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.2021.146

XÂY DỰNG MÔ HÌNH DỰ BÁO CHUỖI THỜI GIAN CHO DỮ LIỆU KHOẢNG DỰA VÀO BÀI TOÁN PHÂN TÍCH CHÙM

Võ Văn Tài*, Trần Ngọc Nhã Hân và Từ Ngọc Thảo
Khoa Khoa học Tự nhiên, Trường Đại học Cần Thơ
**Người chịu trách nhiệm về bài viết: Võ Văn tài (email: vvtai@ctu.edu.vn)*

Thông tin chung:
Ngày nhận bài: 25/04/2021
Ngày nhận bài sửa: 05/06/2021
Ngày duyệt đăng: 29/10/2021

Title:
Building the time series forecasting model for interval data based on cluster analysis problem

Từ khóa:
Chuỗi thời gian, dữ liệu khoảng, dự báo, phân tích chùm

Keywords:
Cluster analysis, forecasting model, interval data, time series

ABSTRACT

This study is to propose using the overlap distance to evaluate the similarity of two intervals. Based on the distance and cluster analysis problem for discrete elements, the study has built the forecasting model for time series with interval data. The proposed model is presented in detailed steps, and is illustrated by the numerical examples. It is also applied in forecasting the salty peak at stations of the main rivers in Ca Mau province. The proposed model can be quickly implemented by a procedure established in the software Matlab.

TÓM TẮT

Nghiên cứu này đề xuất sử dụng khoảng cách chồng lấp trong đánh giá sự tương tự của hai khoảng. Dựa trên khoảng cách này và bài toán phân tích chùm cho các phần tử rời rạc, mô hình dự báo cho chuỗi thời gian với dữ liệu khoảng được xây dựng trong nghiên cứu. Mô hình đề nghị đã trình bày cụ thể các bước và được minh hoạ bởi một ví dụ số. Nó cũng áp dụng trong dự báo đỉnh mặn tại các trạm đo trên các con sông chính của tỉnh Cà Mau. Mô hình đề nghị có thể thực hiện nhanh chóng bởi một chương trình được thiết lập trên phần mềm Matlab.

1. GIỚI THIỆU

Dự báo là sự tiên đoán những vấn đề sẽ xảy ra trong tương lai dựa trên một cơ sở nào đó. Đây là một vấn đề luôn nhận được sự quan tâm của nhiều nhà khoa học, nhà quản lý bởi vì nó có một vai trò rất quan trọng trong thực tế. Tuy nhiên cho đến nay, dự báo vẫn là bài toán chưa có lời giải cuối cùng (Abbasov & Mamedova, 2003; Tai, 2019). Trong thống kê, dựa trên dữ liệu quá khứ, mô hình để dự báo cho tương lai được thiết lập. Đối với dữ liệu dạng chuỗi, hồi quy và chuỗi thời gian là hai mô hình được áp dụng phổ biến ngày nay.

Khi xây dựng mô hình hồi quy, dữ liệu phải thoả những điều kiện mà thực tế nó rất khó đáp ứng. Vì thế hiệu quả dự báo khi sử dụng mô hình hồi quy hiện nay còn nhiều hạn chế (Abreu et al., 2013; Chen, 1996). Mô hình chuỗi thời gian phát triển dựa trên sự đặc thù của dữ liệu dạng này nhằm khắc phục những hạn chế của mô hình hồi quy. Nó được phát triển theo hai hướng: Chuỗi thời gian mờ và chuỗi thời gian không mờ. Chuỗi thời gian không mờ được sử dụng phổ biến ngày nay với mô hình tiêu biểu là tự hồi quy tích hợp trung bình trượt (ARIMA) mà các tham số được tìm bằng phương pháp Box-Jenkins (Box & Jenkins, 1970). Tuy nhiên, thực tế cho thấy chuỗi thời gian không mờ cũng chỉ hiệu quả khi dữ liệu có tính dừng (Aladag et al., 2012; Chen & Hsu, 2004). Thực tế dữ liệu rất khó thoả mãn điều kiện này, nên dự báo của các mô hình chuỗi thời gian không mờ thường không tốt. Hạn chế chính của các mô hình chuỗi thời gian không mờ

là việc dự báo dựa trên sự liên kết của các phần tử bằng một biểu thức toán học mà không có sự linh động theo mức độ của ngôn ngữ tuỳ thuộc vào từng chuỗi (Eren et al., 2014). Để khắc phục điều này chuỗi thời gian mờ đã được đề xuất. Chuỗi thời gian mờ cũng được phát triển theo hai hướng (i) mờ hoá dữ liệu để tạo ra sự liên kết của các phần tử, sau đó sử dụng một mô hình nào đó để dự báo cho tương lai từ các số liệu đã mờ hoá và (ii) xây dựng mô hình trực tiếp từ dữ liệu nguồn để dự báo cho tương lai. Theo hướng (i) có rất nhiều nhà thống kê quan tâm với rất nhiều công trình được công bố liên tục (Huarng, 2001; Singh, 2007; Song & Chissom, 1993). Theo hướng (ii) hiện có rất ít các mô hình được đề xuất. Theo hướng này có hai mô hình tiêu biểu được đề cập trong các ứng dụng gần đây. Abbasov and Manedova (2003) đề xuất mô hình dự báo dân số nước Áo. Mô hình này xây dựng tập nền là giá trị lớn nhất và nhỏ nhất của dữ liệu biến đổi hai thời gian liên tiếp và mối quan hệ mờ của thời gian tương lai với quá khứ theo cấp độ ngôn ngữ của thang đo Likert (7 cấp độ). Nguyên tắc dự báo của mô hình này là lấy dữ liệu gần nhất cộng cho sự biến đổi trung bình của quá khứ (dương hoặc âm) mà nó được thiết lập từ hai bước trên. Mô hình này có hiệu quả cho tập dữ liệu dân số được xét, nhưng không hiệu quả cho các tập dữ liệu khác vì các tham tham số trong mô hình chưa được xác định hiệu quả. Cải tiến mô hình của Abbasov and Manedova (2003), Tai (2019) đã đề nghị một mô hình chuỗi thời gian mờ. Mô hình này không những khảo sát các tham số trong mô hình của Abbasov and Manedova (2003) để tìm tham số tối ưu mà còn đưa ra một quy tắc dự báo mới. Nó cũng đã chứng minh sự hiệu quả qua nhiều tập dữ liệu đối chứng.

Các mô hình ở trên chỉ xây dựng cho dữ liệu chuỗi thời gian dạng điểm. Trong thực tế chúng ta lưu trữ nhiều chuỗi thời gian dạng khoảng như nhiệt độ, lượng mưa, huyết áp của một người (Tai et al., 2020). Hơn nữa khi dự báo chúng ta cũng muốn có kết quả dạng khoảng tin cậy thay vì dạng điểm. Thực tế này đòi hỏi chúng ta phải dự báo dạng khoảng cho chuỗi thời gian. Theo sự hiểu biết của chúng tôi, vấn đề này hầu như chưa được quan tâm. Trong trường hợp này, các giá trị hai đầu mút của các khoảng như hai chuỗi độc lập có thể sử dụng, sau đó sử dụng các mô hình đã có để dự báo. Vì đánh giá sự tương tự của các khoảng có những độ đo riêng so với các phần tử rời rạc nên phương pháp này thường gặp nhiều hạn chế. Bài viết này đề nghị độ đo được gọi là khoảng cách chồng lấp để đánh giá sự tương tự của các khoảng, sau đó dựa vào bài toán phân tích chùm để đề nghị mô hình dự báo cho dữ liệu khoảng. Cụ thể nghiên cứu sử dụng tập nền là sự biến đổi của hai khoảng thời gian liên tiếp được chuẩn hoá về thang đo 100. Sử dụng khoảng cách chồng lấp chia tập nền thành các chùm để từ đó xây dựng một nguyên tắc mới trong dự báo.

Phần tiếp theo của bài viết được trình bày như sau. Phần 2 trình bày những vấn đề liên quan. Mô hình đề nghị được trình bày trong Phần 3. Ví dụ minh hoạ và áp dụng của mô hình đề nghị được cho trong phần 4. Phần 5 là kết luận của bài viết.

2. MỘT SỐ VẤN ĐỀ LIÊN QUAN

2.1. Một số phép toán trên dữ liệu khoảng

Định nghĩa 1. Cho hai khoảng $A=\left[a,\hat{a}\right], B=\left[b,\hat{b}\right]$. Khi đó ta có các phép toán sau:

$$A+B=\left[a+b,\hat{a}+\hat{b}\right], A-B=\left[a-\hat{b},\hat{a}-b\right].$$

$$AB=\left[\min\left\{ab,a\hat{b},\hat{a}b,\hat{a}\hat{b}\right\},\max\left\{ab,a\hat{b},\hat{a}b,\hat{a}\hat{b}\right\}\right].$$

$$\frac{A}{B}=a.(1/b) \text{ với } \frac{1}{B}=\left[\frac{1}{b};\frac{1}{\hat{b}}\right].$$

2.2. Khoảng cách của các khoảng

Cho 2 khoảng có p - chiều *A* và *B:*

$$A=\left(a^1,a^2,...,a^p\right)=\left(\left[a_1,\hat{a}_1\right],\left[a_2,\hat{a}_2\right],...,\left[a_p,\hat{a}_p\right]\right),$$

$$B=\left(b^1,b^2,...,b^p\right)=\left(\left[b_1,\hat{b}_1\right],\left[b_2,\hat{b}_2\right],...,\left[b_p,\hat{b}_p\right]\right).$$

Định nghĩa 2. Các khoảng cách phổ biến giữa *A* và *B*:

Khoảng cách Euclide:

$$d_E(A,B)=\left[\sum_{i=1}^{p}[(a_i-b_i)^2+(\hat{a}_i-\hat{b}_i)^2]\right]^{\frac{1}{2}}.$$

Khoảng cách Hausdoff:

$$d_H\left(A,B\right)=\max_{a^i\in A}\left\{\min_{b^i\in B}\left\{d_E\left(a^i,b^i\right)\right\}\right\}.$$

Khoảng cách City-block:

$$d_C\left(A,B\right)=\sum_{i=1}^{p}(|a_i-b_i|+|\hat{a}_i-\hat{b}_i|).$$

Định nghĩa 3 Khoảng cách chồng lấp:

$$d_O(A,B) = D(A,B).\left(1-\frac{O(A,B)}{2r_a+1}\right), \quad (1)$$

với

$$r_a = \frac{1}{p}\sum_{i=1}^{p}|a_i - \hat{a}_i|,$$

$O(A,B)$ là độ chồng lấp giữa hai khoảng *A* và *B*.

$D(A,B)$ là khoảng cách Hausdorff.

Trong trường hợp *A* và *B* là hai khoảng một chiều: $A=[a,\hat{a}], B=\left[b,\hat{b}\right]$, đặt

$$c_a = \frac{a+\hat{a}}{2}, \; c_b = \frac{b+\hat{b}}{2}, \; r_a = \frac{a-\hat{a}}{2}, r_b = \frac{b-\hat{b}}{2}.$$

Khi đó sự chồng lấp giữa *A* và *B* được xem xét trong 5 trường hợp sau:

(i) Nếu *A* nằm hoàn toàn trong *B*: $\|c_a - c_b\| \le r_b - r_a$ thì $O(A,B) = 2r_a + 1.$

(ii) Nếu *B* nằm hoàn toàn trong *A*: $\|c_a - c_b\| \le r_a - r_b$ thì $O(A,B) = 2r_b.$

(iii) Nếu *B* chồng lấp *A* và nằm bên trái của *A*: $r_a = r_b = 0$ thì $O(A,B) = 0.$

(iv) Nếu *B* chồng lấp *A* và nằm bên phải của *A*: $\|r_a - r_b\| < \|c_a - c_b\| < r_a + r_b$ thì $O(A,B) = r_a + r_b - |c_a - c_b|.$

(v) Nếu *B* không chồng lấp với *A* và nằm bên trái hoặc nằm bên phải so với *A*: $|c_a - c_b| \ge r_a + r_b$ thì $O(A,B) = |c_a - c_b| - (r_a - r_b).$

Do đó công thức (1) được cụ thể như sau:

$$d_O(A,B) = \begin{cases} 0 & (i) \\ \left(|c_a - c_b| + r_a - r_b\right)\left(1 - \dfrac{2r_b}{2r_a+1}\right) & (ii) \\ |c_a - c_b| & (iii) \\ \left(|c_a - c_b| + r_a - r_b\right)\left(1 - \dfrac{r_a + r_b - |c_a - c_b|}{2r_a+1}\right) & (iv) \\ \left(|c_a - c_b| + r_a - r_b\right)\left(1 - \dfrac{|c_a - c_b| - (r_a + r_b)}{2r_a+1}\right) & (v) \end{cases} \quad (2)$$

2.3. Tham số đánh giá mô hình

Cho tập dữ liệu X_i và giá trị dự báo tương ứng $\hat{X}_i$, khi đó để đánh giá mô hình dự báo ta có thể sử dụng tham số sai số phần trăm tuyệt đối trung bình:

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

Cho một mô hình được xây dựng, tham số *MAPE* càng nhỏ thì mô hình xây dựng sẽ càng tốt. Khi xây dựng mô hình dự báo khoảng thì tham số *MAPE* của mô hình sẽ được tính là tổng MAPE của hai biên.

3. THUẬT TOÁN ĐỀ NGHỊ

Cho chuỗi số liệu $X_i = [a_i, b_i]$ tương ứng với thời gian $t_i, i = \overline{1,N}.$ Thuật toán đề nghị bao gồm 7 bước sau:

***Bước 1*:** Chuẩn hóa dữ liệu khoảng trên thang đo 100:

$$Y_i = \left[\hat{a}_i, \hat{b}_i\right], i = \overline{1,N},$$

trong đó $\hat{a}_i = \dfrac{a_i.100}{\max\limits_{1\le i\le N}\{a_i\}}, \hat{b}_i = \dfrac{b_i.100}{\max\limits_{1\le i\le N}\{b_i\}}.$

Bước 2: Tính sự biến đổi giữa hai thời gian liên tiếp

$$Z_i = [l_i, h_i] = Y_{i+1} - Y_i, i = \overline{1,N-1}$$

với $l_i = \hat{a}_{i+1} - \hat{b}_i$ và $h_i = \hat{b}_{i+1} - \hat{a}_i$.

Bước 3: Tìm số chùm thích hợp cho $Z = \{Z_i, i = \overline{1, N-1}\}$ bởi thuật toán ***FCAI*** (*Fuzzy Clustering Algorithm for Intervals)* như sau:

Bước 3.1: Thiết lập dãy trọng tâm $V^{(0)} = \left\{v_1^{(0)}, v_2^{(0)}, ..., v_{N-1}^{(0)}\right\} = \left\{Z_1, Z_2, ..., Z_{N-1}\right\}$ và lấy giá trị $\varepsilon = 0,0001$.

Bước 3.2: Xây dựng dãy trọng tâm mới được thiết lập từ công thức

$$v_i^{(t+1)} = \sum_{j=1}^{N-1} \frac{f\left(v_i^{(t)}, v_j^{(t)}\right)}{\sum_{k=1}^{N-1} f\left(v_i^{(t)}, v_k^{(t)}\right)} v_j^{(t)}; i = \overline{1, N-1},$$

trong đó

$$f\left(v_i^{(t)}, v_j^{(t)}\right) = \begin{cases} \exp\left[-d\left(v_i^{(t)}, v_j^{(t)}\right)/\sigma\right] & \text{khi } d\left(v_i^{(t)}, v_j^{(t)}\right) \le \mu\alpha_{\text{ij}}(t), \\ 0 \text{ khi } d\left(v_i^{(t)}, v_j^{(t)}\right) \le \mu\alpha_{\text{ij}}(t), \end{cases}$$

với

$$\alpha_{\text{ij}}(0) = 1, \alpha_{\text{ij}}(t) = \frac{\alpha_{\text{ij}}(t-1)}{1 + \alpha_{\text{ij}}(t-1) f\left(v_i^{(t-1)}, v_j^{(t-1)}\right)}, t \ge 1,$$

$$\mu = \frac{1}{\binom{2}{N-1}} \sum_{i<j} d_O\left(v_i^{(0)}, v_j^{(0)}\right), \sigma = \sqrt{\frac{1}{\binom{2}{N-1}} \sum_{i<j} \left[d_O\left(v_i^{(0)}, v_j^{(0)}\right) - \mu\right]^2}.$$

Bước 3.3: Tính $v = \left\|v^{(t)} - v^{(t-1)}\right\|$. Lặp lại Bước 3.2 đến khi $v < \varepsilon$.

Bước 3.4: Tìm số phần tử của $v^{(t)}$. Nếu có c phần tử trong $v^{(t)}$ thì ta chia *Z* thành *c* chùm.

Bước 4. Bắt đầu với ma trận sau:

$$U^{(0)} = \left[\mu_{ik}^{(0)}\right]_{c \times N-1},$$

với $\mu_{ik} = 1$ nếu khoảng thứ k thuộc chùm i, ngược lại $\mu_{ik} = 0$.

Bước 5: Xây dựng dãy ma trận phân vùng $U^{(t)}$ thông qua thuật toán ***IFCM*** *(Interval Fuzzy c-Means)* sau**:**

Bước 5.1: Tính phần tử đại diện cho từng chùm:

$$\text{w}_i = \frac{\sum_{k=1}^{N-1} (\mu_{ik})^2 z_k}{\sum_{k=1}^{N-1} (\mu_{ik})^2}, \mu_{ik} \in U^{(0)}, 1 \le i, j \le c, 1 \le k \le N-1.$$

Bước 5.2: Tính các bình phương khoảng cách chồng lấp:

$$D_{ik}^2 = d_O^2\left(z_k, \text{w}_i\right).$$

Bước 5.3: Thiết lập ma trận phân vùng mới $U^{(t)}$ với các phần tử được tính như sau:

- Nếu $D_{ik} > 0$ với $i = \overline{1, c}$ thì

$$\mu_{ik}^{(t)} = \frac{1}{\sum_{j=1}^{c} \left(\frac{D_{ik}}{D_{jk}}\right)^2}.$$

- Nếu tồn tại một số hữu hạn điểm $D_{ik} = 0$ thì $\mu_{ik}^{(t)} = 0$ và tại các điểm còn lại $D_{ik} > 0$ thì $\mu_{ik}^{(t)}$ là ngẫu nhiên sao cho $\sum_{i=1}^{c} \mu_{ik}^{(t)} = 1$.

Bước 5.4: Tính $U = \left\| f\left(U^{(t)}\right) - f\left(U^{(t-1)}\right)\right\|$.

với $f\left(U^{(t)}\right) = \sum_{i=1}^{c} \sum_{k=1}^{N} \mu_{ik}^{(t)} . d_O(\text{w}_i, \text{z}_k)$.

Lặp lại các Bước 5.1, 5.2 và 5.3 cho đến khi $U < \varepsilon$.

Bước 6: Tính khoảng mờ

$$F\left(Z_i\right)=\left[F(l_i),F(h_i)\right]=\sum_{j=1}^{c}\mu^{(t)}{}_{\text{ij}}.v_j,\left(i=\overline{1,N-1}\right)$$

và

$$m=\frac{1}{N-1}\sum_{i=1}^{N-1}F\left(Z_i\right),$$

trong đó $F(l_i)=\sum_{j=1}^{c}\mu_{\text{ij}}.v_j^l$ và

$$F(h_i)=\sum_{j=1}^{c}\mu_{\text{ij}}.v_j^h, i=\overline{1,N-1}.$$

Bước 7: Mờ hóa dữ liệu quá khứ và dự báo cho tương lai:

$$F\left(X_i\right)=\frac{\left(F\left(Z_{i-1}\right)+Y_{i-1}\right).\max\limits_{1\le i\le N}\left\{X_i\right\}}{100}, i=\overline{2,N}.$$

$$F\left(X_{N+1}\right)=\frac{\left(Y_N+m\right).\max\limits_{1\le i\le N}\left\{X_i\right\}}{100}.$$

4. VÍ DỤ MINH HỌA VÀ ÁP DỤNG

4.1. Ví dụ minh hoạ thuật toán

Để minh họa cho thuật toán đề nghị, bộ số liệu về nhiệt độ ở Hà Nội thời gian từ ngày 26/4/2020 đến ngày 14/5/2020 được sử dụng. Tập dữ liệu này được cho trong Cột *X* của Bảng 1.

Bảng 1. Thời tiết Hà Nội từ ngày 26/4/2020 đến ngày 13/5/2020 và các giá trị được tính từ mô hình đề nghị

Ngày	*X*	*Y*	*Z*	*FZ*	*FX*
26/04	[18;23]	[64,29;63,89][67,86;83,33]	[3,97;19,04]	[-0,41;2,34]	-
27/04	[19;30]	[71,43;80,56]	[-11,90;12,69]	[-1,14;0,62]	[17,88;23,84]
28/04	[20;29]	[75;80,56]	[-5,55;9,12]	[-1,14;0,62]	[18,67;30,22]
29/04	[21;29]	[78,57;80,56]	[-1,98;5,55]	[-1,14;0,61]	[19,67;29,22]
30/04	[22;29]	[82,15;77,78]	[1,58;-0,79]	[-0,39;2,39]	[20,67;29,22]
01/05	[23;28]	[85,71;88,89]	[7,93;6,74]	[-1,04;0,86]	[21,88;29,86]
02/05	[24;32]	[92,86;88,89]	[3,96;3,17]	[-0,68;1,70]	[22,70;28,31]
03/05	[26;32]	[92,86;97,22]	[3,96;4,36]	[-0,94;1,10]	[23,80;32,61]
04/05	[26;35]	[96,43;100]	[-0,79;7,14]	[-1,04;0,86]	[25,73;32,39]
05/05	[27;36]	[100;97,22]	[0,00;0,79]	[-0,82;1,38]	[25,70;35,31]
06/05	[28;35]	[100;97,22]	[2,77;-2,77]	[-0,52;2,08]	[26,76;36,50]
07/05	[28;35]	[96,43;97,22]	[-0,79;-2,77]	[-0,52;2,08]	[27,85;35,75]
08/05	[27;35]	[92,86;97,22]	[-4,36;0,79]	[-0,37;2,43]	[27,85;35,75]
09/05	[26;35]	[92,86;94,44]	[-4,36;1,58]	[-1,04;0,86]	[26,89;35,87]
10/05	[26;34]	[92,86;94,44]	[-1,58;1,58]	[-0,38;2,42]	[25,70;35,31]
11/05	[26;34]	[89,29;80,56]	[-5,15;-12,30]	[-0,38;2,42]	[25,89;34,87]
12/05	[25;29]	[92,86;91,67]	[12,30;2,38]	[-1,15;0,61]	[25,89;34,87]
13/05	[26;33]				[24,67;29,22]

Các bước của thuật toán lần lượt được trình bày như sau:

Bước 1: Chuẩn hóa số liệu theo thang đo 100, ta được Cột *Y* của Bảng 1.

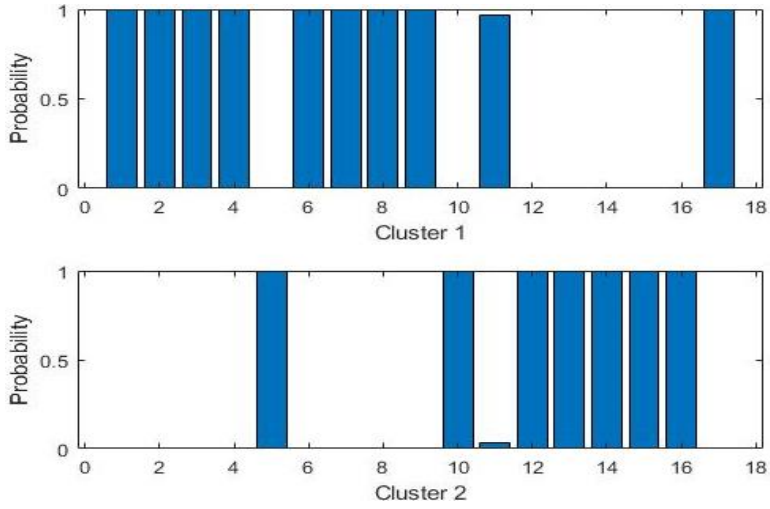
Bước 2: Tính sự biến đổi giữa hai thời gian liên tiếp, ta nhận được Cột *Z* của Bảng 1.

Bước 3: Sau khi chạy thuật toán ***FCAI,*** dữ liệu Z hội tụ về hai trọng tâm:

$$v_1=\left[v_1^l;v_1^h\right]=\left[1,47;6,77\right]$$

$$v_2=\left[v_2^l;v_2^h\right]=\left[-2,09;-1,59\right]$$

nên nó được chia thành hai chùm như minh hoạ bởi Hình 1.



Hình 1. Sự hội tụ của Z thành 2 chùm

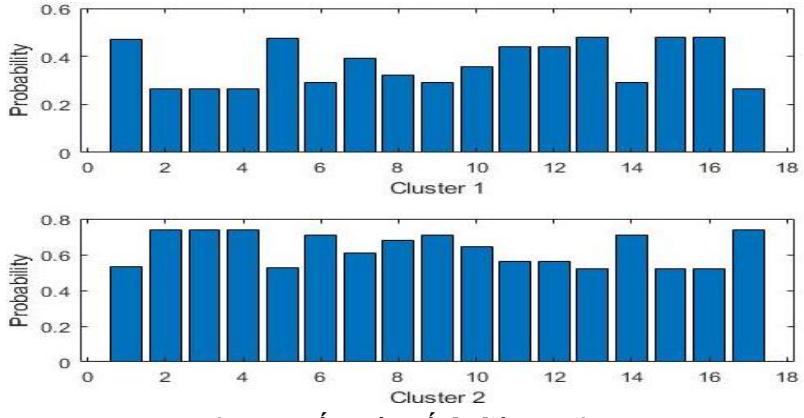
Bước 4: Từ kết quả của Bước 3, ta có ma trận phân vùng ban đầu như sau:

$$U^{(0)} = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

Bước 5: Xây dựng dãy ma trận phân vùng $U^{(t)}$ theo thuật toán ***IFCM*.** Qua 24 vòng lặp ta thu được ma trận như sau:

$$U^{(25)} = \begin{bmatrix} 0,47 & 0,26 & 0,26 & 0,26 & 0,48 & 0.29 & 0,39 & 0,32 & 0,29 & 0,36 & 0,44 & 0,44 & 0,48 & 0,29 & 0,48 & 0,48 & 0,26 \\ 0,53 & 0,74 & 0,74 & 0,74 & 0,52 & 0.71 & 0,61 & 0,68 & 0,71 & 0,64 & 0,56 & 0,56 & 0,52 & 0,71 & 0,52 & 0,52 & 0,74 \end{bmatrix}$$

Kết quả trên được minh hoạ bởi Hình 2.



Hình 2. Xác suất phân bố dữ liệu vào 2 chùm

Bước 6: Tính khoảng mờ

$$F\left(Z_1\right) = \left[F(l_1), F(h_1)\right] = \sum_{j=1}^{2} \mu_{1j}.v_j ,$$
$$= \left[-0,41; 2,34\right]$$

trong đó

$$F(l_1) = \sum_{j=1}^{2} \mu_{1j}.v_j^l = 0,47.1,46 + 0,52.(-2,08)$$
$$= -0,41 .$$

$$F(h_1) = \sum_{j=1}^{2} \mu_{1j}.v_j^h = 0,47.6,77 + 0,52.(-1,58) = 2,34$$

.

Tương tự cho các giá trị tiếp theo cho đến khi hết các giá trị của chuỗi, ta nhận được Cột $F(Z)$ của Bảng 1. Khi đó ta tính được

$$m = \frac{1}{17}\sum_{i=1}^{17} F(Z_i) = [-0,77; 1,49].$$

Bước 7: Mờ hóa dữ liệu và dự báo giá trị tương lai tiếp theo:

$$F\left(X_2^l\right) = \frac{\left(F\left(Z_1^l\right) + Y_1^l\right).\max\limits_{1\le i\le 17}\left\{X_i^l\right\}}{100} = \frac{(-0,41 + 64,29).28}{100} = 17,88$$

$$F\left(X_2^h\right) = \frac{\left(F\left(Z_1^h\right) + Y_1^h\right).\max\limits_{1\le i\le 17}\left\{X_i^h\right\}}{100} = \frac{(2,34 + 63,89).36}{100} = 25,84$$

Tương tự cho các giá trị còn lại đến dữ liệu thứ 19

$$F\left(X_{19}^l\right) = \frac{\left(Y_{18}^l + m^l\right).\max\limits_{1\le i\le 17}\left\{X_i^l\right\}}{100} = \frac{(92,86 - 0,77).28}{100} = 25,78$$

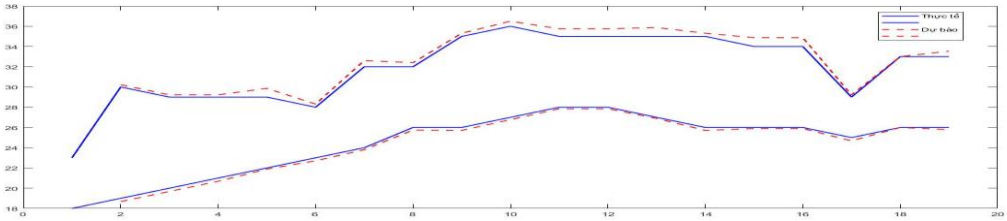
.

$$F\left(X_{19}^h\right) = \frac{\left(Y_{18}^h + m^h\right).\max\limits_{1\le i\le 17}\left\{X_i^h\right\}}{100} = \frac{(91,67 + 1,49).36}{100} = 33,53$$

.

Vậy giá trị dự báo cho ngày tiếp theo là $F(X_{19}) = [25,78; 33,54]$.

Các kết quả thực hiện được cho bởi Cột $F(X)$ của Bảng 1. Khoảng thực tế và dự báo được thể hiện bởi Hình 3.



Hình 3. Đồ thị cho số liệu thực tế và dự báo thời tiết Hà Nội từ ngày 26/4/2020 đến 14/5/2020

Tính tham số MAPE của mô hình nhận được kết quả như sau:

$$MAPE^L = \frac{100}{17}\sum_{i=1}^{17}\frac{\left|\hat{X}_i^L - X_i^L\right|}{X_i^L} = 0,94$$

,

$$MAPE^H = \frac{100}{17}\sum_{i=1}^{17}\frac{\left|\hat{X}_i^H - X_i^H\right|}{X_i^H} = 1,58$$

$$MAPE = MAPE^L + MAPE^H = 0,94 + 1,58 = 2,52$$

.

Hình 3 cũng như tham số MAPE cho thấy mô hình thực hiện rất tốt khi giá trị dự báo rất gần với giá trị thực tế.

4.2. Áp dụng

Phần này sử dụng mô hình đề nghị để dự báo khoảng tin cậy về đỉnh mặn (‰) tại các trạm đo Gành Hào và Cửa lớn của tỉnh Cà Mau. Số liệu thực hiện lần lượt được cho bởi Bảng 2.

Bảng 2. Số liệu đỉnh mặn tại hai trạm Cửa lớn và Gành Hào giai đoạn 2000-2017

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

Bảng 3. Khoảng tin cậy 95% cho số liệu đỉnh mặn tại hai trạm đo

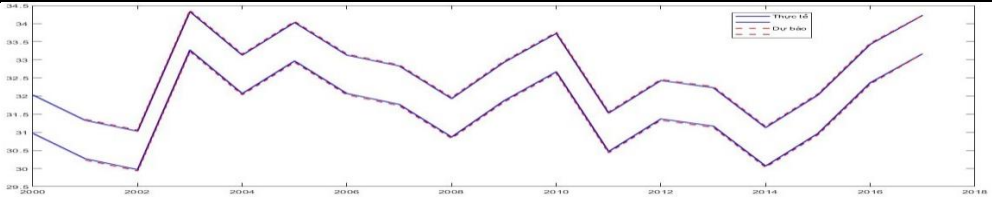
Năm	Gành Hào	Cửa Lớn	Năm	Gành Hào	Cửa Lớn
2000	(30,97; 32,03)	(28,08; 31,12)	2009	(31,87; 32,93)	(26,78; 29,82)
2001	(30,27; 31,33)	(27,88; 30,92)	2010	(32,67; 33,73)	(35,60; 38,63)
2002	(29,97; 31,03)	(32,88; 35,92)	2011	(30,47; 31,53)	(26,88; 29,92)
2003	(33,27; 34,33)	(33,58; 36,62)	2012	(31,37; 32,43)	(25,78; 28,82)
2004	(32,07; 33,13)	(32,78; 35,82)	2013	(31,17; 32,23)	(31,58; 34,62)
2005	(32,97; 34,03)	(34,58; 37,62)	2014	(30,07; 31,13)	(29,78; 32,82)
2006	(32,07; 33,13)	(30,08; 33,12)	2015	(30,97; 32,03)	(31,58; 34,62)
2007	(31,77; 32,83)	(31,38; 34,42)	2016	(32,37; 33,43)	(34,38; 37,42)
2008	(30,87; 31,93)	(29,98; 33,02)	2017	(33,17; 34,23)	(34,98; 38,02)

Từ Bảng 2, nghiên cứu tiến hành ước lượng khoảng tin cậy 95% cho số liệu ở hai trạm. Kết quả ước lượng được cho bởi Bảng 3.

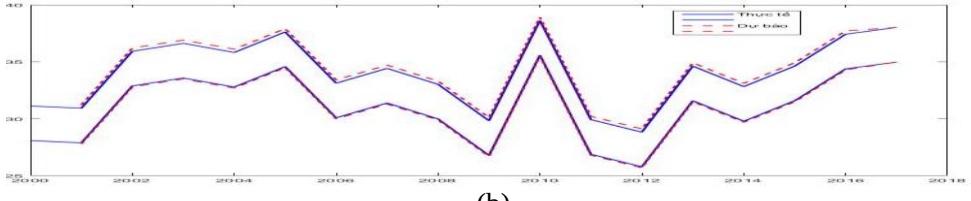
Khi sử dụng số liệu Bảng 3 và áp dụng mô hình đề nghị, ta có Bảng 4 và Hình 3 như sau:

Bảng 4. Số liệu mờ hóa của mô hình đề nghị đỉnh mặn tại hai trạm đo

Năm	Gành Hào	Cửa Lớn	Năm	Gành Hào	Cửa Lớn
2000	(30,93;32,06)	(27,99;31,41)	2009	(31,83;32,96)	(26,69;30,11)
2001	(30,23;31,35)	(27,79;31,21)	2010	(32,63;33,76)	(35,49;38,91)
2002	(29,93;31,06)	(32,79;36,21)	2011	(30,43;31,56)	(26,79;30,21)
2003	(33,23;34,35)	(33,49;36,91)	2012	(31,33;32,46)	(25,69;29,11)
2004	(32,03;33,16)	(32,69;36,11)	2013	(31,13;32,36)	(31,49;34,91)
2005	(32,93;34,06)	(34,49;37,91)	2014	(30,03;31,16)	(29,69;33,11)
2006	(32,03;33,16)	(29,99;33,31)	2015	(30,93;32,06)	(31,49;34,91)
2007	(31,73;32,86)	(31,29;34,71)	2016	(32,33;33,46)	(34,29;37,71)
2008	(30,83;31,96)	(29,89;33,31)	2017	(33,13;34,19)	(34,89;37,92)



(a)



(b)

Hình 3. Số liệu thực tế và mờ hóa đỉnh mặn tại Gành Hào (a), Cửa Lớn (b) của mô hình đề nghị

So sánh mô hình đề nghị với khoảng cách chồng lấp, khoảng cách Hausdoff, khoảng cách City-block và phương pháp ARIMA (cho 2 chuỗi đầu mút của các khoảng), ta có Bảng 4.

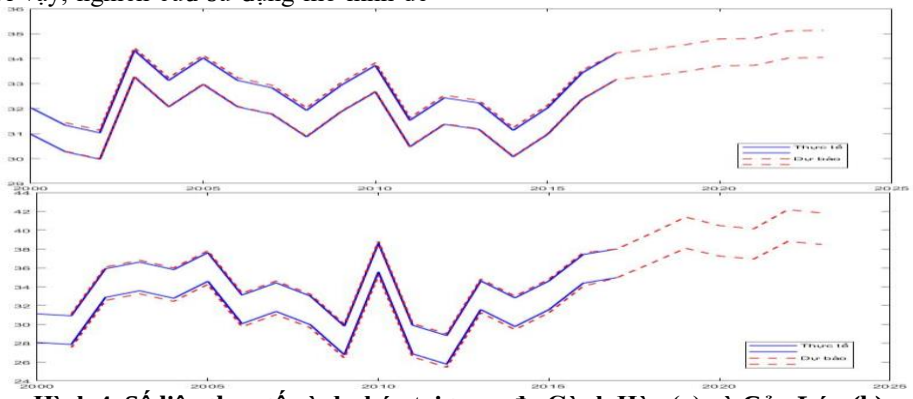
Bảng 4. Tham số MAPE của các mô hình

Trạm đo	Đề nghị	Hausdoff	City-block	ARIMA
Gành Hào	0,20	1,20	2,07	5,71
Cửa Lớn	1,15	2,35	4,95	14,13

Bảng 4 cho thấy mô hình đề nghị có các giá trị MAPE nhỏ nhất nên có kết quả thực hiện tốt nhất và kết quả này tốt hơn nhiều so với các mô hình còn lại. Chính vì vậy, nghiên cứu sử dụng mô hình đề nghị để dự báo khoảng đỉnh mặn đến năm 2023 cho hai trạm. Kết quả thực hiện được trình bày ở Bảng 5 và được minh họa ở Hình 4.

Bảng 5. Dự báo đỉnh mặn tại 2 trạm đo giai đoạn 2018-2023

Năm	Gành Hào	Cửa Lớn
2018	(33,31; 34,37)	(36,50; 39,67)
2019	(33,49; 34,56)	(38,09; 41,40)
2020	(33,71; 34,79)	(37,26; 40,49)
2021	(33,73; 34,81)	(36,94; 40,15)
2022	(34,03; 35,12)	(38,82; 42,19)
2023	(34,05; 35,14)	(38,49; 41,83)



Hình 4. Số liệu thực tế và dự báo tại trạm đo Gành Hào (a) và Cửa Lớn (b)

Bảng 5 và Hình 4 cho thấy kết quả dự báo tốt vì khoảng dự báo bao phủ và rất sát với thực tế. Trong tương lai đến năm 2023, đỉnh mặn của hai trạm có khuynh hướng tiếp tục tăng.

5. KẾT LUẬN

Nghiên cứu đã đề xuất một mô hình dự báo cho chuỗi thời gian dạng khoảng. Mô hình này có nhiều cải tiến từ chuỗi rời rạc sang chuỗi khoảng. Đầu tiên là việc sử dụng khoảng cách chồng lấp làm độ đo đánh giá sự tương tự của các khoảng mà nó có ưu điểm hơn các khoảng cách phổ biến như khoảng cách Hausdoff, khoảng cách City-block và khoảng cách Euclide. Thứ hai là việc sử dụng dữ liệu biến đổi của chuỗi làm tập nền thay vì dữ liệu gốc. Cuối cùng là sự cải tiến bài toán phân tích chùm mờ cho phần tử rời rạc áp dụng cho dữ liệu khoảng để từ đó đề xuất nguyên tắc dự báo mới. Mô hình đề nghị có thể thực hiện hiệu quả trên số liệu thực qua chương trình được thiết lập trên phần mềm Matlab. Ví dụ minh hoạ và áp dụng cho thấy sự hợp lý của mô hình đề nghị.

TÀI LIỆU THAM KHẢO

Abbasov, A. & 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., & 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., & 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. & 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. & 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. & Hung, W. (2015). An automatic clustering algorithm for probability density functions. *Journal of Statistical Computation and Simulation, 85*(1), 3047–3063.

Eren, B., Vedide, R., Uslu, U., & Erol, E. (2014). A modified 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.

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

Song, Q. & Chissom, B. S. (1993). Fuzzy time series and its models. *Fuzzy sets and systems, 54*(3), 269–277.

Tai, V. V. (2019). An improved fuzzy time series forecasting model using variations of data. *Fuzzy Optimization and Decision Making, 18*(2), 151-173.

Tai, V. V., Dinh, P. T., Thao, N. T. & Tuan, L. H (2020). An automatic clustering for interval data using the genetic algorithm. *Annals of Operations Research*. Doi.org/10.1007/s10479-020-03606-8.