỗi thời gian $\left\{X_t, t=\overline{1,n}\right\}$. Dựa vào bài toán phân tích chùm, mô hình FTS được đề nghị với 6 bước như sau:

Bước 1: Tìm phần trăm biến đổi dữ liệu giữa hai thời gian liên tiếp.

$$Y=\left\{Y_{t+1}:Y_{t+1}=\frac{X_{t+1}-X_t}{X_t}.100,\ t=\overline{1,n-1}\right\} \quad (5)$$

Bước 2: Xác định số chùm thích hợp cho tập Y bởi các bước sau:

Bước 2.1: Khởi tạo dãy trọng tâm ban đầu $Z^{(0)}=\left\{v_2^{(0)},v_3^{(0)},\ldots,v_n^{(0)}\right\}=\{Y_2,Y_3,\ldots,Y_n\}$ và một số dương ε rất nhỏ.

Bước 2.2: Cập nhật dãy trọng tâm theo công thức (6):

$$v_i^{(t+1)}=\frac{\sum_{j=2}^{n}K_\lambda\left(v_i^{(t)},v_j^{(t)}\right),v_i^{(t)}}{\sum_{j=2}^{n}K_\lambda\left(v_i^{(t)},v_j^{(t)}\right)},\quad j=\overline{2,n}, \quad (6)$$

trong đó:

$$K_\lambda\left(v_i^{(t)},v_j^{(t)}\right)=\begin{cases}\exp(-d/\lambda) & khi\quad d\left(v_i^{(t)},v_j^{(t)}\right)\le d_S,\\ 0 & khi\quad d\left(v_i^{(t)},v_j^{(t)}\right)>d_S,\end{cases}$$

với $d_S=\frac{2}{n(n-1)}\sum_{i<j}d\left(v_i^{(t)},v_j^{(t)}\right)$ là trung bình các khoảng cách Euclide của các điểm dữ liệu và λ là một số nguyên phụ thuộc vào d_s.

Trên thực tế, λ xác định số khoảng chia của tập nền. Nếu $\lambda\to 0$ thì tập dữ liệu có kích thước n sẽ có n khoảng chia. Ngược lại, nếu $\lambda\to\infty$ thì tập dữ liệu có duy nhất một khoảng chia. Trong bài viết này, chúng tôi chọn $\lambda=d_S/48$.

Bước 2.3: Lặp lại Bước 2.2 cho đến khi $\max_i d\left(v_i^{(t)},v_j^{(t)}\right)<\varepsilon$.

Sau khi các bước lặp kết thúc, mỗi phần tử trong dữ liệu sẽ hội tụ về phần tử đại diện của chùm mà phần tử đó thuộc. Khi thuật toán dừng lại, chúng ta sẽ có k phần tử đại diện tương ứng với k chùm cần phân tích.

Bước 3: Xác định các phần tử thuộc vào trọng tâm cho mỗi chùm bằng thuật toán phân tích chùm k-trung bình như sau:

Bước 3.1: Chia phần tử thành k chùm với số lượng phần tử thuộc mỗi chùm ban đầu được lấy từ kết quả ở Bước 2. Tính trọng tâm của mỗi chùm.

Bước 3.2: Tính khoảng cách từ mỗi phần tử đến trọng tâm của mỗi chùm. Nếu khoảng cách từ một phần tử đến trọng tâm của chùm nó đang thuộc là nhỏ nhất thì ta giữ phần tử đó trong chùm ban đầu. Nếu tồn tại một chùm khác mà khoảng cách từ phần tử đang xét đến trọng tâm của chùm đó là nhỏ nhất thì ta gán phần tử đang xét vào chùm này, bỏ phần tử trong chùm nó đang thuộc. Nếu phần tử được chuyển đến chùm khác thì phải tính lại trọng tâm của hai chùm mới có sự thay đổi.

Bước 3.3: Lặp lại bước 3.2 cho đến khi khoảng cách từ một phần tử bất kỳ đến trọng tâm chùm nó đang thuộc nhỏ hơn khoảng cách đến trọng tâm các chùm khác.

Sau khi thuật toán k-trung bình kết thúc, số phần tử cụ thể và trọng tâm $c_i\ \left(i=\overline{1,k}\right)$ của mỗi chùm sẽ được xác định.

Bước 4: Xác định tập nền U cho dữ liệu dựa vào công thức:

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

Chia tập nền U thành k khoảng như sau:

$$U_1 = \left[c_1 - \frac{c_2 - c_1}{2}; \frac{c_1 + c_2}{2}\right),$$

$$U_i = \left[\frac{c_{i-1} + c_i}{2}; \frac{c_i + c_{i+1}}{2}\right) \quad i = \overline{2, k-1},$$

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

Bước 5: Xác định các tập mờ A_i tương ứng với các khoảng U_i và các mối quan hệ mờ như sau:

$$T_i = \begin{cases} \Delta_1 \cdot \dfrac{1.5}{\dfrac{1}{|a_1|} + \dfrac{0.5}{|a_2|}} & khi \quad i = 1 \\ \Delta_i \cdot \dfrac{2}{\dfrac{1}{|a_{i-1}|} + \dfrac{1}{|a_i|} + \dfrac{0.5}{|a_{i+1}|}} & khi \quad 2 \le i < k, \\ \Delta_k \cdot \dfrac{1.5}{\dfrac{0.5}{|a_{k-1}|} + \dfrac{1}{|a_k|}} & khi \quad i = k \end{cases} \quad (7)$$

trong đó: a_i là điểm giữa của A_i và $\Delta_i = \begin{cases} -1 & khi \quad a_i < 0 \\ 1 & khi \quad a_i > 0 \end{cases}$.

Bước 6: Dự báo cho tập Y theo quy tắc:

$$FY_t = T_i \text{ nếu } Y_t \in U_i, \ i = \overline{1, k}; \ t = \overline{2, n}.$$

Bước 7: Tính toán giá trị dự báo FX_t theo công thức:

$$FX_{t+1} = \left(\frac{FY_{t+1}}{100} + 1\right).X_t, \quad t = \overline{1, n}. \quad (8)$$

3 VÍ DỤ MINH HỌA VÀ MỘT SỐ ĐÁNH GIÁ

3.1 Ví dụ minh họa

Trong phần này dữ liệu tuyển sinh (Enrollment) của trường Đại học Alabama (1971-1992) được sử dụng để minh họa cho thuật toán đề nghị. Dữ liệu này đã được sử dụng trong nhiều nghiên cứu về mô hình FTS như Song and Chisson (1993), Chen (1996), Huarng (2001), Singh (2007), … Nó thường được lấy làm dữ liệu chuẩn để để so sánh hiệu quả của các mô hình FTS. Bảy bước của mô hình đề nghị được trình bày như sau:

Bước 1: Tính phần trăm biến đổi *Y* của chuỗi giữa hai năm liên tiếp, ta có Bảng 1.

Bảng 1: Phần trăm biến đổi Y giữa hai năm liên tiếp của dữ liệu Enrollment

Năm	Thực tế	Y	Năm	Thực tế	Y
1971	13055	−	1982	15433	−5,8274
1972	13563	3,8912	1983	15497	0,4147
1973	13867	2,2414	1984	15145	−2,2714
1974	14696	5,9782	1985	15163	0,1189
1975	15460	5,1987	1986	15984	5,4145
1976	15311	−0,9638	1987	16859	5,4742
1977	15603	1,9071	1988	18150	7,6576
1978	15861	1,6535	1989	18970	4,5179
1979	16807	5,9643	1990	19328	1,8872
1980	16919	0,6664	1991	19337	0,0466
1981	16388	−3,1385	1992	18876	−2,3840

Bước 2: Sử dụng thuật toán tìm số chùm thích hợp cho 21 phần tử của Y. Sau 12 bước lặp, chúng ta có dãy trọng tâm mới như sau:

3,8970 1,8955 5,9572 5,3882 -0,9637 1,8948 1,8948
5,9572 0,4366 -3,1368 -5,8274 0,4366 -2,3282 0,1586
5,3882 5,3882 7,6576 4,5140 1,8948 0,1586 -2,3282

Vì có 13 trọng tâm khác nhau nên chuỗi được chia thành 13 chùm.

Bước 3: Sử dụng thuật toán phân tích chùm 13-trung bình để xác định các phần tử và tính trọng tâm cho mỗi chùm ta được kết quả như Bảng 2.

Bảng 2: Các phần tử và trọng tâm của 13 chùm

Chùm	Phần tử	Trọng tâm (c_i)
1	{−5,8274}	−5,8274
2	{−3,1368}	−3,1385
3	{−2,2714, −2,3840}	−2,3277
4	{−0,9638}	−0,9638
5	{0,1189, 0,0466}	0,0827
6	{0,6664, 0,4147}	0,5405
7	{1,9071, 1,6535, 1,8872}	1,8159
8	{2,2414}	2,2414
9	{3,8912}	3,8912
10	{4,5179}	4,5179
11	{5,1987, 5,4145, 5,4742}	5,3625
12	{5,9782, 5,9643}	5,9713
13	{7,6576}	7,6576

Bước 4: Với $c_1 = -5.8274$, $c_2 = -3.1385$, $c_{12} = 5.9713$ và c_{13}=7.6576, ta xác định được tập nền cho dữ liệu Y như sau:

$$U=\left[c_1-\frac{c_2-c_1}{2};c_{13}+\frac{c_{13}-c_{12}}{2}\right]=[-7.1719;\ 8.5008].$$

Chia U thành 13 khoảng U_1, U_2, …, U_{13} được kết quả:

U_1=[−7.1719;−4.4830), U_2=[−4.4830;−2.7331), U_3=[−2.7331;−1.6457), U_4=[−1.6457;−0.4405), U_5=[−0.4405;0.3116), U_6=[0.3116;1.1782), U_7=[1.1782;2.0287), U_8=[2.0287;3.0663), U_9=[3.0663;4.2046), U_{10}=[4.2046;4.9402), U_{11}=[4.9402;5.6669), U_{12}=[5.6669;6.8144), U_{13}=[6.8144;8.5008].

Bước 5: Xác định các tập mờ A_i tương ứng với các khoảng U_i và tìm các điểm giữa a_i, $i=\overline{1,13}$. Sau đó, tính toán các giá trị T_i theo công thức (7). Các kết quả được trình bày tại Bảng 3.

Bảng 3: Các giá trị a_i và T_i

Chùm	Tập mờ	a_i	T_t	Chùm	Tập mờ	a_i	T_t
1	A_1	−5,8274	−4,8359	8	A_8	2,5475	2,3756
2	A_2	−3,6080	−3,3822	9	A_9	3,6354	3,4442
3	A_3	−2,1894	−1,8611	10	A_{10}	4,5724	4,4394
4	A_4	−1,0431	−0,2236	11	A_{11}	5,3035	5,2906
5	A_5	−0,0645	−0,1200	12	A_{12}	6,2407	6,2537
6	A_6	0,7449	0,2125	13	A_{13}	7,6576	7,1188
7	A_7	1,6035	1,3413				

Bước 6: Giải mờ cho tập Y.

Vì Y_2=3.8912 thuộc vào khoảng U_7 nên $FY_2=T_7$=1.3413. Tương tự, ta tính được FY_3, …, FY_{22}. Kết quả được trình bày tại Bảng 4.

Bảng 4: Kết quả mờ hóa dữ liệu Enrollment của thuật toán đề nghị

Thực tế	Y	FY	FX	Thực tế	Y	FY	FX
13055	-	-	-	15433	−5,8274	−4,8359	15595
13563	3,8912	3,4442	13505	15497	0,4147	0,2125	15466
13867	2,2414	2,3756	13885	15145	−2,2714	−1,8611	15209
14696	5,9782	6,2537	14734	15163	0,1189	−0,1200	15127
15460	5,1987	5,2906	15474	15984	5,4145	5,2906	15965
15311	−0,9638	−0,2236	15425	16859	5,4742	5,2906	16830
15603	1,9071	1,3413	15516	18150	7,6576	7,1188	18059
15861	1,6535	1,3413	15812	18970	4,5179	4,4394	18956
16807	5,9643	6,2537	16853	19328	1,8872	1,3413	19224
16919	0,6664	0,2125	16843	19337	0,0466	−0,1200	19305
16388	−3,1385	-3,3822	16347	18876	−2,3840	−1,8611	18977

Bước 7: Dự báo

Sử dụng công thức (8), ta tính được:

$$FX_2=\left(\frac{FY_2}{100}+1\right).X_1=13505.$$

Tương tự, ta tính được FX_3, …, FX_{22}. Các kết quả cũng được trình bày tại Bảng 4.

3.2 Một số đánh giá

Bên cạnh dữ liệu Enrollment, bài viết này sử dụng nhiều tập dữ liệu chuẩn khác nhau để so sánh mô hình đề nghị với các mô hình phổ biến khác. Đó là các tập dữ liệu Actual và NYSE. Đây là những tập dữ liệu phổ biến được sử dụng để đánh giá hiệu quả của các mô hình trong nhiều bài báo (Tai, 2018). 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 khác nhau, trong đó có mô hình đề nghị. Tập kiểm tra được sử dụng để đánh giá hiệu quả của việc mờ hóa. Dự báo từ số liệu gốc, số liệu mờ hóa cho mô hình ARIMA (ARIMAR, ARIMAP), dự báo từ số liệu gốc, số liệu mờ hóa cho mô hình AM (AMR, AMP), dự báo từ số liệu gốc, số liệu mờ hóa cho mô hình IFTS (IFTSR, IFTSP) để so sánh với số liệu thực của tập kiểm tra.

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	Chen	611,320	3,869	519885
	Singh	275,200	1,705	102719
	Heuristic	488,910	3,101	310840
	Chen-Hsu	259,590	1,633	104770
	AM	552,780	3,368	399917
	Mô hình đề nghị	**63,225**	**0,401**	**5857**
Actual	Chen	12,546	8,860	238,85
	Singh	6,859	4,911	62,537
	Heuristic	8,0540	5,576	98,544
	Chen-Hsu	10,940	8,190	164,160
	AM	10,092	6,832	151,310
	Mô hình đề nghị	**1,094**	**0,794**	**2,148**
NYSE	Chen	76,880	0,702	8257,224
	Singh	37,650	0,343	1818,369
	Heuristic	43,802	0,400	3016,226
	Chen-Hsu	42,328	0,384	2635,617
	AM	84,688	0,768	9558,289
	Mô hình đề nghị	**4,233**	**0,038**	**38,317**

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	ARIMAR	1988,600	10,42	5838400
	AMR	589,0860	3,592	481669,5
	IFTSR	548,710	2,855	390870
	ARIMAP	**299,09**	**1,575**	**149717**
	AMP	523,618	2,732	306607
	IFTSP	390,133	2,054	250537
Actual	ARIMAR	15,919	7,399	380,78
	AMR	15,766	7,908	404,69
	IFTSR	15,766	7,908	404,69
	ARIMAP	**13,932**	**6,639**	**333,32**
	AMP	35,865	16,912	1402,16
	IFTSP	35,865	16,912	1402,16
NYSE	ARIMA	180,612	1,581	40122,40
	AMR	217,474	1,912	52185,98
	IFTSR	217,474	1,912	52185,98
	ARIMAP	165,870	1,450	37399,24
	AMP	**145,144**	**1,277**	**24332,03**
	IFTSP	145,155	1,277	24335,26

Bảng 5 cho thấy trong giai đoạn mờ hóa của tập huấn luyện, mô hình đề nghị luôn cho kết quả tốt nhất với cả 3 bộ dữ liệu. Bảng 6 cho thấy với tập kiểm tra của cả ba bộ số liệu, khi sử dụng số liệu mờ hóa từ mô hình đề nghị, những mô hình dự báo đều cho kết quả tốt hơn mô hình gốc (ARIMAP, AMP, IFTSP), trong đó ARIMAP cho kết quả tốt nhất với Enrollment và AMP cho kết quả tốt nhất với Actual.

4 ÁP DỤNG TRONG DỰ BÁO ĐỈNH MẶN TỈNH TRÀ VINH

Đồng bằng sông Cửu Long nói chung và tỉnh Trà Vinh nói riêng là một trong những nơi chịu ảnh hưởng nặng nề của biến đổi khí hậu. Trong các ảnh hưởng đó, sự xâm nhập mặn được xem là nghiêm trọng nhất. Nó ảnh hưởng rất nhiều đến đời sống nhân dân, việc trồng trọt và nuôi thủy sản. Để hạn chế tác hại này, trước hết chúng ta phải có được dự báo đúng mức độ sự xâm nhập mặn, từ đó có biện pháp đối phó phù hợp. Các số liệu về đỉnh mặn là cơ sở khoa học không thể thiếu cho các biện pháp hiệu quả đối phó với sự xâm nhập mặn. Mặc dù được sự quan tâm của các nhà quản lý và các nhà khoa học, nhưng việc dự báo xâm nhập mặn của tỉnh Trà Vinh cũng như các địa phương khác trong cả nước còn nhiều hạn chế. Trong phần này, bài viết sử dụng dữ liệu quá khứ (Bảng 7) và mô hình đề nghị để dự báo đỉnh mặn tại hai trạm đo chính của tỉnh Trà Vinh là Cầu Quan và Trà Vinh.

Bảng 7: Số liệu đỉnh mặn (‰) tại hai trạm đo chính của tỉnh Trà Vinh

Năm	Cầu Quan	Trà Vinh	Năm	Cầu Quan	Trà Vinh
2002	6,3	7,9	2010	11,8	10,8
2003	7,9	11,3	2011	8,3	11,1
2004	10,6	8,3	2012	4,4	9,1
2005	10,9	10,7	2013	9,2	12,4
2006	9,7	9,0	2014	5,9	6,0
2007	8,9	9,5	2015	8,5	8,9
2008	10,0	9,9	2016	10,4	10,7
2009	6,3	9,9	2017	11,5	11,7

Việc thực hiện được chia thành hai giai đoạn như sau:

i) Đánh giá mô hình: Chia dữ liệu ban đầu thành tập huấn luyện và tập kiểm tra với tỉ lệ lần lượt là 80% và 20%. Mờ hóa dữ liệu tập huấn luyện bằng mô hình đề nghị, sau đó sử dụng các mô hình ARIMA và AM để dự báo trên số liệu gốc và số liệu mờ hóa cho tập kiểm tra. Kết quả thực hiện được so sánh cho cả hai tập huấn luyện và kiểm tra.

ii) Dự báo: Sử dụng toàn bộ dữ liệu để mờ hóa bằng mô hình đề nghị. Sau đó, dùng mô hình ARIMA trên bộ số liệu này để dự báo đỉnh mặn tại các trạm đo cho đến năm 2025.

Thực hiện i) trên tập huấn luyện và tập kiểm tra, ta lần lượt có Bảng 8 và Bảng 9

Bảng 8: Kết quả thực hiện cho tập huấn luyện của hai trạm đo

Trạm	Phương pháp	MAE	MAPE	MSE
Cầu Quan	ARIMAR	1,770	24,887	5,146
	AMR	4,360	62,285	19,096
	IFTSR	4,360	62,285	19,096
	ARIMAP	1,212	15,622	1,811
	AMP	3,435	55,084	12,714
	IFTSP	3,435	55,084	12,714
Trà Vinh	ARIMAR	0,791	8,575	1,061
	AMR	2,880	32,861	11,260
	IFTSR	2,880	32,861	11,260
	ARIMAP	1,051	11,692	1,702
	AMP	2,595	31,396	9,425
	IFTSP	2,758	37,498	12,025

Bảng 9: Kết quả thực hiện cho tập kiểm tra của hai trạm đo

Trạm	Phương pháp	MAE	MAPE	MSE
Cầu Quan	ARIMAR	2,537	23,659	8,343
	AMR	2,633	25,629	7,296
	IFTSR	2,633	25,629	7,296
	ARIMAP	1,332	11,957	3,830
	AMP	3,431	32,678	13,579
	IFTSP	3,431	32,678	13,579
Trà Vinh	ARIMAR	3,110	31,562	11,887
	AMR	7,433	69,526	60,896
	IFTSR	7,433	69,526	60,896
	ARIMAP	0,960	8,963	1,312
	AMP	6,508	60,419	48,557
	IFTSP	7,084	65,861	57,061

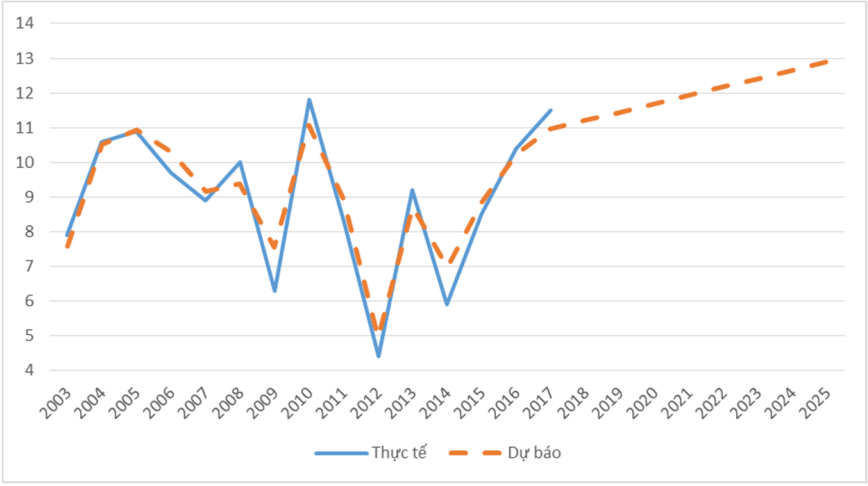
Bảng 8 và Bảng 9 cho thấy số liệu dự báo đỉnh mặn tại trạm đo Cầu Quan trên tập huấn luyện và tập kiểm tra theo mô hình đề nghị đều cho kết quả tốt nhất. Đối với trạm đo Trà Vinh, mặc dù số liệu dự báo trên tập huấn luyện của mô hình đề nghị không phải là tốt nhất nhưng trên tập kiểm tra thì nó đạt được kết này. Chính vì vậy, mô hình ARIMAP được chọn để dự báo cho tương lai.

Sử dụng toàn bộ dữ liệu để mờ hóa từ mô hình đề nghị, lấy số liệu nhận được từ kết quả này dự báo đỉnh mặn đến năm 2025 cho hai trạm bằng mô hình ARIMA, ta có được Bảng 10.

Bảng 10: Kết quả dự báo đỉnh mặn tại các trạm đo

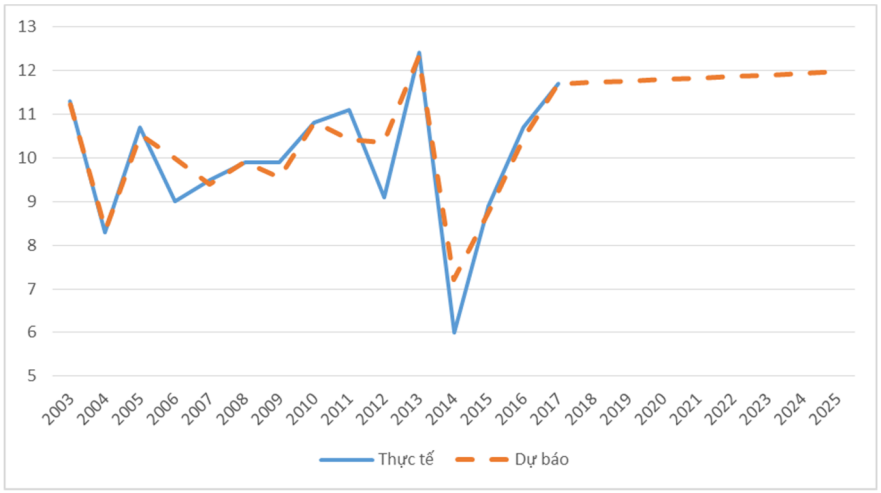
Năm	Cầu Quan	Trà Vinh
2018	11,205	11,727
2019	11,446	11,761
2020	11,687	11,795
2021	11,929	11,829
2022	12,170	11,863
2023	12,410	11,897
2024	12,652	11,931
2025	12,893	11,965

Bảng 10, Hình 1 và Hình 2 cho thấy dữ liệu đỉnh mặn ở cả hai trạm không có sự biến động nhiều, có xu hướng tăng trong những năm tiếp theo. Trong đó, đỉnh mặn tại trạm đo Cầu Quan được dự báo là tăng mạnh hơn.



Hình 1: Đồ thị đỉnh mặn thực tế và dự báo tại trạm Cầu Quan

Mặc dù kết quả thực hiện từ tập kiểm tra (Bảng 9) cho thấy mô hình đề nghị có độ tin cậy tương đối tốt, tuy nhiên chúng ta cần phải so sánh từ số liệu thực tế tương lai mới có thể đánh giá được hiệu quả thực sự của mô hình đề nghị.



Hình 2: Đồ thị đỉnh mặn thực tế và dự báo tại trạm Trà Vinh

5 KẾT LUẬN

Bài báo này đề nghị một mô hình mờ hóa dữ liệu chuỗi thời gian dựa trên cải tiến từ những bước chính của các mô hình hiện tại. Các bước của mô hình đề nghị được trình bày cụ thể về mặt lý thuyết và được minh họa chi tiết trên số liệu thực. Thông qua các tham số MSE, MAE và MAPE, thực hiện cho nhiều bộ số liệu chuẩn, mô hình đề nghị đã cho kết quả tốt hơn các mô hình đang được sử dụng phổ biến. Áp dụng thực tế của mô hình đề nghị có thể thực hiện tương tự trong dự báo cho rất nhiều vấn đề thực tế khác. Điều này đã minh chứng cho ý nghĩa của vấn đề được nghiên cứu. Trong tương lai, chúng tôi tiếp tục thử nghiệm mô hình đề nghị để dự báo cho nhiều vấn đề thực tế đang đòi hỏi cấp thiết.

TÀI LIỆU THAM KHẢO

Abbasov, A. and Mamedova, M., 2003. Application of fuzzy time series to population forecasting. Vienna University of Technology. 1: 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.

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.

Eren, B., Vedide, R., Uslu, U., and Erol, E., 2014. A modied genetic algorithm for forecasting fuzzy time series. Applied Intelligence, 41: 453–463.

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

Huarng, K. and Yu, T., 2006. Ratio-based lengths of intervals toimprove fuzzy time series forecasting. IEEE Trans Syst ManCybern-Part B: Cybern. 36: 328–340.

Oliveira, D. J and Ludermir, T. B., 2014. A distributed PSO-ARIMA-SVR hybrid system for time series forecasting. In 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC). 3867–3872.

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 Thao, N.T., 2018. Similar coefficient of cluster for discrete elements. The Indian Journal of Statistics, 80(1): 19 – 36.

Tai V.V., 2018. An improved fuzzy time series forecasting model using variations of data. Fuzzy Optimization and Decision Making DOI: 10.1007/s10700-018-9290-7

Zhiqiang, Z. and Qiong, Z., 2012. Fuzzy time series forecasting based on k-means clustering. Open Journal of Applied Sciences. 25:100–103.