Tạp chí Khoa học Trường Đại học Cần Thơ
Số chuyên đề: Khoa học tự nhiên
website: ctujsvn.ctu.edu.vn

DOI:10.22144/ctu.jvn.2022.100

XÂY DỰNG MÔ HÌNH DỰ BÁO KHOẢNG CHO CHUỖI THỜI GIAN DỰA TRÊN SỰ CẢI TIẾN TRONG THIẾT LẬP MỐI QUAN HỆ MỜ

Võ Văn Tài*, Võ Thị Huệ Chi và Huỳnh Thị Yến Nhi
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: 24/02/2022
Ngày nhận bài sửa: 21/03/2022
Ngày duyệt đăng: 06/04/2022

Title:
Building the interval forecasting model for time series based on the improvement in establishing the fuzzy relationship

Từ khóa:
Dữ liệu ảnh, dữ liệu khoảng, khoảng cách chồng lấp, phân tích chùm

Keywords:
Clustering analysis, image data, interval data, overlap distance

ABSTRACT

Time series is a type data stored normally, and has huge demand in forecasting for many real problems. This study proposes the predictive model for interval time series based on the improvement in establishing the fuzzy relationship. In this model, the universal set is the consecutive changes of two time intervals, and the divided number for intervals is determined by the fuzzy cluster analysis algorithm for interval data. Based on the fuzzy relationship between the elements of the universal set and the divided intervals, the principle for interpolating the historical data and forecasting for the future are established. The proposed model is detailed step-by-step, and illustrated by numerical examples. It is also applied in forecasting the temperature in Ha Noi to illustrate the practical application. The illustrative example and practical application show the suitability of the proposed model as well as its advantages in comparison with popular models.

TÓM TẮT

Chuỗi thời gian là một kiểu dữ liệu được lưu trữ phổ biến và có nhu cầu dự báo rất lớn cho nhiều vấn đề thực tế. Nghiên cứu này đề nghị mô hình dự báo cho chuỗi thời gian khoảng dựa trên sự cải tiến trong thiết lập mối quan hệ mờ. Trong mô hình này, tập nền là sự biến đổi liên tiếp của hai khoảng thời gian và số lượng khoảng chia cho nó được xác định từ thuật toán phân tích chùm mờ dành cho dữ liệu khoảng. Dựa trên mối quan hệ mờ giữa những phần tử của tập nền và các khoảng được chia, một nguyên tắc mờ hoá dữ liệu quá khứ và dự báo cho tương lai được thiết lập. Mô hình đề nghị được trình bày chi tiết các bước và được minh hoạ bởi ví dụ số. Nó cũng được áp dụng trong dự báo nhiệt độ ở Hà Nội để minh hoạ cho áp dụng thực tế. Ví dụ minh hoạ và áp dụng thực tế cho thấy sự phù hợp của mô hình đề nghị cũng như thuận lợi của nó trong so sánh với các mô hình phổ biến.

1. GIỚI THIỆU

Chuỗi thời gian là kiểu dữ liệu được lưu trữ theo thời gian như giờ, ngày, tháng, quý hay năm. Đây là kiểu dữ liệu phổ biến được ghi nhận ở tất các lĩnh vực và có nhu cầu dự báo rất lớn trong thực tế, trong sự phát triển kinh tế xã hội của mỗi quốc gia. Trong thống kê, dự báo cho chuỗi thời gian có thể thực hiện bằng mô hình hồi quy, hoặc chuỗi thời gian và gần đây là chuỗi thời gian mờ (Tai, 2019; Tai and Thuy, 2020). Mô hình hồi quy đã được nghiên cứu từ lâu và được sử dụng phổ biến ngày nay. Mặc dù có

nhiều cải tiến, nhiều mô hình cụ thể trong hình thức tham số và phi tham số, nhưng hồi quy vẫn có rất nhiều hạn chế trong ứng dụng. Khi xây dựng mô hình hồi quy, chúng ta phải giả sử nhiều điều kiện về số liệu mà thực tế rất khó thoả mãn nên nó chỉ thích hợp cho rất ít trường hợp cụ thể (Abreu and Ambe, 2013; Yanpeng, 2020). Khắc phục điều này, mô hình chuỗi thời gian, mô hình dành riêng cho dữ liệu chuỗi đã được đề nghị. Mô hình này được đánh giá có nhiều ưu điểm hơn hồi quy qua nhiều tập dữ liệu (Yanpeng et al., 2020). Tuy nhiên với điều kiện dừng và sai số là một ồn trắng, mô hình chuỗi thời gian cũng hạn chế cho nhiều tập dữ liệu (Aladag et al., 2012). Khắc phục yếu điểm của chuỗi thời gian, chuỗi thời gian mờ đã được đề xuất (Chen and Hsu, 2012; Ghosh et al., 2015).

Mô hình chuỗi thời gian mờ (fuzzy time series, (FTS)) được phát triển theo hai hướng: (a) Mờ hoá dữ liệu gốc để tạo ra sự liên kết của nó, sau đó áp dụng một mô hình đã biết để dự báo cho tương lai và (b) xây dựng mô hình dự báo trực tiếp cho tương lai. So với (b), hướng nghiên cứu (a) được quan tâm nhiều hơn bởi các nhà thống kê. Có rất nhiều công trình nghiên cứu được công bố theo hướng này trong các năm qua (Huarng, 2001; Ming, 2002; Own and Yu, 2005; Tai et al., 2021). Theo hướng này, một mô hình đề nghị gồm 3 bước chính (i) Xác định tập nền và khoảng chia cho tập nền, (ii) Tìm mối quan hệ giữa các phần tử trong tập nền và (iii) Xây dựng luật mờ hoá dữ liệu. Một yếu điểm chung của các mô hình trên là chúng chỉ mờ hoá được dữ liệu quá khứ mà không dự báo được cho tương lai. Bởi vì mục đích cuối cùng là dự báo, do đó sau khi thực hiện các mô hình này, chúng phải nhờ đến một mô hình khác để thực hiện. Khi chúng ta xây dựng các mô hình mờ hoá dữ liệu, chắc chắn sẽ có những sai số nhất định, do đó nghiên cứu (a) còn rất nhiều thách thức khi áp dụng vào thực tế, có thể mất thông tin trong giai đoạn mờ hoá, nên khi dự báo cho tương lai, hướng nghiên cứu thứ nhất thường chỉ cho kết quả tốt về một khuynh hướng phát triển của dữ liệu (tăng hoặc giảm). Khi tương lai có khuynh hướng biến đổi vừa có tăng, vừa có giảm phức tạp thì kết quả dự báo rất hạn chế (Tai, 2019).

Trong quá trình nghiên cứu, kết quả cho thấy so với (a), FTS phát triển theo hướng (b) chưa được quan tâm nhiều. Theo hướng này, các mô hình tiêu biểu được nhắc đến là Abbasov-Mamedova (2003), Tai (2019) và Tinh (2020). Mô hình FTS theo hướng (b) gồm 4 bước chính, trong đó 3 bước đầu tương tự như hướng (a). Bước thứ tư là xây dựng một luật dự báo cho tương lai từ các mối quan hệ chuỗi các phần tử quá khứ được thiết lập bởi ba bước trên. Đây là bước khó nhất, cũng là bước quan trọng nhất cho việc thiết lập một mô hình dự báo tốt. Mô hình của Abbasov và Mamedova (2003) và Tai (2019) thiết lập luật dự báo dựa trên mức độ quan hệ của các phần tử theo cấp độ ngôn ngữ (thường là thang đo Likert 5 cấp độ, 7 cấp độ hoặc 9 cấp độ). Trong mô hình của Tinh (2020) luật dự báo được thiết lập dựa vào số mờ tam giác, hình thang hoặc các số mờ cải tiến từ các số mờ này. Các mô hình này có hạn chế là việc chọn các tham số theo kinh nghiệm mà không có một nguyên tắc chung. Tham số được chọn cho chuỗi này có thể không tốt cho chuỗi kia. Do đó, trong thực tế, chúng chỉ hiệu quả cho những chuỗi cụ thể mà không phải là tất cả. Tai et al. (2021) đã chia chuỗi thành các nhóm theo mức độ biến đổi của các phần tử chuỗi và xây dựng nguyên tắc dự báo theo mối quan hệ mờ của bài toán phân tích chùm. Hạn chế của mô hình này là chỉ dự báo cho chuỗi điểm mà không thực hiện dự báo chuỗi khoảng.

Một vấn đề khác được đặt ra trong thực tế là nhu cầu dự báo khoảng cho chuỗi thời gian như giá trị cao nhất và thấp nhất. Trong trường hợp này, chúng ta có thể xem mỗi chuỗi khoảng gồm 2 chuỗi độc lập: chuỗi cận trên và chuỗi cận dưới để thực hiện. Hai chuỗi này được sử dụng trong thực tế khi dự báo vì giữa hai chuỗi không độc lập và độ đo đánh giá sự tương tự của các khoảng hoàn toàn khác độ đo đánh giá sự tương tự của điểm nên cách làm trên là rất hạn chế (Sato-Ilic, 2011). Một số nghiên cứu quan trọng gần đây về vấn đề dự báo này là Andre et al. (2008), Tao (2015), Tai and Dinh (2021). Tuy nhiên, nghiên cứu của Andre et al. (2008) và Tao (2015) chỉ là sự mờ hoá chuỗi khoảng. Để dự báo khoảng cho tương lai, chúng cần sử dụng những mô hình dự báo điểm khác. Tai and Dinh (2021) đã phát triển mô hình dự báo khoảng từ mô hình dự báo điểm của Tai et al. (2021). Mặc dù mô hình này đã có sử dụng khoảng cách chồng lấp nhưng nguyên tắc dự báo của nó dựa vào bài toán phân tích chùm mờ mà không dựa vào cấp độ ngôn ngữ nên nó thường không tốt khi chuỗi khoảng có sự biến đổi phức tạp, không theo quy luật của quá khứ.

Trong nghiên cứu này, việc xây dựng mô hình chuỗi thời gian khoảng được xem xét dựa trên sự cải tiến từ mô hình dành cho dữ liệu điểm. Trong mô hình này, sự biến đổi của hai khoảng liên tiếp được tìm hiểu đầu tiên để xây dựng tập nền và chia tập nền thành các khoảng với số lượng thích hợp qua thuật toán phân tích chùm. Thuật toán phân tích chùm được xây dựng dựa vào khoảng cách chồng lấp. Dựa trên mô hình mờ hoá dữ liệu điểm của

Singh (2007), luật mờ hoá dữ liệu quá khứ và dự báo tương lai cho khoảng được thiết lập. Mô hình đề nghị được minh hoạ chi tiết qua ví dụ số và có thể áp dụng nhanh chóng cho chuỗi thực tế. Các chuỗi được áp dụng cho thấy sự hợp lý của mô hình đề nghị cũng như ưu điểm của nó trong so sánh với các mô hình phổ biến.

2. CÁC VẤN ĐỀ LIÊN QUAN

2.1. Khoảng cách

Cho 2 khoảng có p - chiều *A* và *B:*

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

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

Khi đó ta có các khoảng cách phổ biến giữa *A* và *B* như sau:

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 Hausdorff:

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

Khoảng cách City-block:

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

Khoảng cách chồng lấp:

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

với

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

$O\left(A,B\right)$ là vùng chồng lấp giữa hai khoảng *A* và *B,*

$d_H\left(A,B\right)$ là khoảng cách Hausdorff.

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

$\left\|c_a-c_b\right\|\le r_b-r_a$ thì $O(A,B)=2r_a+1.$

(ii) Nếu *B* nằm hoàn toàn trong *A*:

$\left\|c_a-c_b\right\|\le r_a-r_b$ 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*:

$\left\|r_a-r_b\right\|<\left\|c_a-c_b\right\|<r_a+r_b$ thì $O(A,B)=r_a+r_b-\left|c_a-c_b\right|.$

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

Do đó công thức (1) được cụ thể bởi (2):

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

2.2. Các tham số đánh giá mô hình xây dựng

Chúng ta kí hiệu $\{X_{il}\}$ và $\{X_{ir}\}$, i = 1, 2, …, *n* lần lượt là cận dưới và cận trên của dữ liệu gốc $\{X_t\}$. Đặt $\left\{X_{il}\right\}$ và $\left\{X_{ir}\right\}$ lần lượt là giá trị dự báo của $\{X_{il}\}$ và $\{X_{ir}\}$. Khi đó, những tham số sau để đánh giá mô hình dự báo cho chuỗi thời gian:

* Sai số bình phương trung bình:

$$MSE=\frac{1}{2}\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_{il}-X_{il}\right)^2+\frac{1}{n}\sum_{i=1}^{n}\left(X_{ir}-X_{ir}\right)^2\right].$$

* *Sai số tuyệt đối trung bình*:

$$MAE=\frac{1}{2}\left[\frac{1}{n}\sum_{i=1}^{n}\left|\left(X_{il}-X_{il}\right)\right|+\frac{1}{n}\sum_{i=1}^{n}\left|\left(X_{ir}-X_{ir}\right)\right|\right].$$

* Sai số phần trăm tuyệt đối trung bình:

$$MAPE=\frac{1}{2}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left|X_{il}\right|-\left|X_{il}\right|}{X_{il}}.100\right)+\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left|X_{ir}\right|-\left|X_{ir}\right|}{X_{ir}}.100\right)\right].$$

Khi xây dựng mô hình dự báo, các tham số trên càng nhỏ thì mô hình xây dựng sẽ càng tốt và ngược lại.

3. MÔ HÌNH DỰ BÁO ĐỀ NGHỊ

Cho một chuỗi $\{X\} = \{x_t = (a_t, b_t), t = 1, 2, \ldots, n\}$. Thuật toán đề xuất cho chuỗi thời gian này gồm các bước như sau:

Bước 1: Tính toán sự biến đổi của hai khoảng thời gian liên tiếp để có chuỗi Y:

$$Y = \{Y_t = x_t - x_{t-1} = (c_t; d_t), t = 2,3,\ldots,n\}.$$

Bước 2: Chia chuỗi $\{Y\}$ thành c chùm phù hợp $w_1, w_2, \ldots, w_c$, bởi thuật toán phân tích chùm mờ cho khoảng (FCAI) như sau:

Bước 2.1. Gọi trọng tâm ban đầu của các chùm là

$$V^{(0)} = \{v_1^{(0)}, v_2^{(0)}, \ldots, v_N^{(0)}\} = \{Y_1, Y_2, \ldots, Y_N\} \text{ và cho } \varepsilon = 0,0001.$$

Bước 2.2. Cập nhật trọng tâm mới của chùm bởi công thức (3).

$$v_i^{(t+1)} = \sum_{j=1}^{N} \frac{f\left(v_i^{(t)}, v_j^{(t)}\right)}{\sum_{k=1}^{N} f\left(v_i^{(t)}, v_k^{(t)}\right)} v_j^{(t)},\ i = \overline{1,N}, \qquad (3)$$

trong đó

$$f\left(v_i^{(t)}, v_j^{(t)}\right) = \begin{cases} e^{-\left(\frac{d_0\left(v_i^{(t)}, v_j^{(t)}\right)}{\lambda}\right)} & \text{khi } d_0\left(v_i^{(t)}, v_j^{(t)}\right) \le \mu.\alpha_{ij}(t), \\ 0 & \text{khi } d_0\left(v_i^{(t)}, v_j^{(t)}\right) > \mu.\alpha_{ij}(t), \end{cases}$$

với

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

$$\mu = \frac{1}{C_N^2} \sum_{i<j} d\left(v_i^{(0)}, v_j^{(0)}\right),$$

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

Bước 2.3. Tính $v = \max_i \left\{ \left| v^{(t+1)} - v^{(t)} \right| \right\}$ và lặp lại Bước 2.2 cho đến khi $v < \varepsilon$.

Tìm phần tử của $v^{(t)}$. Nếu có m phần tử trong $v^{(t)}$ thì chuỗi $\{Y\}$ sẽ được chia thành m chùm.

Bước 3. Gọi Y_L và Y_R lần lượt là chuỗi cận dưới và cận trên của Y, đặt

$$U_L = \left[\min\{Y_L\}; \max\{Y_L\}\right],$$
$$U_R = \left[\min\{Y_R\}; \max\{Y_R\}\right],$$

lần lượt là tập nền của Y_L và Y_R. Chia U_L thành m khoảng bằng nhau $A_1, A_2, \ldots, A_m$ và U_R thành m khoảng bằng nhau $B_1, B_2, \ldots, B_m$.

Vào thời gian t và $t+1$, nếu $y_t \in Y_L \in A_i$ và $y_{t+1} \in Y_L \in A_j$, khi đó quan hệ mờ được chỉ ra là $A_i \to A_j$. Khi đó A_i được gọi là trạng thái hiện tại, và A_j được gọi là trạng thái kế tiếp. Sau đó, chúng ta đếm số quan hệ mờ cho mỗi trường hợp trong chuỗi Y_L. Thực hiện tương tự cho chuỗi cận trên Y_R.

Bước 4. Xây dựng luật dự báo cho chuỗi cận dưới

Quy tắc dự báo cho quan hệ mờ $A_i \to A_j$ với 3 cấp độ thời gian $t-2, t-1, t, t = 4,5,\ldots,n$ như sau:

Bước 4.1. Tính toán các giá trị:

$$c = \left\| y_t - y_{t-1} \right| - \left| y_{t-1} - y_{t-2} \right\|,$$
$$Z_{1l} = y_t + \frac{c}{2}, Z_{1r} = y_t - \frac{c}{2},$$
$$Z_{2l} = y_t + c, Z_{2r} = y_t - c,$$
$$Z_{3l} = y_t + \frac{c}{4}, Z_{3r} = y_t - \frac{c}{4},$$
$$Z_{4l} = y_t + 2c, Z_{4r} = y_t - 2c,$$
$$Z_{5l} = y_t + \frac{c}{6}, Z_{5r} = y_t - \frac{c}{6},$$
$$Z_{6l} = y_t + 3c, Z_{6r} = y_t - 3c,$$

Bước 4.2. Gọi $L[A_j]$ và $U[A_j]$ lần lượt là cận dưới và cận trên của khoảng A_j. Chúng ta có quy tắc tính R và S cho 12 trường hợp mà phụ thuộc vào cận dưới và cận trên của khoảng A_j như sau:

– Nếu $L[A_i] \le Z_{1l} \le U[A_i]$ thì $R_1 = Z_{1l}, S_1 = 1$, ngược lại $R_1 = 0, S_1 = 0$.

– Nếu $L[A_i] \le Z_{1r} \le U[A_i]$ thì $R_2 = Z_{1r}, S_2 = 1$, ngược lại $R_2 = 0, S_2 = 0$.

– Nếu $L[A_i] \le Z_{2l} \le U[A_i]$ thì $R_3 = Z_{2l}, S_3 = 1$, ngược lại $R_3 = 0, S_3 = 0$.

– Nếu $L[A_i] \le Z_{2r} \le U[A_i]$ thì $R_4 = Z_{2r}, S_4 = 1$, ngược lại $R_4 = 0, S_4 = 0$.

– Nếu $L[A_i] \le Z_{3l} \le U[A_i]$ thì $R_5 = Z_{3l}, S_5 = 1$, ngược lại $R_5 = 0, S_5 = 0$.

– Nếu $L[A_i] \le Z_{3r} \le U[A_i]$ thì $R_6 = Z_{3r}, S_6 = 1$, ngược lại $R_6 = 0, S_6 = 0$.

– Nếu $L[A_i] \le Z_{4l} \le U[A_i]$ thì $R_7 = Z_{4l}, S_7 = 1$, ngược lại $R_7 = 0, S_7 = 0$.

– Nếu $L[A_i] \le Z_{4r} \le U[A_i]$ thì $R_8 = Z_{4r}, S_8 = 1$, ngược lại $R_8 = 0, S_8 = 0$.

– Nếu $L[A_i] \le Z_{5l} \le U[A_i]$ thì $R_9 = Z_{5r}, S_9 = 1$, ngược lại $R_9 = 0, S_9 = 0$.

– Nếu $L[A_i] \le Z_{5r} \le U[A_i]$ thì $R_{10} = Z_{5r}, S_{10} = 1$, ngược lại $R_{10} = 0, S_{10} = 0$.

– Nếu $L[A_i] \le Z_{6l} \le U[A_i]$ thì $R_{11} = Z_{6l}, S_{11} = 1$, ngược lại $R_{11} = 0, S_{11} = 0$.

– Nếu $L[A_i] \le Z_{6r} \le U[A_i]$ thì $R_{12} = Z_{6r}, S_{12} = 1$, ngược lại $R_{12} = 0, S_{12} = 0$.

Tính toán hai giá trị *R* và *S* như sau:

$$R = R_1 + R_2 + \cdots + R_{12}, S = S_1 + S_2 + \cdots + S_{12}.$$

R_i và S_i, $i = 1, 2, ..., 12$ được coi như các thành phần để xác định trọng lượng ảnh hưởng đến mức độ của $A_i \to A_j$. Nó phụ thuộc vào 12 giá trị của Z_{jl} và $Z_{jr}, j = 1, 2, ..., 6,$ và giúp mô hình đề xuất xác định giá trị cho sự biến đổi từ quan hệ mờ $A_i \to A_j$ của thời gian *t* tới *t* + 1.

Bước 4.3. Xác định giá trị của sự biến đổi từ $A_i \to A_j$ của thời gian t tới $t+1$ bởi (4).

$$V_1 = \frac{R + M\left[A_j\right]}{S + 1}, \quad (4)$$

trong đó $M[A_j]$ là điểm giữa của khoảng A_j.

Bước 4.4. Dự đoán giá trị biến đổi từ *t* tới $t+1$.

Nếu quan hệ mờ $A_i \to A_j$ xuất hiện k lần $(k \ge 1)$ cho hai thời gian liên tiếp khác nhau, khi đó chúng ta thực hiện Bước 5.1, Bước 5.2, Bước 5.3 và Bước 5.4 cho mối quan hệ mờ để thu được $V_1, V_2, \ldots, V_k$. Khi đó, giá trị cho biến đổi từ $A_i \to A_j$ được xác định bởi (5).

$$V = \frac{V_1 + V_2 + \cdots + V_k}{k}. \quad (5)$$

Bước 4.5. Tính toán giá trị dự báo cho V_t:

$$x_{lt+1} = X_t + V_t, \quad (6)$$

ở đây X_t là giá trị của chuỗi $X(t)$ ở thời gian *t*.

Bước 5. Thực hiện hoàn toàn tương tự cho chuỗi cận trên để có x_{rt+1}. Khi đó, giá trị dự báo tại thời điểm *t* + 1 là $x_{t+1} = [x_{lt+1}; x_{rt+1}]$.

Để dự báo cho khoảng tiếp theo, ta nhập giá trị vừa dự báo trước đó vào chuỗi gốc ban đầu và lặp lại các bước phía trên để thực hiện.

4. VÍ DỤ MINH HỌA VÀ ÁP DỤNG

4.1. 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 được sử dụng trong nhiều nghiên cứu (Huarng, 2001; Singh, 20007; Ghosh, 2013). Số liệu được cho bởi Bảng 1:

Bảng 1. Số liệu tuyển sinh của trường Đại học Alabama giai đoạn 1971-1992

Năm	X_t	Năm	X_t
1971	13055	1982	15433
1972	13563	1983	15497
1973	13867	1984	15145
1974	14696	1985	15163
1975	15460	1986	15984
1976	15311	1987	16859
1977	15603	1988	18150
1978	15861	1989	18970
1979	16807	1990	19328
1980	16919	1991	19337
1981	16388	1992	18876

Ước lượng khoảng tin cậy 95% cho chuỗi dữ liệu Bảng 1, ta có Cột "Khoảng *X*" của Bảng 2.

Bảng 2. Dữ liệu khoảng và các giá trị cho một số bước của mô hình đề nghị

Năm	Khoảng X	Y_l	Y_r	A_i
1971	(12249,61; 13860,39)	-	-	-
1972	(12757,61; 14368,39)	508	508	A_3
1973	(13061,61; 14672,39)	304	304	A_3
1974	(13890,61; 15501,39)	829	829	A_4
1975	(14654,61; 16265,39)	764	764	A_4
1976	(14505,61; 16116,39)	-149	-149	A_2
1977	(14797,61;16408,39)	292	292	A_3
1978	(15055,61; 16666,39)	258	258	A_3
1979	(16001,61;17612,39)	946	946	A_4
1980	(16113,61;17724,39)	112	112	A_2
1981	(15582,61;17193,39)	-531	-531	A_1
1982	(14627,61;16238,39)	-955	-955	A_1
1983	(14691,61;16302,39)	64	64	A_2
1984	(14339,61;15950,39)	-352	-352	A_2
1985	(14357,61;15968,39)	18	18	A_2
1986	(15178,61;16789,39)	821	821	A_4
1987	(16053,61;17664,39)	875	875	A_4
1988	(17344,61;18955,39)	1291	1291	A_4
1989	(18164,61;19775,39)	820	820	A_4
1990	(18522,61;20133,39)	358	358	A_3
1991	(18531,61;20142,39)	9	9	A_2
1992	(18070,61;19681,39)	-461	-461	A_1

Bước 1. Tính sự biến đổi của hai khoảng thời gian liên tiếp, ta có Cột Y_l và Y_r của Bảng 2.

Bước 2. Sau 5 vòng lặp, 22 khoảng ban đầu hội tụ về 4 khoảng

v_1 = (-1435,1; 1786,5), v_2 = (-754,7; 2466,9), v_3 = (-2011,7; 1209,8),v_4 =(-2539,6; 1681,9), nên ta chia chuỗi Y thành 4 khoảng.

Bước 3. Ta có tập nền là U = [-955; 1291]. Ta chia tập nền thành 4 khoảng bằng nhau và tính điểm giữa mỗi khoảng, ta được Bảng 3.

Bảng 3. Các khoảng chia và điểm giữa của mỗi khoảng

Các khoảng chia	Giá trị	Điểm giữa	Giá trị
U_1	$[-955,0;-393,5]$	u_1^0	-674,25
U_2	$[-393,5;168,0]$	u_2^0	-112,75
U_3	$[168,0;729,5]$	u_3^0	448,75
U_4	$[729,5;1291,0]$	u_4^0	1010,25

Từ Bảng 2, ta xác định được mối quan hệ giữa mỗi điểm thời gian và các khoảng chia, với kết quả được tổng kết bởi Bảng 4

Bảng 4. Mối quan hệ mờ của các nhóm

Mối quan hệ	Số lượng	Mối quan hệ	Số lượng
$A_1 \to A_1$	`1	$A_2 \to A_4$	1
$A_1 \to A_2$	1	$A_3 \to A_2$	1
$A_2 \to A_1$	2	$A_3 \to A_3$	2
$A_2 \to A_2$	2	$A_3 \to A_4$	2
$A_2 \to A_3$	1	$A_4 \to A_2$	2
$A_4 \to A_3$	1	$A_4 \to A_4$	4

Bước 4. Xây dựng luật dự báo cho chuỗi cận dưới:

Bước 4.1. Giả sử chúng ta cần dự báo giá trị của chuỗi cận dưới vào năm 1976. Dựa vào Bảng 1 và Bảng 3, ta thấy có sự khác biệt của năm 1975 rơi vào tập mờ A_4 có ba quan hệ mờ logic được thiết lập bao gồm: $A_4 \to A_2$ với tần suất là 2; $A_4 \to A_3$ có tần suất là 1 và $A_4 \to A_4$ có tần suất là 4. Tính toán các giá trị sau:

$$c = \big\|y_t - y_{t-1}\big| - \big|y_{t-1} - y_{t-2}\big\| = \big\|764 - 829\big| - \big|829 - 304\big\| = 460,$$

$$Z_{1l} = y_t + \frac{c}{2} = 764 + \frac{460}{2} = 994,$$

$$Z_{1r} = y_t - \frac{c}{2} = 764 - \frac{460}{2} = 534,$$

$$Z_{2l} = y_t + c = 764 + 460 = 1224,$$

$$Z_{2r} = y_t - c = 764 - 460 = 304,$$

$$Z_{3l} = y_t + \frac{c}{4} = 764 + \frac{460}{4} = 879,$$

$$Z_{3r} = y_t - \frac{c}{4} = 764 - \frac{460}{4} = 649,$$

$$Z_{4l} = y_t + 2c = 764 + 2.460 = 1684,$$

$$Z_{4r} = y_t - 2c = 764 - 2.460 = -156,$$

$$Z_{5l} = y_t + \frac{c}{6} = 764 + \frac{460}{6} = 840,67,$$

$$Z_{5r} = y_t - \frac{c}{6} = 764 - \frac{460}{6} = 687,33,$$

$$Z_{6l} = y_t + 3c = 764 + 3,460 = 2144,$$

$$Z_{6r} = y_t - 3c = 764 - 3,460 = -616.$$

Bước 4.2.

Xét quan hệ mờ $A_4 \to A_3$, ta có

$$L[A_3] = 168,0; U[A_3] = 729,5; M[A_3] = 448,75.$$

Áp dụng nguyên tắc trong Bước 4.3, ta tính được:

$$R_1 = 0, S_1 = 0,\ R_2 = 534,\ S_2 = 1,\ R_3 = 0, S_3 = 0,$$
$$R_4 = 304,\ S_4 = 1,\ .\ R_5 = 0, S_5 = 0.\ R_6 = 649,\ S_6 = 1,$$
$$R_7 = 0, S_7 = 0.\ R_8 = 0, S_8 = 0,\ R_9 = 0, S_9 = 0,$$
$$R_{10} = 687.33, S_{10} = 1\ R_{11} = S_{11} = 0\ R_{12} = S_{12} = 0.$$

Khi đó

$$R = R_1 + R_2 + \cdots + R_{12} = 534 + 304 + 649 + 687,33 = 2174,33.$$
$$S = S_1 + S_2 + \cdots + S_{12} = 1+1+1+1 = 4.$$
$$V_1 = \frac{R + M[A_3]}{S+1} = \frac{2174,33 + 448,75}{5} = 524,616.$$

Tương tự, xét quan hệ mờ $A_4 \to A_2$, ta tính được R = -156 và S = 1, do đó giá trị cho sự biến đổi của quan hệ mờ $A_4 \to A_2$ là

$$V_2 = \frac{R + M[A_2]}{S+1} = \frac{-156 - 112.75}{2} = -134,75.$$

Cũng tương tự khi xét quan hệ mờ $A_4 \to A_4$, ta nhận được:

$$R = R_1 + R_2 + \cdots + R_{12} = 994 + 1224 + 879 + 840,67 = 3937,67,$$
$$S = 4.$$

Do đó giá trị của sự biến đổi quan hệ mờ $A_4 \to A_4$ là:

$$V_3 = \frac{R + M[A_4]}{S+1} = \frac{3937,67 + 1010,25}{5} = 989,584.$$

Bước 4.3. Dự đoán giá trị biến đổi từ t tới $t+1$.

Vì $A_4 \to A_2$ với tần suất là 2, $A_4 \to A_3$ có tần suất là 1 và $A_4 \to A_4$ có tần suất là 4, nên giá trị của sự biến đổi từ 1975 đến 1976 là:

$$V = \frac{V_1 + 2V_2 + 4V_3}{7}$$
$$= \frac{524,616 + 2.(-134,375) + 4.989,54}{7} = 602,004.$$

Bước 4.4. Tính toán giá trị dự báo cho V_t bằng công thức (7), ta có

$$x_{l1976} = X_{l1975} + V = 14654,61 + 602,004$$
$$= 15256,614.$$

Bước 5. Tính toán tương tự Bước 5 cho chuỗi cận trên ta thu được giá trị dự báo vào năm 1976 là

$$x_{r1976} = X_{r1975} + V = 16265,39 + 602,004$$
$$= 16867,394.$$

Khi đó giá trị dự báo tại thời điểm 1976 là

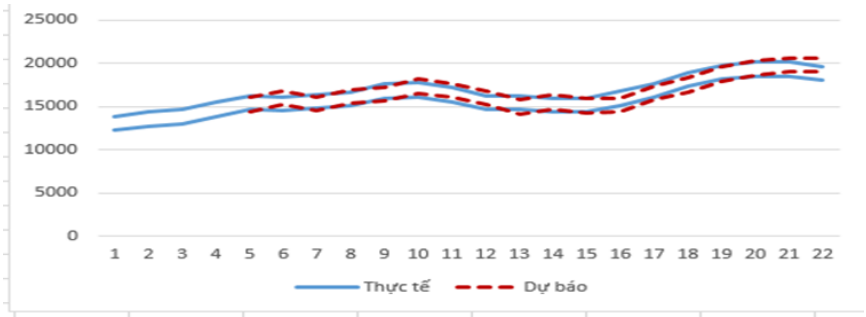
$$x_{1976} = (x_{l1976}; x_{r1976}) = (15256,64; 16867,42).$$

Để dự báo cho giá trị tiếp theo ta nhập giá trị vừa dự báo trước đó vào chuỗi gốc ban đầu và lặp lại các bước phía trên. Tính toán tương tự, ta có kết quả Bảng 5.

Bảng 5. Kết quả dự báo từ mô hình đề nghị cho dữ liệu Enrollment

Năm	Thực tế	Dự báo
1971	(12249,61;13860,39)	-
1972	(12757,61;14368,39)	-
1973	(13061,61;14672,39)	-
1974	(13890,61;15501,39)	-
1975	(14654,61;16265,39)	(14447,90;16058,68)
1976	(14505,61;16116,39)	(15256,64;16867,42)
1977	(14797,61;16408,39)	(14497,90;16108,68)
1978	(15055,61;16666,39)	(15326,18;16936,96)
1979	(16001,61;17612,39)	(15632,91;17243,69)
1980	(16113,61;17724,39)	(16576,37;18187,15)
1981	(15582,61;17193,39)	(16112,62;17723,4)
1982	(14627,61;16238,39)	(15207,35;16818,13)
1983	(14691,61;16302,39)	(14142,30;15753,08)
1984	(14339,61;15950,39)	(14729,31;16340,09)
1985	(14357,61;15968,39)	(14286,08;15896,86)
1986	(15178,61;16789,39)	(14378,67;15989,45)
1987	(16053,61;17664,39)	(15792,34;17403,12)
1988	(17344,61;18955,39)	(16713,11;18323,89)
1989	(18164,61;19775,39)	(17996,86;19607,64)
1990	(18522,61;20133,39)	(18711,51;20322,29)
1991	(18531,61;20142,39)	(19050,15;20660,93)
1992	(18070,61;19681,39)	(19033,29;20644,07)

Kết quả thực tế và dự báo được cho bởi Hình 1.



Hình 1. Đồ thị cho giá trị thực tế và dự báo của chuỗi Enrollment

So sánh mô hình đề nghị với mô hình ARIMA, Abbasov-Manedova (AM) và mô hình của Tai (2019), ta có Bảng 6.

Bảng 6. So sánh mô hình đề nghị với các mô hình phổ biến cho tập dữ liệu Enrollment

Mô hình	MAE	MSE	MAPE
Đề nghị	445,019	253446,123	2,806
AM	523,172	307230,00	3,302
Tai (2019)	459,379	301510,00	2,925
ARIMA	512,844	385067,244	3,342

Bảng 6 cho thấy rằng tất cả các tham số của mô hình đề nghị đều nhỏ hơn các mô hình phổ biến khác.

4.2. Áp dụng

Số liệu được sử dụng là nhiệt độ thấp nhất và cao nhất theo ngày của thành phố Hà Nội từ 26/4/2020 đến 22/05/2020. Số liệu cụ thể được trình bày trong Bảng 7.

Khi thực hiện tương tự ví dụ minh hoạ theo các bước của thuật toán đề nghị, ta có kết quả dự báo được cho bởi Bảng 8.

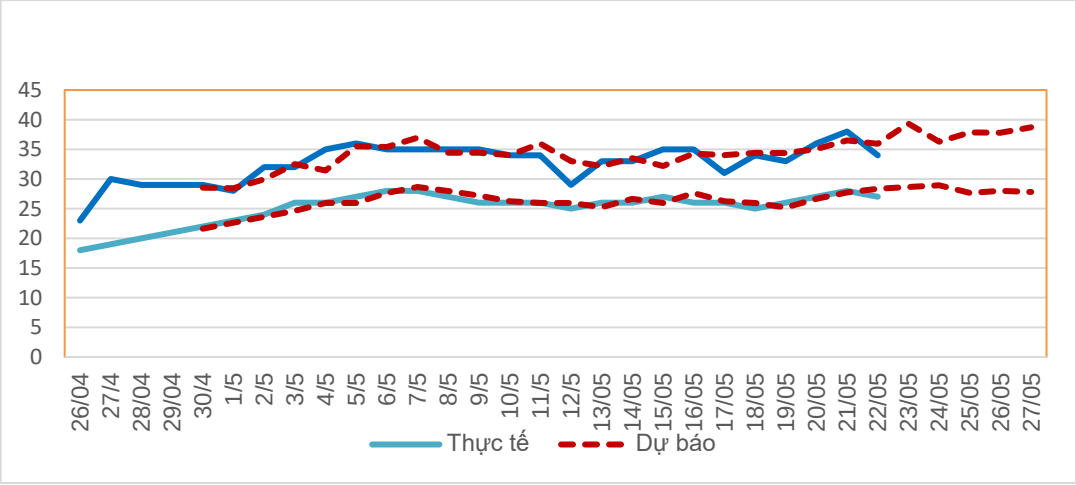
Bảng 7. Dữ liệu khoảng về nhiệt độ của Hà Nội từ 26/4/2020 đến 22/05/2020

Ngày	Nhiệt độ	Ngày	Nhiệt độ	Ngày	Nhiệt độ
26/04	(18; 23)	05/05	(27; 36)	14/05	(26; 33)
27/4	(19; 30)	06/05	(28; 35)	15/05	(27; 35)
28/04	(20;29)	07/05	(28; 35)	16/05	(26; 35)
29/04	(21; 29)	08/05	(27; 35)	17/05	26; 31)
30/4	(22; 29)	09/05	(26; 35)	18/05	(25; 34)
01/05	(23; 28)	10/05	(26; 34)	19/05	(26; 33)
02/05	(24; 32)	11/05	(26; 34)	20/05	(27; 36)
03/05	(26; 32)	12/05	(25; 29)	21/05	(28; 38)
04/05	(26; 35)	13/05	(26; 33)	22/05	(27; 34)

Bảng 8. Kết quả dự báo từ mô hình đề nghị cho dữ liệu nhiệt độ ở Hà Nội

Ngày	Thực tế	Dự báo	Ngày	Thực tế	Dự báo
26/4	(18; 23)	-	12/5	(26; 34)	(25,969; 33,015)
27/4	(19; 30)	-	13/5	(25; 29)	(25,261; 32,200)
28/4	(20;29)	-	14/5	(26; 33)	(26,664; 33,520)
29/4	(21; 29)	-	15/5	(27; 36)	(25,969; 32,200)
30/4	(22; 29)	(21,628; 28,559)	16/5	(26; 33)	(27,628; 34,284)
01/5	(23; 28)	(22,628; 28,442)	17/5	(27; 35)	(26,261; 34,015)
02/5	(24; 32)	(23,628; 29,960)	18/5	(26; 35)	(25,969; 34,400)
03/5	(26; 32)	(24,629; 32,520)	19/5	26; 31)	(25,205; 34,387)
04/5	(26; 35)	(25,950; 31,441)	20/5	(25; 34)	(26,664; 35,060)
05/5	(18; 23)	(25,969; 35,520)	21/5	(26; 33)	(27,744; 36,520)
06/5	(27; 36)	(27,664; 35,389)	22/5	(27; 36)	(28,371; 35,941)
07/5	(28; 35)	(28,664; 36,960)	23/5	-	(28,647; 39,341)
08/5	(28; 35)	(27,969; 34,442)	24/5	-	(28,944; 36,284)
09/5	(27; 35)	(27,205; 34,442)	25/5	-	(27,670; 37,866)
10/5	(26; 35)	(26,261; 34,015)	26/5	-	(27,995; 37,793)
11/5	(26; 34)	(25,969; 35,960)	27/5	-	(27,824; 38,716)

Giá trị thực tế và dự báo từ mô hình đề nghị được cho bởi Hình 2.



Hình 2. Đồ thị cho giá trị thực tế và dự báo của chuỗi nhiệt độ ở Hà Nội

So sánh mô hình đề nghị với các mô hình phổ biến khác, ta nhận được Bảng 9:

Bảng 9. So sánh sai số các mô hình dự báo nhiệt độ ở Hà Nội

Mô hình	MAE	MSE	MAPE
Đề nghị	**1,29**	**2,79**	**4,14**
AM	3,17	18,51	10,95
Tai (2019)	1,58	3,82	5,07
ARIMA	2,33	8,43	7,12

Từ Bảng 8, mô hình đề nghị cho kết quả rất nổi bật với cả 3 tham số đánh giá và có sự khác biệt có ý nghĩa so với các phương pháp khác. Kết quả từ Hình 2 cũng cho thấy giá trị dự báo tương đối gần với giá trị thực, nên chúng có thể được áp dụng trong thực tế. Trong tương lai nhiệt độ của Hà Nội có sự tăng lên, đặc biệt là cận trên.

5. KẾT LUẬN

Dự báo luôn nhận được sự quan tâm của các nhà quản lý và nhà khoa học vì nó mang lại nhiều lợi ích. Cho đến nay, nó vẫn là bài toán chưa có lời giải cuối cùng. Dựa vào khoảng cách chồng lấp để đánh giá sự tương tự của các khoảng trong chuỗi, thuật toán phân tích chùm mờ để xác định khoảng chia và và mối quan hệ mờ để đề xuất mô hình dự báo cho chuỗi thời gian khoảng. Mô hình đề nghị được kiểm chứng trên những chuỗi cụ thể và cho kết quả tốt trong so sánh với các mô hình phổ biến khác. Với chương trình được thiết lập trên phần mềm R, mô hình đề nghị có thể thực hiện nhanh chóng cho dữ liệu thực. Trong thời gian tới, mô hình đề nghị cho nhiều chuỗi thực được áp dụng để kiểm tra sự hiệu quả cũng như giải quyết được những đòi hỏi từ thực tế.

TÀI LIỆU THAM KHẢO

Abbasov, A., & Mamedova, M. (2003). Application of fuzzy time series to population forecasting. *Vienna University of Technology, 1*, 545–552. https://doi.org/10.1080/18756891.2013.808426

Abreu, P. H., & Ambe, H. M. (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), 337-345.

Andre, L. S. M., Francisco, A. T., & Teresa, B. L. (2008). Forecasting models for interval-valued time series. *Neurocomputing*, 71, 3344–3352. https://doi.org/10.1016/j.neucom.2008.02.022

Ghosh, H., Chowdhury, S., & Prajneshu, S. (2015). An improved fuzzy time series method of forecasting based on L–R fuzzy. *Journal of Applied Statistics*, *43*(6), 1128–1139. https://doi.org/10.1080/02664763.2015.1092111

Huarng, K. (2001). Heuristic models of fuzzy time series for forecasting. *Fuzzy Sets and Systems*, *123*(3), 369-386. https://doi.org/10.1016/S0165-0114(00)00093-2

Ming, C. S. (2002). Forecasting enrollments based on high-order fuzzy time series. *Fuzzy Sets and Systems*, *33*(1), 1-16. https://doi.org/10.1080/019697202753306479

Sato-Ilic, M. (2011). Symbolic clustering with interval-valued data, *Procedia Computer Science*, *6*, 358 – 363. https://doi.org/10.1016/j.procs.2011.08.066

Singh, S. R. (2007). A simple method of forecasting based on fuzzy time series. *Applied Mathematics and Computation*, *186*(1), 330–339. https://doi.org/10.1016/j.amc.2006.07.128

Tai, V. V. (2019). An improved fuzzy time series forecasting model using variations of data, *Fuzzy Optimization and Decision Making*, *18*(2), 151-173. https://doi.org/10.1007/s10700-018-9290-7

Tai, V. V., & Thuy, L. T. T. (2020). A fuzzy time series model based on improved fuzzy function and cluster analysis problem. *Communications in Mathematics and Statistics*. https://doi.org/10.1007/s40304-019-00203-5.

Tai, V. V., & Dinh P. T. (2021). Interval forecasting model for time series based on the fuzzy clustering technique. *IOP Conference Series: Materials Science and Engineering*, https://doi.org/10.1088/1757-899X/1109/1/012030

Tai, V. V., Luan, N. H., & Thuy, L. T. (2021). A forecasting model for time series based on improvements from fuzzy clustering problem. *Annals of Operations Research*. https://doi,org/10,1007/s10479-021-04041-z.

Tao, X., Yukun, B., Zhongyi, H., & Raymond, C. (2015). Forecasting interval time series using a fully complex-valued RBF neural network with DPSO and PSO algorithms. *Information* Sciences, *305*, 77-92.

https://doi.org/10.1016/j.ins.2015.01.029

Tinh, N. V. (2020). 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), 1 – 26. https://doi.org/10.1142/S1469026820500170

Yanpeng, Z., Hua, Q., Weipeng, W., & Jihong, Z. (2020). A Novel fuzzy time series forecasting model based on multiple linear regression and time series clustering. *Mathematical Problems in Engineering*, https://doi.org/10.1155/2020/9546792