TRƯỜNG ĐẠI HỌC CẦN THƠ
KHOA KHOA HỌC TỰ NHIÊN

PGS.TS. VÕ VĂN TÀI

# CHUYÊN ĐỀ THỐNG KÊ KINH TẾ
# MSHP: TN482

2024

# Chương 1

# LÀM TRƠN VÀ MỜ HOÁ CHUỖI THỜI GIAN TRONG DỰ BÁO KINH TẾ

## 1.1 Chuỗi thời gian không mờ

### 1.1.1 Giới thiệu

Chuỗi thời gian là tập hợp gồm các số liệu có cùng khái niệm và phạm vi được thu thập liên tục và thường kỳ (ngày, tuần, tháng, quý, năm) trong một thời gian dài và được sắp xếp theo một thứ tự thời gian.

Các giá trị của chuỗi tuần tự theo thời gian của đại lượng $Y$ được ký hiệu $Y_1, Y_2, \ldots, Y_t, \ldots, Y_n$, với $Y$ là giá trị được khảo sát ở thời điểm t.

Các số liệu thời gian có thể được thu thập theo một tần suất quan sát nhất định tuỳ đặc điểm từng đối tượng nghiên cứu. Ví dụ như theo ngày (chứng khoáng, lãi suất, tỉ giá ngoại tệ,...), theo tuần (lương tuần, lượng cung tiền,...), theo tháng (tỉ lệ thất nghiệp, doanh số, lượng mưa, sản lượng công nghiệp,...), theo quý (doanh thu, bình quân thu nhập,...) theo năm (Ngân sách chính phủ, sản lượng nông sản, thuỷ hải sản, giá trị nhập xuất khẩu, tốc độ tăng trưởng kinh tế,...).

### 1.1.2 Một số định nghĩa

**a) Hàm tự tương quan**

Trong chuỗi thời gian thường các giá trị ở những thời điểm khác nhau có mối tương quan với nhau. Sự tương quan này được đánh giá bằng hệ số tự tương quan.

**Định nghĩa 1.1** *Tự tương quan là sự tương quan giữa một biến với chính nó theo những độ trễ thời gian khác nhau.*

Ta tính hệ số tự tương quan của biến $Y_t$ với độ trễ $k$ theo công thức

$$\rho_k = \frac{\sum_{t=k+1}^{n}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2} \tag{1.1}$$

trong đó

$Y_t$ là biến ngẫu nhiên có luật phân phối xác suất nào đó,

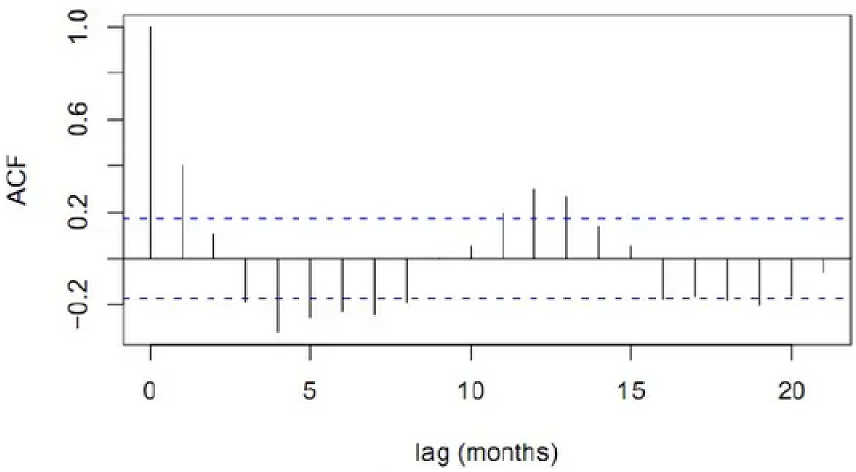
$\rho_k$ là hệ số tương quan của $Y$ ở độ trễ $k$,

$\bar{Y}$ là trung bình của $Y_t$.

Nếu $\rho_k \neq 0$ thì giữa $Y_t$ và $Y_{t-k}$ có sự tương quan với nhau.

Để biểu diễn sự tự tương quan của một biến theo nhiều độ trễ khác nhau một cách trực quan, ta dùng hàm tự tương quan.

**Định nghĩa 1.2** *Hàm tự tương quan ACF là một đồ thị biểu diễn các hệ số tự tương quan theo các độ trễ khác nhau.*




**Hình** 1.1: Đồ thị của hàm tự tương quan

**b) Hàm tự tương quan riêng**

Xét quá trình dừng $\{X_t; t \in \mathbb{Z}\}$ có kỳ vọng 0. Trong trường hợp tổng quát nhất, $X_t$ tương quan với $X_{n+1}$ thông qua $\{X_2, \ldots, X_n\}$.

Mối tương quan trực tiếp (nếu có) giữa $X_1$ và $X_{n+1}$ mà không qua các biến ngẫu nhiên trung gian $X_2, \ldots, X_n$ gọi là tương quan riêng cấp $n$ giữa $X_1$ và $X_{n+1}$, kí hiệu $\alpha(n)$.

Một cách hợp lý để "khử" các tương quan giữa $X_t$ và $\{X_{t+1}, \ldots, X_n\}$ cũng như khử các tương quan giữa $X_{n+1}$ và $\{X_{t+1}, \ldots, X_n\}$ là tìm xấp xỉ tốt nhất $\hat{X}_t$ của $X_t$ và $\hat{X}_{n+1}$ của $X_{n+1}$ và loại $\hat{X}_t$ ra khỏi $X_t$, $\hat{X}_{n+1}$ ra khỏi $X_{n+1}$ và tìm tương quan các phần còn lại $X_t - \hat{X}_t$ và $X_{n+1} - \hat{X}_{n+1}$.

**Định nghĩa 1.3** *Tự tương quan riêng cấp n giữa* $X_1$ *và* $X_{n+1}$*, kí hiệu* $\alpha(n)$ *là tương quan giữa* $X_1 - \hat{X}_1$ *và* $X_{n+1} - \hat{X}_{n+1}$*, tức là*

$$\alpha(n) = \operatorname{corr}(X_1 - \hat{X}_1, X_{n+1} - \hat{X}_{n+1}) = \frac{\operatorname{cov}(X_1 - \hat{X}_1, X_{n+1} - \hat{X}_{n+1})}{\sqrt{\operatorname{var}(X_1 - \hat{X}_1)}.\sqrt{\operatorname{var}(X_{n+1} - \hat{X}_{n+1})}} \tag{1.2}$$

Ta có $\alpha(1) = \operatorname{corr}(X_1 X_2) = \rho_1$.

Trong định nghĩa trên ta chưa chú ý đến tính dừng của quá trình $X_t$ có kỳ vọng 0 và có $\mathrm{cov}(X_k, X_l)$ không phụ thuộc vào $k, l$ mà chỉ phụ thuộc $|k-l|$. Một cách tổng quát, ta có định nghĩa sau:

**Định nghĩa 1.4** *Cho quá trình dừng $X_t, t \in \mathbb{Z}$, có kỳ vọng 0. Tự tương quan riêng cấp n của quá trình $X_t$, kí hiệu $\alpha(n)$, được cho bởi:*

$$\alpha(n) = \mathrm{corr}(X_t - \hat{X}_t, X_{t-n} - \hat{X}_{t-n}) = \frac{\mathrm{cov}(X_t - \hat{X}_t, X_{t-n} - \hat{X}_{t-n})}{\sqrt{\mathrm{var}(X_t - \hat{X}_t)}.\sqrt{\mathrm{var}(X_{t-n} - \hat{X}_{t-n})}} \tag{1.3}$$

tức là

$$\alpha(n) = \frac{\mathrm{cov}(X_t - \hat{X}_t, X_{t-n} - \hat{X}_{t-n})}{\sqrt{\mathrm{var}(X_t - \hat{X}_t)}}$$

hay

$$\alpha(n) = \frac{\mathrm{cov}(X_t - \hat{X}_t, X_{t-n} - \hat{X}_{t-n})}{\sqrt{\mathrm{var}(X_{t-n} - \hat{X}_{t-n})}} \tag{1.4}$$

trong đó $\hat{X}_t$ là xấp xỉ tuyến tính của $X_t$.

**c) Độ trễ**

Trong phân tích chuỗi thời gian, ta có thể gặp hiện tượng biến phụ thuộc ở thời kỳ $t$ phuộc thuộc vào chính biến đó ở thời kỳ $t-1$, hay nói một cách dễ hiểu biến ngày hôm nay phụ thuộc vào chính biến đó ngày hôm qua. Trong tình huống này thì ta nói biến đó có độ trễ là 1, định nghĩa tương tự cho độ trễ thứ $2, \ldots, k$.

Chẳng hạn khi nghiên cứu mối quan hệ tiêu dùng và thu nhập , ta thấy rằng tiêu dùng ở thời điểm hiện tại chẳng những phụ thuộc vào thu nhập mà còn phụ thuộc vào tiêu dùng ở thời điểm trước đó.

**d) Nhiễu trắng**

Quá trình $\{u_t\}_{t=-\infty}^{\infty}$ được gọi là nhiễu trắng nếu mỗi thành phần của chuỗi có kỳ vọng bằng 0, phương sai không đổi và không tự tương quan, tức là:

$$E(u_t) = 0 \tag{1.5}$$

$$Var(u_t) = \sigma^2 \tag{1.6}$$

$$Cov(u_t, u_{t+s}) = 0, s \neq 0 \tag{1.7}$$

Đôi khi thay điều kiện $Cov(u_t, u_{t+s}) = 0, s \neq 0$ bằng điều kiện mạnh hơn:

$$u_t, u_\tau \text{ độc lập với nhau với } t \neq \tau. \tag{1.8}$$

Quá trình thỏa mãn các điều kiện (1.5), (1.6), (1.7) được gọi là nhiễu trắng độc lập.

Nếu các điều kiện (1.5), (1.6), (1.8) được thỏa mãn và $u_t \sim N(0, \sigma^2)$ thì quá trình ngẫu nhiên được gọi là nhiễu trắng Gauss.

***Lưu ý:** Từ (1.8) có thể suy ra (1.7), điều ngược lại thì không đúng. Nhiễu trắng là một chuỗi dừng.

**e) Chuỗi dữ liệu dừng**

Một chuỗi dữ liệu được gọi là dừng nếu nó có những đặc điểm sau:

- Thể hiện trở lại trạng thái trung bình trong đó dữ liệu dao động xung quanh một giá trị trung bình cố định trong dài hạn.
- Có một phương sai xác định không thay đổi theo thời gian.
- Có một giản đồ tự tương quan ACF với các hệ số tự tương quan giảm dần khi độ trễ tăng lên. Nghĩa là hệ số tự tương quan bậc một khác 0 một cách có ý nghĩa thống kê, nhưng các hệ số tự tương quan bậc hai hoặc bậc ba bằng 0. Như vậy khi quan sát giản đồ tự tương quan, ta thấy các hệ số tự tương quan giảm xuống bằng 0 một cách rất nhanh sau 2 hoặc 3 độ trễ.

Theo ngôn ngữ thống kê, các đặc điểm trên của một chuỗi thời gian $Y_t$ được thể hiện như sau:

- $E(Y_t)$ là một hằng số cho tất cả các thời điểm $t$ khi $E(Y_t) = \mu$.
- $Var(Y_t)$ là một hằng số cho tất cả các thời điểm $t$ khi $Var(Y_t) = \sigma^2$.
- $Cov(Y_t, Y_{t+k})$ là một hằng số cho tất cả các thời điểm $t$ và $k$ ở thời điểm khác 0. Lưu ý, giá trị của hiệp phương sai giữa hai giai đoạn chỉ phụ thuộc vào khoảng cách giữa hai giai đoạn

  $$\gamma_k = Cov(Y_t, Y_{t+k}) = E[(Y_t - \mu)(Y_{t+k} - \mu)],$$

  trong đó $\gamma_k$ là hiệp phương sai ở độ trễ $k$, nghĩa là hiệp phương sai giữa các giá trị $Y_t$ và $Y_{t+k}$. Tóm lại, nếu một chuỗi dừng thì giá trị trung bình, phương sai và hiệp phương sai (ở các độ trễ khác nhau) sẽ giống nhau, điều này có ý nghĩa là các đại lượng này không thay đổi theo thời gian.

### 1.1.3 Các mô hình phổ biến

**a) Mô hình tự hồi quy**

Trong mô hình này, biến phụ thuộc được hồi quy theo các biến trễ của nó. Đơn giản nhất là mô hình sau:

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + u_t. \tag{1.9}$$

trong đó

$\phi_0, \phi_1$ lần lượt là hệ số chặn và hệ số góc,

$u_t$ là một nhiễu trắng.

Khi đó, $Y_t$ được gọi là một quá trình tự hồi quy bậc nhất, ký hiệu AR(1). Giá trị của $Y$ tại thời điểm $t$ phụ thuộc vào chính giá trị của nó ở thời điểm trước đó, các giá trị $Y$ này được xem là sự chênh lệch so với giá trị trung bình.

Trong phương trình (1.1), ta ràng buộc $-1 < \varphi_1 < 1$ để đảm bảo tính dừng của chuỗi thời gian $Y_t$. Nếu $|\varphi_1| > 1$, thì $Y_t$ sẽ có xu hướng ngày càng lớn hơn và vì thế có thể trở thành một chuỗi gia tăng đột biến (chuỗi không dừng).

Mô hình AR($p$) có dạng như sau:

$$Y_t = \varphi_0 + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + ... + \varphi_p Y_{t-p} + u_t. \tag{1.10}$$

Hay dạng rút gọn của nó là $Y_t = \varphi_0 + \sum\limits_{i=1}^{p} \varphi_1 Y_{t-1} + u_t$.

Chuỗi trong mô hình AR($p$) là dừng khi $\sum\limits_{i=1}^{p} \varphi_1 < 1$.

Để xác định độ trễ $p$ ta sử dụng giản đồ tự tương quan theo cách: ACF sẽ có xu hướng bằng 0 ngay sau độ trễ $p$ và hệ số tương quan riêng PACF sẽ có xu hướng khác 0 một cách có ý nghĩa thống kê cho đến độ trễ $p$ đó.

**b) Mô hình trung bình di động**

Ta có mô hình

$$Y_t = \mu + \beta_1 u_t + \beta_2 u_{t-1}, \tag{1.11}$$

trong đó

$\mu$ là hằng số,

$\beta_1, \beta_2$ hệ số ước lượng mức ảnh hưởng của $u_t, u_{t-1}$ lên $Y_t$,

$u_t$ là một nhiễu trắng,

$Y_t$ tại thời điểm $t$ là một hằng số cộng với giá trị trung bình của sai số hiện tại và sai số trong thời điểm trước đó.

Mô hình (1.3) được gọi là một quá trình trung bình di động bậc 1. Kí hiệu MA(1).

Tương tự, ta có dạng: $Y_t = \mu + \beta_0 u_1 + \beta_1 u_{t-1} + \beta_2 u_{t-2}$ được gọi là một quá trình di động bậc 2, kí hiệu MA(2).

Tổng quát, ta có

$$Y_t = \mu + \beta_0 u_1 + \beta_1 u_{t-1} + \beta_2 u_{t-2} + ... + \beta_q u_{t-q}, \tag{1.12}$$

được gọi là quá trình di động bậc $q$. Kí hiệu MA($q$).

Tóm lại, một quá trình trung bình di động chỉ đơn giản là một kết hợp tuyến tính các sai số ngẫu nhiên nhiễu trắng.

Để xác định độ trễ $q$ ta sử dụng giản đồ tự tương quan theo cách: ACF có xu hướng khác 0 một cách có ý nghĩa thống kê cho đến độ trễ $q$ và bằng 0 ngay sau độ trễ $q$ đó. Trong khi đó PACF ngay sau độ trễ $q$ có xu hướng giảm dần về 0.

**c) Mô hình tự hồi quy và trung bình di động**

Trong trường hợp này, dãy số $Y$ bao hàm những đặc điểm của cả AR và MA nên người ta gọi nó là ARMA. Vì vậy, dãy số $Y_t$ được gọi là một quá trình ARMA(1,1) nếu có dạng sau:

$$Y_t = \theta + \phi_1 Y_{t-1} + \beta_0 u_t + \beta_1 u_{t-1}. \tag{1.13}$$

Mô hình (1.5) được gọi là mô hình tự hồi quy và trung bình trượt ARMA(1,1), bởi vì nó bao gồm một quá trình tự hồi quy và quá trình trung bình trượt. Trong mô hình này, $\theta$ đại diện cho một giá trị cố định nào đó.

Nếu kết hợp mô hình AR($p$) với mô hình MA($q$) ta có mô hình ARMA($p, q$) có dạng như sau:

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + + \phi_p Y_{t-p} + u_t + \beta_1 u_{t-1} + \beta_2 u_{t-2} + \ldots + \beta_q u_{t-q}.$$

Trong trường hợp này ta cần xác định độ trễ của $p$ và $q$ thích hợp theo cách như đã trình bày ở các phần trên.

Nói tóm lại trong một quá trình ARMA$(p, q)$ sẽ có quá trình tự hồi quy bậc $p$ và quá trình trung bình trượt bậc $q$.

**d) Mô hình trung bình di động tổng hợp với tự hồi quy**

Mô hình ARIMA là mô hình dự báo cho chuỗi thời gian, mô hình này chỉ dựa trên bản thân chuỗi dữ liệu theo thời gian của chi tiêu muốn dự báo. Một điều đáng lưa ý là để có thể áp dụng mô hình ARIMA thì dãy số liệu nghiên cứu của chúng ta đòi hỏi phải có tính dừng. Tuy nhiên, trong thực tế đa số các dãy số liệu thường không có tính dừng, chính vì vậy chúng ta phải chuyển chúng sang dừng trước khi áp dụng mô hình này.

Mô hình ARIMA gồm hai thành phần:

- AR: Tự hồi quy,

- MA: Trung bình trượt.

Kí hiệu tổng quát của mô hình này là ARIMA $(p, d, q)$ với $p$ là bậc tự hồi quy; $d$ là bậc sai phân (độ lệch), số lần lấy sai phân để chuyển dãy số $Y_t$ từ không dừng thành dừng, $q$ là bậc trung bình trượt.

Phương trình khái quát của mô hình ARIMA $(p, d, q)$ được trình bày dưới dạng sau:

$$Z_t = \delta + \alpha_1 Z_{t-1} + \alpha_2 Z_{t-2} + \ldots + \alpha_p Z_{t-p} + \beta_1 \varepsilon_{t-1} + \ldots + \beta_q \varepsilon_{t-q} + e_t, \tag{1.14}$$

trong đó

$\alpha$: tham số tự hồi quy,

$\beta_i$: tham số trung bình di động,

$\delta = \mu(\beta_1 + \beta_2 + \ldots + \beta_q)$,

$\mu$ giá trị trung bình của chuỗi thời gian,

$e_t$: sai số dự báo, $e_t = \hat{Y}_t - Y$.

***Lưu ý:** Ta có hai trường hợp đặc biệt sau của dữ liệu tại thời điểm $t$:

$Z_t = \delta + \alpha_1 Z_{t-1} + \alpha_2 Z_{t-2} + \ldots + \alpha_p Z_{t-p} + e_t$ là mô hình tự hồi quy thuần ARIMA$(p, d, 0)$ hay AR$(p)$.

$Z_t = \delta + \beta_1 \varepsilon_{t-1} + \ldots + \beta_q \varepsilon_{t-q} + e_t$ là mô hình trung bình di động thuần ARIMA$(0, d, q)$ hay MA$(q)$.

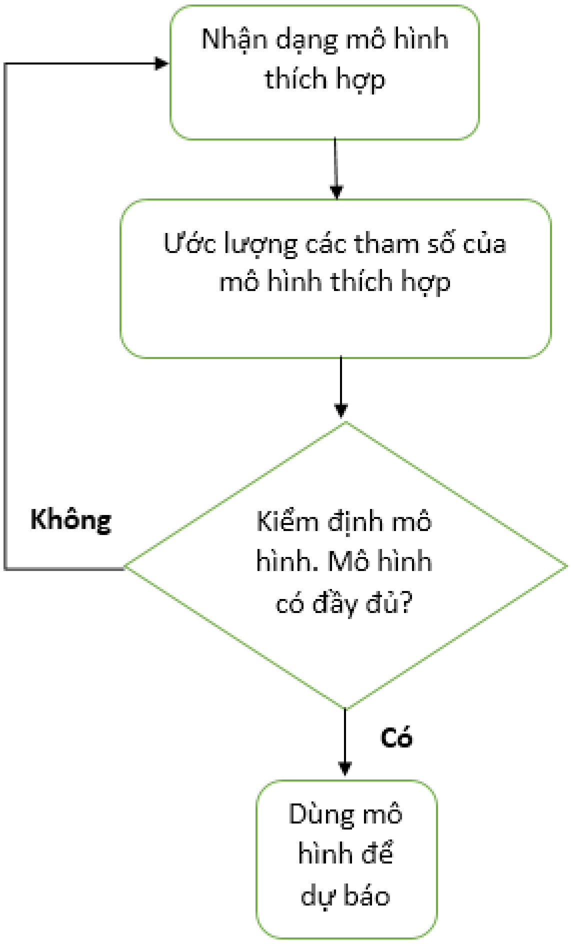
### 1.1.4 Phương pháp Box-Jenkins

Phương pháp phân tích chuỗi thời gian Box và Jenkins đề xuất đã tạo một bước tiến cho các công cụ dự báo. Phương pháp này không dựa vào một hoặc nhiều phương trình mà dựa vào việc phân tích tính ngẫu nhiên của một chuỗi thời gian. Chuỗi thời gian có thể được giải thích ở thời điểm hiện tại, trong quá khứ, các trễ và yếu tố ngẫu nhiên.

Phương pháp Box - Jenkins sử dụng mô hình ARIMA để dự báo một biến bằng cách chỉ xem xét mô hình của chuỗi dữ liệu quá khứ đó. Nó thích hợp cho việc dự báo ngắn hạn.

Khác các phương pháp khác ở chỗ nó không giả định bất kỳ mô hình cụ thể nào trong chuỗi dữ liệu quá khứ sẽ được dự báo. Nó sử dụng phương pháp lặp đi lặp lại để nhận dạng một mô hình thỏa mãn nhất từ nhiều mô hình. Mô hình được chọn sẽ được kiểm chứng với dữ liệu quá khứ để xem có chính xác hay không.

Mục tiêu của phương pháp này là tìm trong số tất cả các mô hình ARIMA, một mô hình thích hợp nhất với số liệu của hiện tượng nghiên cứu. Phương pháp này bao gồm các bước cơ bản được tóm tắt trong sơ đồ sau:




**Hình** 1.2: Quy trình thực hiện của phương pháp Box-Jenkins

### *Bước* 1: Nhận dạng mô hình

Đây là bước quan trọng và cũng là bước khó nhất. Nó cho phép nhận biết được trong tất cả các mô hình ARIMA, mô hình nào có khả năng thích hợp nhất. Điều đó có nghĩa ở bước này ta phải xác định một cách chính xác giá trị của $p, d$ và $q$. Các nguyên tắc sau đây nhằm xác định các thông số $p, d$ và $q$ của mô hình ARIMA.

### a) Xác định d cho mô hình ARIMA

Đặc trưng quan trọng của quá trình ARIMA là quá trình dừng. Nếu mô hình không có tính dừng thì ta tính sai phân.

$$\triangle Y_t = Y_t - Y_{t-1}. \tag{1.15}$$

Tiếp tục lấy sai phân đến lần thứ $d$ thì được mô hình có tính dừng tốt nhất.

Xác định bậc $p$ và $q$ của mô hình ARIMA.

**b) Xác định p, d của mô hình ARMA nhờ vào biểu đồ tương quan**

Nếu biểu đồ tương quan riêng phần (PACF) có giá trị cao tại trễ 1,2,...,p và giảm nhiều sau $p$ hay chỉ có $p$ giá trị đầu tiên là khác 0 (thường lấy $p = 3$ là lớn nhất) và các giá trị của biểu đồ tự tương quan giảm từ từ ta có thể tiên đoán có một AR($p$).

Nếu biểu đồ tự tương quan (ACF) chỉ có $q$ giá trị đầu tiên là khác 0 (thường lấy $q = 3$ là lớn nhất) và các giá trị của biểu đồ tương quan riêng phần giảm từ từ ta có thể tiên đoán có một MA($q$).

Dạng ACF và PACF cho các mô hình ARMA($p, q$) có thể được tóm tắt bởi bảng 1.1:

**Bảng** 1.1: Các dạng ACF và PACF cho mô hình ARMA($p, q$)

| **Mô hình** | **ACF** | **PACF** |
|---|---|---|
| MA(1) | Có ý nghĩa ở độ trễ thứ nhất | Bằng 0 ngay lập tức |
| MA($q$) | Có đỉnh ở $q$ | Giảm dần |
| AR(1) | Bằng 0 ngay lập tức | Có ý nghĩa ở độ trễ thứ nhất |
| AR($p$) | Giảm dần | Có đỉnh ở $p$ |
| ARMA(1,1) | Bằng 0 sau độ trễ thứ nhất | Bằng 0 sau độ trễ thứ nhất |
| ARMA($p, q$) | Bằng 0 ngay độ trễ thứ $q$ | Bằng 0 ngay độ trễ thứ $p$ |

Trong thực hành, phương pháp phân tích đồ thị chỉ cho ta tìm được $p, q$ trong các trường hợp đơn giản. Trong trường hợp tổng quát, ta có thể áp dụng các tiêu chuẩn AIC hoặc SIC để xác định các thông số $p, q$ trong một mô hình ARIMA.

Trong trường hợp lí tưởng, giá trị chọn $p, q$ tương ứng với các trường hợp cho ta các giá trị AIC và SIC cực tiểu. Trong khi áp dụng ta có thể gặp trường hợp ở đó giá trị $p, q$ đề nghị không làm cho hai tiêu chuẩn này đồng thời cực tiểu. Tuy vậy, thường các tiêu chuẩn này cho ta các giá trị $p, q$ tối ưu không khác nhau lớn. Trong trường hợp này ta sẽ khảo sát từng tổ hợp $(p, q)$ cụ thể để quyết định chọn mô hình hợp lí nhất.

**Bước 2: Ước lượng các tham số của mô hình**

Dựa vào mô hình đã chọn, ta dùng các phần mềm thống kê để chạy mô hình và xác định các hệ số ước lượng một cách dễ dàng, hoặc xác định các hệ số ước lượng thông thường ta dùng phương pháp bình phương nhỏ nhất.

**Bước 3: Kiểm định mô hình**

Sau khi các thông số của mô hình được xác định, ta sẽ thực hiện kiểm định trên các kết quả của ước lượng thu được.

Các hệ số của mô hình phải khác 0 (kiểm định $f$). Nếu có một hay nhiều hệ số không thỏa mãn ta sẽ loại bỏ ra khỏi mô hình AR hay MA đang xét.

Phân tích giá trị sai số được thực hiện từ 2 tiêu chuẩn sau:

+ Giá trị trung bình của sai số bằng 0, trong trường hợp ngược lại ta nên thêm một hằng số vào mô hình.

+ Dãy giá trị sai số là một sai số ngẫu nhiên trắng. Các giá trị thống kê của Box-Pierce và của Ljung-Box cho phép kiểm định tính chất này. Nếu nó không phải là một sai số ngẫu nhiên trắng ta kết luận mô hình không hoàn chỉnh và phải thêm vào mô hình các bổ sung cần thiết.

Bước kiểm định mô hình rất quan trọng, ta phải trở lại bước 1 nếu mô hình đề nghị không thích hợp.

**Bước 4: Dự báo**

Một khi mô hình đã được kiểm định, ta có thể dùng nó để dự báo cho tương lai. Một trong những lý do làm cho mô hình ARIMA được thông dụng là nhờ sự thành công của nó trong việc dự báo.

Trong hầu hết các trường hợp, dự báo bằng phương pháp này cho kết quả đáng tin tưởng hơn các phương pháp dự báo cổ điển trước đó, nhất là dự báo ngắn hạn.

## 1.2 Tiêu chuẩn đánh giá mô hình

Cho một chuỗi dự liệu thực tế $\{X_i\}$ và giá trị dự báo của biến tương ứng $\{\hat{X}_i\},\ i = 1, 2, \ldots, n$, với $n$ là số lượng mẫu.

### 1.2.1 Sai số trung bình

$$ME = \frac{1}{n}\sum_{i=1}^{n}(\hat{X}_i - X_i). \tag{1.16}$$

Giá trị của $ME$ nằm trong khoảng $(-\infty, +\infty)$. $ME$ cho biết xu hướng lệch trung bình của giá trị dự báo so với giá trị thực tế, nhưng không phản ánh độ lớn của sai số. $ME$ dương cho biết giá trị dự báo vượt quá giá trị quan trắc và ngược lại. Mô hình được xem là "hoàn hảo" (không thiên lệch về một phía nào cả) nếu $ME = 0$.

### 1.2.2 Sai số tuyệt đối trung bình

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{X}_i - X_i\right|. \tag{1.17}$$

Giá trị MAE nằm trong khoảng $(0, +\infty)$. $MAE$ biểu thị biên độ trung bình của sai số mô hình nhưng không nói lên xu hướng lệch của giá trị dự báo và thực tế. Khi $MAE = 0$, giá trị của mô hình hoàn toàn trùng khớp với giá trị thực tế, mô hình được xem là "lý tưởng".

MAE là một thước đo rất hữu ích khi người phân tích muốn đo lường sai số có cùng một đơn vị tính với dữ liệu gốc. MAE càng nhỏ thì càng tốt tức là giá trị dự báo càng chính xác.

### 1.2.3 Sai số bình phương trung bình

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{X}_i - X_i)^2. \tag{1.18}$$

$MSE$ là trung bình của tổng bình phương của hiệu giữa các giá trị mô hình và thực tế, phản ánh mức độ dao động của sai số. Mô hình là "lí tưởng" nếu $MSE = 0$.

### 1.2.4 Sai số phần trăm tuyệt đối trung bình

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left|\hat{X}_i - X_i\right|}{X_i}\right).100. \tag{1.19}$$

$MAPE$ là thước đo của sai số theo tỷ lệ phần trăm. $MAPE$ có độ nhạy tỷ lệ và không nên được sử dụng khi làm việc với dữ liệu âm lượng thấp. Lưu ý rằng vì "thực tế" nằm trong mẫu số của phương trình (1.25), MAPE không được xác định khi nhu cầu thực tế bằng 0. Hơn nữa, khi giá trị thực tế không bằng 0, nhưng khá nhỏ, $MAPE$ thường sẽ đạt giá trị cực trị. Độ nhạy tỷ lệ này làm cho $MAPE$ gần như vô giá trị như một thước đo lỗi cho dữ liệu âm lượng thấp.

### 1.2.5 Tiêu chuẩn Akaike Information Criterion

Tiêu chuẩn Akaike Information Criterion (AIC) được tính theo công thức sau:

$$AIC = n + ln(\sum e_t^2) + 2z,$$

trong đó:

$n$ là số quan sát của chuỗi,

$z$ là số tham số ước lượng,

$e_t = Y_t - F_t$.

### 1.2.6 Tiêu chuẩn Schwarz Information Criterion

Tiêu chuẩn Schwarz Information Criterion (AIC) được tính theo công thức sau:

$$SIC = n + ln(\sum e_t^2) + zln(n),$$

trong đó các số $n, z$ và $e_t$ có ý nghĩa như ở tiêu chuẩn $AIC$.

## 1.3 Làm trơn số liệu

### 1.3.1 Phương pháp trung bình trượt

Phương pháp này cho phép phát hiện ra quy luật dao động theo chu kỳ của chuỗi.

**a. Trung bình trượt đơn**

Công thức tính giá trị trung bình trượt đơn với khoảng trượt là $k$ được xác định theo công thức:

$$M_t = \frac{1}{k}(x_t + x_{t-1} + x_{t-2} + \ldots + x_{t-k+1}), \tag{1.20}$$

trong đó $M_t$ là trung bình của $k$ quan sát $x_t, x_{t-1}, x_{t-2}, \ldots, x_{t-k+1}$.

Tại mỗi thời kỳ, quan sát cũ nhất bị loại ra và thêm vào một quan sát gần nhất. Ta có thể dùng phương pháp luân phiên để tính trung bình trượt đơn cho thuận tiện hơn. Phương trình (1.6) có thể viết:

$$M_t = \frac{1}{k}(x_t + (x_{t-1} + x_{t-2} + \ldots + x_{t-k+1} + x_{t-k}) - x_{t-k}).$$

Mà $M_{t-1} = (x_{t-1} + x_{t-2} + \ldots + x_{t-k+1} + x_{t-k})$.
Do đó $M_t = M_t - 1 + \dfrac{x_t - x_{t-k}}{k}$.

Nghĩa là có thể tính giá trị của $M_t$ từ giá trị của $M_t$ từ giá trị trước đó $M_{t-1}$.

**b. Trung bình trượt kép**

Trung bình trượt kép là trung bình trượt của trung bình trượt đơn, nghĩa là từ trung bình trượt đơn vừa tính lấy trung bình trượt một lần nữa. Khi đó ta có:

$$M_t^{[2]} = \frac{1}{k}(M_t + M_{t-1} + \ldots + M_{t-k+1}), \tag{1.21}$$

trong đó $M_t^{[2]}$ là trung bình trượt kép (chỉ số [2] ở trên là bậc của trung bình trượt chứ không phải là số mũ); $M_t + M_{t-1} + \ldots + M_{t-k_1}$ là các trung bình trượt đơn.

Tương tự như trên ta cũng có công thức luân phiên cho trung bình trượt kép

$$M_t^{[2]} = M_{t-1}^{[2]} + \frac{1}{k} + (M_t - M_{t-k}).$$

**c. Trung bình trượt trung tâm**

Khác với trung bình trượt đơn và kép ở trên, trung bình trượt trung tâm của thời kỳ hiện tại $t$ được lấy với cả thời kỳ trước và sau, đối xứng qua $t$. (Ví dụ: trung bình trượt của 5 thời kỳ thì lấy trung bình của hai thời kỳ trước, thời kỳ hiện tại và hai thời kỳ sau).

Quá trình thực hiện như sau:

Chọn thời kỳ để tính trung bình $L$. $L$ được chọn tùy thuộc mục đích nghiên cứu. $L$ càng lớn thì càng trơn, nhưng bị mất đi những chu kỳ dao động nhỏ hơn đáng lẽ phải có $L$ càng nhỏ thì biểu hiện dao động càng rõ nhưng lại mang quá nhiều nhiễu loạn, khó phát hiện các chu kỳ. Việc tính trung bình phụ thuộc vài $L$ là chẳn hay lẻ.

Nếu $L$ lẻ thì trung bình trượt trung tâm được tính

$$M_{ct} = \frac{1}{L}\left(x_{t-(L-1)/2} + \ldots + x_t + \ldots x_{t+(L-1)/2}\right), \qquad (1.22)$$

trong đó $x_t$ là điểm giữa của khoảng $L$ các quan sát.

Lưu ý rằng khi đó sẽ mất đi số hạng đầu và số hạng cuối.

Nếu $L$ chẵn thì tính theo hai bước:

- Tính trung bình trượt bao quanh khoảng $L$.

$$M_{t_1}^{[1]} = \frac{1}{L}\left(x_{t-\frac{L}{2}+1} + \ldots + x_t + \ldots + x_{t+\frac{L}{2}}\right), \qquad (1.23)$$

$$M_{t_2}^{[1]} = \frac{1}{L}\left(x_{t-\frac{L}{2}+2} + \ldots + x_t + \ldots + x_{t+\frac{L}{2}+1}\right), \qquad (1.24)$$

  trong đó $M_{t_1}^{[1]}, M_{t_2}^{[1]}$ là các trung bình trượt đơn bao quanh khoảng $L$.

- Sau đó tính trung bình trượt đơn của hai giá trị vừa tính $M_{t_1}^{[1]}, M_{t_2}^{[1]}$ và viết tương ứng với $M_{t_2}^{[1]}$ để cho "trung tâm" của trung bình trượt tương ứng với chuỗi gốc.

$$M_t = \frac{M_{t_1}^{[1]} + M_{t_2}^{[1]}}{2}. \qquad (1.25)$$

### 1.3.2 Phương pháp hàm mũ

Làm trơn hàm mũ là kỹ thuật, trong đó liên tục tính toán lại hoặc giải thích lại những biến đổi hoặc dao động gần với thời điểm xem xét. Những dao động này có thể do sai số ngẫu nhiên, hoặc vốn có bên ngoài không dự tính được. Phương pháp làm trơn này cho phép hiệu chỉnh để có kết quả dự báo chính xác hơn. Làm trơn hàm mũ chỉ thực hiên cho chuỗi dừng. Bài toán này được Kolmogorov đề xuất đầu tiên, sau đó được Viner phát triển.

Trong làm trơn hàm mũ một giá trị trơn được ước lượng mới là tổ hợp của giá trị làm trơn (hay ước lượng) của thời kỳ trước cộng với tỉ lệ của sai số ngẫu nhiên được tạo thành trong thời kỳ trước:

$$x'_{t+1} = x_t + \alpha(e_t), \tag{1.26}$$

Phương trình này được viết dưới dạng:

$$s_t = s_{t-1} + \alpha(x_t - s_{t-1}),$$

trong đó

$s_t$ là giá trị làm trơn (hay ước lượng) mới cho thời kỳ tiếp theo,

$s_{t-1}$ là giá trị làm trơn (hay ước lượng) cho thời kỳ trước,

$x_t$là số liệu thực của chuỗi trước,

$x_t - s_{t-1}$ là sai số,

$\alpha$ là trọng số hoặc hằng số làm trơn.

Sau khi loại bỏ số hạng đồng dạng phương trình (1.12) được viết thành:

$$x'_{t+1} = s_t = \alpha x_t + (1-\alpha)s_{t-1}.$$

Trọng số hay hằng số làm trơn không là chung cho mọi số hạng. Những quan trắc gần nhất có trọng số lớn nhất $\alpha$, quan trắc gần tiếp theo có trọng số $\alpha(1-\alpha)$; tiếp tục ta có trọng số $\alpha(1-\alpha)^2, \alpha(1-\alpha)^3, \ldots$. Như vậy số hạng làm trơn tại thời điểm $t$ có thể viết như sau:

$$x'_{t+1} = s_t = \alpha x_t + \alpha(1-\alpha)x_{t-1} + \alpha(1-\alpha)^2 x_{t-2} + \ldots + \alpha(1-\alpha)^{t-1}x_1 + (1-\alpha)^t s_0 \tag{1.27}$$

**a) Làm trơn bằng hàm mũ đơn**

Thực hiện công thức (1.12) ở dạng truy hồi. Ta xuất phát từ giá trị ban đầu $s_0$ đồng thời cần xác định hằng số làm trơn $\alpha$.

-*Hằng số làm trơn:*

Về lý thuyết, $\alpha$ có thể thay đổi từ 0.01 đến 1. Để xác định $\alpha$ ta thường dùng phép thử sai, sao cho tổng bình phương sai số $\sum(x_t - x'_t)^2$ hay $\sum(x_t - s_{t-1})^2$ là nhỏ nhất. Giá trị ban đầu có thể phán đoán bằng sự so sánh giữa phương pháp làm trơn và trung bình trượt, khi đó ta có $\alpha = \dfrac{2}{L+1}$ trong đó $L$ là độ dài thời kỳ làm trơn.

-*Ước lượng ban đầu $s_0$:*

Ước lượng này thường được lấy bằng trung bình số học của cả chuỗi.

**b) Làm trơn bằng hàm mũ kép**

Đối với chuỗi có xu thế tuyến tính thực hiện theo phương trình

$$s_t^{(2)} = \alpha s_t + (1-\alpha)s_{t-1}^{(2)}, \tag{1.28}$$

trong đó $s_t^{(2)}$ là giá trị làm trơn hàm mũ kép; $s_t$ là giá trị làm trơn hàm mũ đơn, tính theo (1.12) ở trên.

Giá trị ban đầu $s_0$ được xác định:

$$s_0 = a - \frac{1-\alpha}{\alpha}b,$$
$$s_0^{(2)} = a - 2\left(\frac{1-\alpha}{\alpha}\right)b,$$

trong đó $a, b$ là các hệ số xác định theo hồi quy tuyến tính giữa các giá trị của chuỗi và của thời gian $t$: $x = a + bt$.

Đối với chuỗi có xu thế đường cong bậc hai, ta tính theo công thức

$$s_t^{(3)} = \alpha s_t^{(2)} + (1-\alpha)s_{t-1}^{(3)},$$

trong đó $s_t^{(3)}$ là giá trị làm trơn hàm mũ kép, $s_t^{(2)}$ và $s_t$ là các giá trị làm trơn tính theo công thức ở trên.

Các ước lượng ban đầu được tính như sau:

$$
\begin{aligned}
s_0 &= a - \frac{1-\alpha}{\alpha} b_1 + \frac{(1-\alpha)(2-\alpha)}{2\alpha^2} 2b_2,\\
s_0^{(2)} &= a - \frac{2(1-\alpha)}{\alpha} b_1 + \frac{2(1-\alpha)(3-2\alpha)}{2\alpha^2} 2b_2,\\
s_0^{(3)} &= a - \frac{3(1-\alpha)}{\alpha} b_1 + \frac{3(1-\alpha)(4-3\alpha)}{2\alpha^2} 2b_2,
\end{aligned}
$$

trong đó $a, b_1, b_2$ là các hệ số hồi quy bậc hai của chuỗi số liệu với thời gian $t = a + b_1 t + b_2 t^2$.

## 1.4 Mô hình của Abbasov-Manedova

### 1.4.1 Mô hình

Đây là mô hình của Abbasov và Mamedova, Đại học Công nghệ Vienna, đề xuất năm 2003. Hai tác giả đã ứng dụng chuỗi thời gian mờ để xây dựng mô hình dự báo gồm 5 bước sau:

**Bước 1.** Xác định tập nền $U$ là đoạn thời gian giữa các biến đổi nhỏ nhất và lớn nhất trong tổng số dữ liệu.

**Bước 2.** Chia tập $U$ thành $n$ đoạn có độ dài bằng nhau, đồng thời tính các giá trị trung bình của từng đoạn $(u_m^i, i = 1, ..., n)$.

**Bước 3.** Mô tả chất lượng của các giá trị biến đổi số liệu như là một biến ngôn ngữ theo Công thức (1.29).

$$
A^t = \left[ \frac{\mu_{A_i}(u_i)}{u_i} \right], u_i \in U, \mu_{A_i} \in [0; 1], \tag{1.29}
$$

$$
\mu_{A_i}(u_i) = \frac{1}{1 + \left[C \times (U - u_m^i)\right]^2},
$$

trong đó

$A^t$ là mờ hóa các biến của năm $t$,

$C$ là hằng số tự chọn sao cho $\mu_{A_i}(u_i) \in [0; 1]$,

$U$ là các biến đổi của từng năm,

$u_m^i$ là giá trị trung bình của từng đoạn thứ $i$.

**Bước 4.** Lựa chọn tham số $w\,(1 < w < n)$ là số năm của dữ liệu ban đầu tương ứng với đoạn thời gian trước khi sang năm có liên quan, tính toán các mối quan hệ mờ của ma trận $R(t)$ theo công thức sau:

$$R(t)\,[i,j] = O^w\,[i,j] \cap K(t)\,[j]\,,$$

hay viết gọn lại

$$R(t) = O^w(t) \otimes K(t) = \begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1j} \\ R_{21} & R_{22} & \cdots & R_{2j} \\ \cdots & \cdots & \cdots & \cdots \\ R_{i1} & R_{i2} & \cdots & R_{\mathrm{ij}} \end{bmatrix}. \tag{1.30}$$

Thiết lập các tập mờ $F(t)$ bằng toán tử max trong các cột của ma trận $R(t)$.

$$F(t) = [\max(\mathrm{R_{11}}, \mathrm{R_{21}}, \ldots, \mathrm{R_{i1}}) \max(\mathrm{R_{21}}, \mathrm{R_{22}}, \ldots, \mathrm{R_{i2}}) \ldots \max(\mathrm{R_{1j}}, \ldots, \mathrm{R_{ij}})]\,, \tag{1.31}$$

trong đó $i = 1, ..., w; j = 1, ..., n$.

**Bước 5.** Giải mờ kết quả thu được hoặc chuyển đổi các giá trị mờ vào các giá trị định tính. Dự báo sự biến đổi của dữ liệu trong năm tới $V(t)$.

$$V(t) = \frac{\sum\limits_{i=1}^{w} \mu_t(u_i) \times u_m^i}{\sum\limits_{i=1}^{w} \mu_t(u_i)}. \tag{1.32}$$

Kết quả dự báo cho năm thứ $t$ được tính theo công thức sau:

$$N(t) = N(t-1) + V(t), \tag{1.33}$$

trong đó

$N(t)$ là số liệu của năm $t$,

$V(t)$ là số liệu biến đổi từ năm $t-1$ đến năm $t$.

### 1.4.2 Ví dụ minh họa

Trong phần này, luận văn sử dụng dữ liệu tuyển sinh của trường Đại học Alabama từ 1971 đến 1992 để làm ví dụ minh họa cho thuật toán, với các tham số được chọn để thực hiện là $n = 7, w = 7, C = 0,0001$. Tập dữ liệu này được cho cụ thể bởi Bảng 1.2.

**Bảng** 1.2: Dữ liệu thực về tuyển sinh của trường Đại học Alabama từ 1971 đến 1992

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

**Bước 1.** Tính sự biến thiên của số liệu tuyển sinh giữa hai năm liên tiếp, kết quả của sự biến đổi này được thể hiện cụ thể ở Bảng 1.3.

**Bảng** 1.3: Kết quả sự biến đổi của hai năm liên tiếp tuyển sinh của trường Đại học Alabama từ 1971 đến 1992

| Năm | Biến đổi | Năm | Biến đổi | Năm | Biến đổi |
|---|---|---|---|---|---|
| 1971 | - | 1979 | 946 | 1987 | 875 |
| 1972 | 508 | 1980 | 112 | 1988 | 1291 |
| 1973 | 304 | 1981 | -531 | 1989 | 820 |
| 1974 | 829 | 1982 | -955 | 1990 | 358 |
| 1975 | 764 | 1983 | 64,0 | 1991 | 9,00 |
| 1976 | -149 | 1984 | -352 | 1992 | -461 |
| 1977 | 292 | 1985 | 18,0 | | |
| 1978 | 258 | 1986 | 821 | | |

Từ Bảng 1.3, ta có giá trị nhỏ nhất là $-955$ và lớn nhất là 1291. Do đó tập nền $U = [-955; 1291]$.

**Bước 2:** Chia tập nền $U$ thành 5 đoạn có độ dài bằng nhau và tính điểm giữa của mỗi đoạn, ta có kết quả ở Bảng 1.4.

**Bảng** 1.4: Các đoạn chia của tập nền $U$ và điểm giữa của chúng theo mô hình AM

| Các khoảng chia | Giá trị | Điểm giữa | Giá trị |
|---|---|---|---|
| $u_1$ | [-955; -634,143 ] | $u_1^0$ | -794,571 |
| $u_2$ | [-634,143; -313,286 ] | $u_2^0$ | -473,714 |
| $u_3$ | [-313,286; 7,571 ] | $u_3^0$ | -152,857 |
| $u_4$ | [7,571429; 328,429] | $u_4^0$ | 168 |
| $u_5$ | [328,429; 649,286] | $u_5^0$ | 488,857 |
| $u_6$ | [649,286; 970,143] | $u_6^0$ | 809,714 |
| $u_7$ | [970,143; 1291] | $u_7^0$ | 1130,571 |

**Bước 3.** Với $w = 7$, áp dụng Công thức (1.29) ta có kết quả ở Bảng 1.5.

**Bảng** 1.5: Mờ hóa số liệu tuyển sinh của trường đại học Alabama giai đoạn 1971-1992

| Kết quả mờ hóa |
|---|
| $A^{1972}$={(0,983316150331/u1),(0,990454367876/u2),(0,995651668927/u3), (0,998845334792/u4),(0,999996335523/u5),(0,999090512817/u6), (0,996139013162/u7)} |
| $A^{1973}$={(0,988075322232/u1),(0,993987968042/u2),(0,997917162775/u3), (0,999815074203/u4),(0,999658395101/u5),(0,997449054577/u6), (0,993214158979/u7)} |
| $A^{1974}$={(0,974317153923/u1),(0,983312551661/u2),(0,990451616079/u3), (0,995649796950/u4),(0,998844365403/u5),(0,999996280626/u6), (0,999091373088/u7)} |
| $A^{1975}$={(0,976284631679/u1),(0,984911775510/u2),(0,991663806094/u3), (0,996460413178/u4),(0,999243536751/u5),(0,999979102477/u6), (0,998658057108/u7)} |
| $A^{1976}$={(0,995849672314/u1),(0,998946716901/u2),(0,999999851224/u3), (0,998996118790/u4),(0,995947869222/u5),(0,990892380329/u6), |

(0,983890727205/u7)}

$A^{1977}$={(0,988331389296/u1),(0,994170992865/u2),(0,998024929846/u3),

(0,999846263638/u4),(0,999612622772/u5),(0,997326883885/u6),

(0,993017083602/u7)}

$A^{1978}$={(0,989042334577/u1),(0,994674455315/u2),(0,998314808745/u3),

(0,999919006560/u4),(0,999467333680/u5),(0,996965350597/u6),

(0,992443723069/u7)}

$A^{1979}$={(0,970594962763/u1),(0,980242346491/u2),(0,988069192744/u3),

(0,993983576449/u4),(0,997914562220/u5),(0,999814296532/u6),

(0,999659449891/u7)}

$A^{1980}$={(0,991848279140/u1),(0,996581116618/u2),(0,999298998686/u3),

(0,999968640983/u4),(0,998581801083/u5),(0,995155530883/u6),

(0,989731655268/u7)}

$A^{1981}$={(0,999305783292/u1),(0,999967184546/u2),(0,998572121548/u3),

(0,995137747016/u4),(0,989705983036/u5),(0,982342252661/u6),

(0,973133538418/u7)}

$A^{1982}$={(0,999742692958/u1),(0,997688993733/u2),(0,993606804308/u3),

(0,987545773857/u4),(0,979578497364/u5),(0,969798378772/u6),

(0,958316966787/u7)}

$A^{1983}$={(0,992682491659/u1),(0,997116969367/u2),(0,999529950846/u3),

(0,999891851697/u4),(0,998198216365/u5),(0,994469854615/u6),

(0,988752205927/u7)}

$A^{1984}$={(0,998045134291/u1),(0,999851878269/u2),(0,999603578436/u3),

(0,997303291898/u4),(0,992979232340/u5),(0,986683910475/u6),

(0,978492557067/u7)}

$A^{1985}$={(0,993440586725/u1),(0,997588002410/u2),(0,999708163560/u3),
(0,999775050613/u4),(0,997787840011/u5),(0,993770929429/u6),
(0,987773193431/u7)}

$A^{1986}$={(0,974563207935/u1),(0,983513509760/u2),(0,990605123449/u3),
(0,995754015261/u4),(0,998898026908/u5),(0,999998726328/u6),
(0,999042572851/u7)}

$A^{1987}$={(0,972881239426/u1),(0,982134673444/u2),(0,989545547218/u3),
(0,995026370636/u4),(0,998511156897/u5),(0,999957379571/u6),
(0,999347258798/u7)}

$A^{1988}$={(0,958316966787/u1),(0,969798378772/u2),(0,979578497364/u3),
(0,987545773857/u4),(0,993606804308/u5),(0,997688993733/u6),
(0,999742692958/u7)}

$A^{1989}$={(0,974593887940/u1),(0,983538548236/u2),(0,990624226896/u3),
(0,995766954844/u4),(0,998904645199/u5),(0,999998942041/u6),
(0,999036383330/u7)}

$A^{1990}$={(0,986889947694/u1),(0,993130036350/u2),(0,997397042863/u3),
(0,999639130273/u4),(0,999828793398/u5),(0,997963697031/u6),
(0,994066747480/u7)}

$A^{1991}$={(0,993584158415/u1),(0,997675286071/u2),(0,999738091267/u3),
(0,999747253896/u4),(0,997702661143/u5),(0,993629410935/u6),
(0,987577046430/u7)}

$A^{1992}$={(0,998888537743/u1),(0,999998383472/u2),(0,999051380529/u3),
(0,996059181493/u4),(0,991058387862/u5),(0,984109439275/u6),
(0,975294810926/u7)}

---

**Bước 4,** Dựa vào các Công thức (1.30) và (1.31), chúng tôi dự báo số lượng tuyển sinh năm 1979 như sau:

$$O^7(1979) = \begin{bmatrix} 0,983316150331 & 0,990454367876 & \ldots & 0,999090512817 & 0,996139013162 \\ 0,988075322232 & 0,993987968042 & \ldots & 0,997449054577 & 0,993214158979 \\ 0,974317153923 & 0,983312551661 & \ldots & 0,999996280626 & 0,999091373088 \\ 0,976284631679 & 0,984911775510 & \ldots & 0,999979102477 & 0,998658057108 \\ 0,995849672314 & 0,998946716901 & \ldots & 0,990892380329 & 0,983890727205 \\ 0,988331389296 & 0,994170992865 & \cdots & 0,997326883885 & 0,993017083602 \end{bmatrix}.$$

$K(1979) = [0,989042334577\ 0,994674455315 \ldots 0,996965350597\ 0,992443723069]$.

Thực hiện toán tử min ta thu được:

$$R(1979) =$$

$$\begin{bmatrix} 0,983316150331 & 0,99045436787 \ldots 0,996965350597 & 0,992443723069 \\ 0,988075322232 & 0,993987968042 \ldots 0,996965350597 & 0,992443723069 \\ 0,974317153923 & 0,983312551661 \ldots 0,996965350597 & 0,992443723069 \\ 0,976284631679 & 0,984911775510 \ldots 0,99696530597 & 0,992443723069 \\ 0,989042334577 & 0,994674455315 \ldots 0,990892380329 & 0,983890727205 \\ 0,988331389296 & 0,994170992865 \ldots 0,996965350597 & 0,992443723069 \end{bmatrix}.$$

Thực hiện toán tử max ta có:

$$F(1979) =$$

$$[0,989042334577\ 0,994674455315\ \ldots\ 0,996965350597\ 0,992443723069].$$

**Bước 5.** Dự báo số lượng tuyển sinh năm 1979 theo các Công thức (1.32) và (1.33), ta có:

$$V(1979) = \frac{0,989042334577 \times (-794,5714) + \cdots + 0,992443723069 \times 1130,5714}{0,989042334577 + 0,994674455315 + \cdots + 0,992443723069}$$
$$= 168,7336.$$

$$N(1979) = 15861 + 168,7336 = 16029,73.$$

Tính toán tương tự cho các năm tiếp theo, kết quả dự báo được trình bày cụ thể ở Bảng 1.7.

**Bảng** 1.7: Dự báo số liệu tuyển sinh bằng mô hình AM

| Năm | Số liệu | | | | | | |
|---|---|---|---|---|---|---|---|
| thực tế | Dự báo | Sai số | Năm | Số liệu | | | |
| thực tế | Dự báo | Sai số | | | | | |
| 1971 | 13055 | - | - | 1982 | 15433 | 16550,90 | 162,8994 |
| 1972 | 13563 | - | - | 1983 | 15497 | 15592,02 | 159,0173 |
| 1973 | 13867 | - | - | 1984 | 15145 | 15664,16 | 167,1629 |
| 1974 | 14696 | - | - | 1985 | 15163 | 15308,78 | 163,7755 |
| 1975 | 15460 | - | - | 1986 | 15984 | 15329,78 | 166,7823 |
| 1976 | 15311 | - | - | 1987 | 16859 | 16155,56 | 171,5553 |
| 1977 | 15603 | - | - | 1988 | 18150 | 17032,69 | 173,6934 |
| 1978 | 15861 | - | - | 1989 | 18970 | 18326,99 | 176,9885 |
| 1979 | 16807 | 16029,73 | 168,7336 | 1990 | 19328 | 19143,29 | 173,2928 |
| 1980 | 16919 | 16981,23 | 174,2266 | 1991 | 19337 | 19497,51 | 169,5060 |
| 1981 | 16388 | 17086,54 | 167,5435 | 1992 | 18876 | 19503,73 | 166,7333 |

Kết quả dự báo số lượng tuyển sinh từ năm 1993 đến năm 1997 được thể hiện ở Bảng 1.8 và Hình **??**.

**Bảng** 1.8: Dự báo tuyển sinh từ năm 1993 đến năm 1997

| Năm | Dự báo |
|---|---|
| 1993 | 19039,72 |
| 1994 | 19207,69 |
| 1995 | 19375,69 |
| 1996 | 19543,69 |
| 1997 | 19711,69 |

Nhìn vào Hình **??** và Bảng 1.8, ta thấy trong 5 năm tới số lượng tuyển sinh sẽ tăng dần nhưng không nhiều.

## 1.5 Mô hình dự báo cải tiến của Singh

### 1.5.1 Mô hình

Cho $X(t), t = 1, 2, ..., n$ là một chuỗi thời gian. Mô hình dự báo chuỗi thời gian mờ (FTSF) đề nghị gồm 6 bước sau:

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

$$K_i = 100X_i/\max\{X_i\}, \quad i = 1, 2, ..., n. \tag{1.34}$$

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

$$E_t = K_{t+1} - K_t, t = 1, 2, ..., n - 1. \tag{1.35}$$

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 và gán tập mờ $A_i,\ i = \overline{1, m}$ với các hàm thuộc như sau:

$$A_1 = \{1/u_1, 0.5/u_2, 0/u_3, ..., 0/u_{m-1}, 0/u_m\},$$
$$A_2 = \{0.5/u_1, 1/u_2, 0.5/u_3, ..., 0/u_{m-1}, 0/u_m\},$$
$$...............$$
$$A_m = \{0/u_1, 0/u_2, 0/u_3, ..., 0.5/u_{m-1}, 1/u_m\}.$$

*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:
  $[A_j]$ là khoảng tương ứng của $u_j$,

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

$l[A_j]$ là độ dài của khoảng $A_j$,

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

- Tính các giá trị:

  $D_i = ||E_i - E_{i-1}| - |E_{i-1} - E_{i-2}||, \quad Z_i = E_i + \dfrac{D_i}{2},$
  $ZZ_i = E_i - \dfrac{D_i}{2}, \quad Y_i = E_i + D_i, \quad YY_i = E_i - D_i,$
  $P_i = E_i + \dfrac{D_i}{4}, \quad PP_i = E_i - \dfrac{D_i}{4}, \quad Q_i = E_i + D_i \times 2,$
  $QQ_i = E_i - D_i \times 2, \quad G_i = E_i + \dfrac{D_i}{6}, \quad GG_i = E_i - \dfrac{D_i}{6},$
  $H_i = E_i + D_i \times 3, \quad HH_i = E_i - D_i \times 3.$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\hat{V}(t) = \frac{\sum P_j \times V_j}{\sum P_j}, \tag{1.36}$$

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,

$$V_j = \frac{R + M[A_j]}{S + 1}. \tag{1.37}$$

*Bước 6:* Tính giá trị dự báo

$$\hat{K}(t) = K(t-1) + \hat{V}(t), \tag{1.38}$$

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 $t$,

### 1.5.2 Ví dụ minh họa

Trong phần phần này chúng tôi lấy số liệu về tuyển sinh của trường đại học Alabama (EAU) 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 $X_i$ của Bảng 1.9.

**Bảng** 1.9: Số liệu gốc, số liệu chuẩn hóa, số liệu biến đổi và tập mờ tương ứng của dữ liệu EAU.

| 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 1.9. Tính sự biến đổi số lượng sinh viên giữa hai năm liên tiếp của $\{K_i\}$, ta có giá trị $\{E_i\}$ của Bảng 1.9. Vì $D_{min} = -4.939, D_{max} = 6.6.76$, 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 1.10.

**Bảng** 1.10: Các đoạn chia của tập nền.

| $u_i$ | $u_m^i$ |
|---|---|
| [-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;1.698] | 1.284 |
| [1.698;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 $A_i$ tương ứng với từng đoạn $u_i$ được cho bởi Bảng 1.9 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 1.11:

**Bảng** 1.11: Mối quan hệ mờ của các $A_i$

| | | | | |
|---|---|---|---|---|
| $A_1 \to A_4 : 1$ | | | | |
| $A_2 \to A_1 : 1$ | $A_2 \to A_4 : 1$ | | | |
| $A_3 \to A_5 : 1$ | | | | |
| $A_4 \to A_2 : 3$ | $A_4 \to A_8 : 1$ | | | |
| $A_5 \to A_5 : 1$ | $A_5 \to A_8 : 2$ | | | |
| $A_6 \to A_4 : 1$ | $A_6 \to A_5 : 1$ | | | |
| $A_7 \to A_3 : 1$ | | | | |
| $A_8 \to A_4 : 1$ | $A_8 \to A_6 : 1$ | $A_8 \to A_7 : 1$ | $A_8 \to A_8 : 1$ | $A_8 \to A_9 : 1$ |
| $A_9 \to A_8 : 1$ | | | | |

**Bước 5:** Tính sự biến đổi

Giả sử chúng ta cần phải dự báo giá trị chuỗi năm 1976. Dựa bào Bảng 1.9, Bảng 1.10 và Bảng 1.11, 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:

$$[A_3] = [-1.620; 0.039], \quad U[A_3] = 0.039, \quad L[A_3] = -1.620,$$

$$M[A_3] = -0.791, \quad E_4 = 4.287, \quad E_3 = 1.572, \quad E_2 = 2.627.$$

Khi đó chúng ta nhận được:

$$
\begin{aligned}
&D_5 = |E_5 - E_4| - |E_4 - E_3| = 1.660, Z_5 = E_5 + \frac{D_5}{2} = 5.140,\\
&ZZ_5 = E_5 - \frac{D_5}{2} = 2.762, Y_5 = E_5 + D_5 = 6.330,\\
&YY_5 = E_5 + D_5 = 1.572, \quad P_5 = E_5 + \frac{D_5}{4} = 4.546,\\
&PP_5 = E_5 - \frac{D_5}{4} = 3.356, \quad Q_5 = E_5 + D_5 \times 2 = 8.709,\\
&QQ_5 = E_5 - D_5 \times 2 = -0.807, \quad G_5 = E_5 + \frac{D_5}{6} = 4.348,\\
&GG_5 = E_5 - \frac{D_5}{6} = 3.555, \quad H_5 = E_5 + D_5 \times 3 = 11.088,\\
&HH_5 = E_5 - D_5 \times 3 = -3.186.
\end{aligned}
$$

Ta xét:

+ $Z_5 > L[A_3]$ và $Z_5 > U[A_3];$ khi đó $R = 0$ và $S = 0$.

+ $ZZ_5 > L[A_3]$ và $ZZ_5 > U[A_3];$ 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 $R = 0 + QQ_5 = -0.8067$ và $S = 0 + 1 = 1$.

+ $H_5 > L[A_3]$ 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 (2.6), 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 (2.5) 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 (2.7):

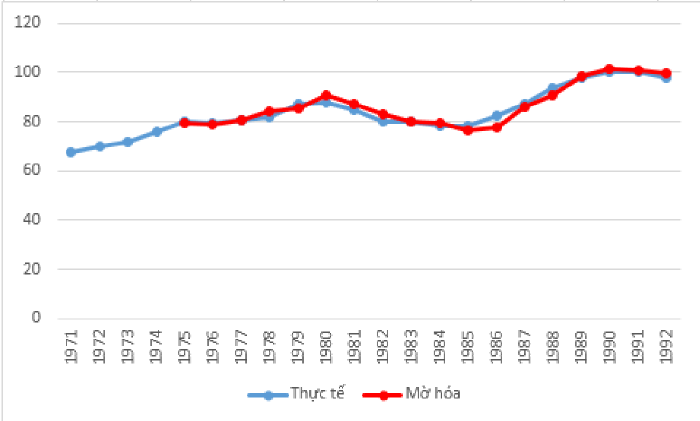
$$\hat{K}(t) = K(t-1) + \hat{V}_{final} = 79.9504 + (-0.7986) = 79.1518$$

Thực hiện tương tự cho các năm tiếp theo cho đến năm 1992 ta được Bảng 1.12:

**Bảng** 1.12: Số lượng tuyển sinh được mờ hóa theo mô hình đề nghị

| Năm | Thực tế | Chuẩn hóa | Mờ hóa | Năm | Thực tế | Chuẩn hóa | Mờ hóa |
|---|---|---|---|---|---|---|---|
| 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 | 100.000 | 100.822 |
| 1981 | 16388 | 84.7494 | 87.0952 | 1992 | 18876 | 97.6160 | 99.5919 |

Từ số liệu Bảng 1.12, áp dụng công thức (1.23) ta tính được giá trị của MAE = 1.6508. Kết quả thực hiện được cho bởi Hình 1.3




**Hình** 1.3: Đồ thị số liệu thực tế và dự báo của dữ liệu Enrollment

## A. CODE MÔ HÌNH DỰ BÁO ARIMA

```
dt=read.csv(file.choose(),header=F)

dt

is.ts(dt)

dt=ts(dt[,1])

dt

is.ts(dt)

library(urca)

library(AnalyzeTS)

summary(ur.df(dt,type="drift"))

dt1=diff(dt)

Descriptives(dt1,plot=TRUE,ans=2,r=0)

summary(ur.df(dt1,type="none"))

dt2=diff(dt1)

Descriptives(dt2,plot=TRUE,ans=2,r=0)

summary(ur.df(dt2,type="none"))

library(AnalyzeTS)

Descriptives(dt,answer=1,plot=TRUE,r=3)
```

```
PrintAIC(dt,order=c(1,1,1),type="ARIMA")
arima(dt,order=c(1,0,1),seas=list(order=c(1,0,1),12))
fit<-arima(dt,order=c(0,1,1))
fit
ts.plot(dt,fitted(fit),col=c("blue","red"),main="",xlab="Thoi gian",ylab="Ty
dong")
legend("topleft",c("Doanh thu","Du bao"),lty=c(1,1),col=c("blue","red"))
pred=predict(fit,n.ahead=3)
pred
ts.plot(dt,pred$pred,lty=c(1,2),ylab="Ty dong", xlab="Thoi gian",col="blue")
legend("topleft",c("Doanh thu","Doanh thu du bao"),lty=c(1,2), col="blue")
# Tìm n
D1=0
D2=0
ts1=diff(st)
ts1<-as.vector(ts1)
min.x= min(ts1)-D1
max.x = max(ts1)-D2
tong_temp = 0
MMSE_list = c()
for (n in 1:11) {
h=((max.x - min.x)/n)
k <- 1:(n+1)
U <- 1:n
for (i in 1:(n+1)){
if (i==1)
```

```
k[i] = min.x
else {
k[i] = min.x + (i-1) * h
U[i-1] = paste("U", i-1, sep="")
}
}

D <- data.frame(U, low = k[1:n], up = k[2:(n+1)])
D$Bw <- (1/2) * (D$low+D$up)

tong_temp = 0
for (v_i in 1:n) {
tong=0
v=ts1[ts1 >= D$low[v_i] & ts1 <= D$up[v_i]]
# if (v_i == n) {
# v=ts1[ts1 > D$low[v_i] & ts1 <= D$up[v_i]]
# } else if (v_i == 1) {
# v=ts1[ts1 >= D$low[v_i] & ts1 < D$up[v_i]]
# } else {
# v=ts1[ts1 > D$low[v_i] & ts1 < D$up[v_i]]
# }

for (i in 1:length(v)){
tong=tong+sum(((v[i]-D$Bw[v_i])2̂)/length(v))
}
tong_temp = tong_temp + tong }
```

```
MMSE = (1/n) * tong_temp
MMSE_list = c(MMSE_list, MMSE)
}
n_best = match(min(MMSE_list, na.rm = TRUE),MMSE_list)
print("N tot nhat la:")
print(n_best)

index = c()
for (i in 1:length(MMSE_list)){
index = c(index, i)
}
df <- data.frame(index,MMSE_list)
colnames(df) <- c('n','MMSE')
library(ggplot2)
ggplot(df, aes(x = n, y = MMSE))+ geom_line() + geom_point() +
scale_x_continuous("n", labels = as.character(index), breaks = index)
print(MMSE_list)
# Tìm w
ts2=diff(st)
ts2
pacf(ts2,lag.max=17,plot=T)
# Tìm C
hesoC1=DOC(st,n=4,w=2,error=0.01,CEF="MAE",
type="Abbasov-Mamedova")
C1=as.numeric(hesoC1[1])
```

## B. MÔ HÌNH CẢI TIẾN CỦA AM

```
mhMAM=fuzzy.ts2(st,n=4,w=5,C=C1,forecast=5,
type="Abbasov-Mamedova",trace =TRUE)
```

**Thuật toán chùm mờ**

```
dl=read.csv(file.choose(),header=TRUE)
library(rdist)
st=dl[,2]
st=matrix(st)
M=max(st)
chuanhoa=round((st*100/M),digits=6)
zt=diff(chuanhoa)
library(rdist)
k=data.frame()
chum=function(zt){
d=pdist(zt,metric="euclidean",p=2)
ds=sum(d)/(dim(d)[1]*dim(d)[1]-dim(d)[1])
for (i in 1:dim(d)[1]){
for ( j in 1:dim(d)[2]){
if (d[i,j]<= ds) {
k[i,j] =exp(-d[i,j]/(ds/50))
} else{
k[i,j]=0
}
}
}
```

```
ztmoi=matrix(nrow=dim(zt)[1],ncol=dim(zt)[2])
for (l in 1:dim(zt)[2]){
for (i in 1:dim(d)[1]){
tu=0
mau=0
for (j in 1:dim(d)[1]){
tu=tu+zt[j,l]*k[i,j]
mau=mau+k[i,j]
}
ztmoi[i,l]=tu/mau
}
}
zt=as.matrix(zt)
sodem=0
epxilanh=0.01
while (max(abs(ztmoi-zt))>= epxilanh){
zt=ztmoi
d= pdist(zt,metric="euclidean",p=2)
for (i in 1:dim(d)[1]){
for (j in 1:dim(d)[1]){
if (d[i,j]<=ds){
k[i,j]=exp(-d[i,j]/(ds/50))
}else{
k[i,j]=0
}
}
```

```
}
for (l in 1:dim(zt)[2]){
for (i in 1:dim(d)[1]){
tu=0
mau=0
for (j in 1:dim(d)[1]){
tu=tu+zt[j,l]*k[i,j]
mau= mau+k[i,j]
}
ztmoi[i,l]=tu/mau
}
}
ztmoi
sodem=sodem+1
}
print(ztmoi)
vchum=unique(ztmoi)
vchum
A=vchum[order(vchum[,1],decreasing=FALSE),]
vchum=matrix(A,nrow=dim(vchum)[1],ncol=dim(vchum)[2],byrow=FALSE)
print(vchum)
print(sodem)
# khoi tao ma tran U
U=matrix(nrow=dim(vchum)[1],ncol=dim(zt)[1])
for (i in 1:dim(vchum)[1]){
for (j in 1:dim(ztmoi)[1]){
```

```
if (ztmoi[j,1]==vchum[i,1]){
U[i,j]=1
}else{
U[i,j]=0
}
}
} print(U)
matranU0=data.frame(U)
matranU0
# TÌm số chùm
chuanhoa=round((st*100/M),digits=6)
chuanhoa
Z=diff(chuanhoa)
Z
diemgiua=data.frame()
for (i in 1:dim(vchum)[1]){
sum=0
t=0
for (j in 1:dim(ztmoi)[1]){
if (ztmoi[j,1]==vchum[i,1]){
sum=sum+Z[j,1]
t=t+1
}
}
diemgiua[1,i]=sum/t
}
```

```
diemgiua=round(diemgiua,digits=4)
print(diemgiua)
mid=matrix(diemgiua,nrow=1,ncol=dim(vchum)[1])
mid
# tim ma tran U qua cac vong lap
z2=t(Z)
n=dim(z2)[1]
N=dim(z2)[2]
c=dim(U)[1]
v=data.frame()
for (k in 1:n){
for (i in 1:c){
tu=0
mau=0
for ( j in 1:N){
tu=tu+(U[i,j]^2)*z2[k,j]
mau=mau+U[i,j]^2
}
v[k,i]=tu/mau
}
}
v #tim d
z2=t(Z)
z2
d2=matrix(0,nrow=c,ncol=N)
for (j in 1:N){
```

```
for (i in 1:c){
d2[i,j]=0
for (k in 1:n){
d2[i,j]=d2[i,j]+(v[k,i]-z2[k,j])^2
}
}
}
d2
d2=round(d2,digits=4)
# cap nhat U
U
Umoi=matrix(nrow=c,ncol=N)
sodem2=0
for ( j in 1:N){
m=0
for(k in 1:c){
if(d2[k,j]==0){
m=m+1
}
if (m==0){
for ( i in 1:c){
tong=0
for (k in 1:c){
tong=tong+d2[i,j]/d2[k,j]
}
Umoi[i,j]=1/tong
```

```
}
}
else
for (l in 1:c){
if (d2[l,j]==0){
Umoi[l,j]=1/m
}else{
Umoi[l,j]=0
}
}
}
}
Umoi
# Tim chuan
chuan=0
lech=matrix(0,ncol=N,nrow=c)
for (i in 1:c){
for(j in 1:N){
lech[i,j]=abs(Umoi[i,j]-U[i,j])
if (lech[i,j]>chuan){
chuan=lech[i,j]
}
}
}
chuan
# lap lai thuat toan
```

```
vl=0
while (chuan>=epxilanh){
U=Umoi
for (k in 1:n){
for (i in 1:c){
tu=0
mau=0
for(j in 1:N){
tu=tu+((U[i,j])^2)*z2[k,j]
mau=mau+(U[i,j])^2
}
v[k,i]=tu/mau
}
}
v
for (j in 1:N){
for (i in 1:c){
d2[i,j]=0
for (k in 1:n){
d2[i,j]=d2[i,j]+(v[k,i]-z2[k,j])^2
}
}
}
d2
for (j in 1:N){
m=0
```

```
for (k in 1:c){
if (d2[k,j]==0){
m=m+1
}
}
if (m==0){
for (i in 1:c){
tong=0
for (k in 1:c){
tong=tong+d2[i,j]/d2[k,j]
}
Umoi[i,j]=1/tong
}
if (m!=0){
for (l in 1:c){
if (d2[l,j]==0){
Umoi[l,j]=1/m
}else{
Umoi[l,j]=0
}
}
}
}
}
Umoi
chuan=0
```

```
for (i in 1:c){
for (j in 1:N){
lech[i,j]=abs(Umoi[i,j]-U[i,j])
if (lech[i,j]>chuan){
chuan=lech[i,j] }
}
if (chuan<epxilanh){
chuan
}
Umoi
}
vl=vl+1
}
Umoi
print(vl)
chuan
Umoi=round(Umoi,digits = 4)
print(Umoi)
# Nhan hai ma tran
mid=as.vector(mid)
mid=as.numeric(mid)
mid=matrix(mid,nrow=1)
f<-round(mid %*% Umoi,digits=4)
f=matrix(f,nrow=N)
# Du doan
fx=matrix(nrow=N,ncol=1)
```

```
for (i in 1:N){
fx[i,1]=chuanhoa[i,1]+f[i,1]
}
dudoan=(fx*M)/100
dudoan
actual=st[2:length(st)]
actual=as.vector(actual)
actual=as.numeric(actual)
actual=matrix(actual,nrow=N)
X=data.frame(dudoan,actual)
# Tim cac sai so
MAE=sum(abs(X[,1]-X[,2]))/dim(X)[1]
MSE=sum((abs(X[,1]-X[,2]))^2)/dim(X)[1]
MAPE=(100/dim(X)[1])*sum(abs((X[,1]-X[,2])/X[,1]))
Error=matrix(c("MAE","MSE","MAPE",round(MAE,3),
round(MSE,round(MAPE,3)),nrow = 2,ncol = 3,byrow = T)
print(Error)
# Du bao moi
m=mean(f)
print(m)
dubaomoi=((chuanhoa[length(st),1]+m)*M)/100
print(dubaomoi)
# Ghep cot
fz=c(0,f)
fx=c(0,dudoan)
as.matrix(fz)
```

```
as.matrix(fx)
Bang2=data.frame(st,fz,fx)
}
chum(zt)
```

## C. MÔ HÌNH CẢI TIẾN CỦA Singh

```
data=read.csv(file.choose(),header=T)
st=data[,2]
zt=diff(st)
ts=as.matrix(zt)
singpublic=function(ts,n,D1,D2){
min.x=min(ts)-D1
max.x=max(ts)-D2
h=((max.x-min.x)/n)
k<-1:(n+1)
U<-1:n
for (i in 1:(n+1)){
if (i==1)k[i]=min.x
else
{k[i]=min.x+(i-1)*h
U[i-1]=paste("U",i-1,sep="")}
}
D<-data.frame(U,low=k[1:n],up=k[2:(n+1)])
D$Bw<-(1/2)*(D$low+D$up)
D=data.frame(D)
# xac dinh tap mo cho ts
loai <- 1:length(ts)
```

```
for (i in 1:length(ts)) {
for (j in 1:n){
if (D$low[j] <= ts[i] & ts[i] <= D$up[j]) {
loai[i] = paste("A", j, sep = "")
break
}
}
}
# Tinh 12 gia tri
Ds <- 0
E <- ts
v=matrix(0,ncol = 12,nrow = length(ts))
for(i in 3:length(ts)){Ds[i] = abs(abs(E[i] - E[i - 1]) - abs(E[i - 1] - E[i - 2]))
a=c(-1/6,-1/4,-1/2,-3,-2,-1,1/2,1/4,1/6,1,2,3)
v[i,]=E[i]+a*(Ds[i])}
print(v)
# xac dinh quan he mo
ts=data.frame(ts,loai)
qhm=0
for(i in 1:dim(ts)[1]-1){qhm[i]=paste(loai[i],loai[i+1],sep="->")}
print(qhm)
# noi suy
count=0
f<-matrix(0,ncol = n*2,nrow =dim(ts)[1]+1)
k<-matrix(0,ncol = n,nrow = n)
for(i in 1:length(qhm)){k[as.numeric(substr(qhm[i],2,2)),
```

```
as.numeric(substr(qhm[i],6,6))]=k[as.numeric(substr(qhm[i],2,2)),
as.numeric(substr(qhm[i],6,6))]+1}
print(k)
for(i in 3:dim(ts)[1]){for(l in 1:n){
R<-0
S<-0
if (k[as.numeric(substr(ts$loai[i],2,2)),l]!=0){
if (D$low[l] <= v[i,1] & v[i,1] <= D$up[l]) {
R = R + v[i,1]
S = S + 1
}
if (D$low[l] <= v[i,2] & v[i,2] <= D$up[l]) {
R = R + v[i,2]
S = S + 1
}
if (D$low[l] <= v[i,3] & v[i,3] <= D$up[l]) {
R = R + v[i,3]
S = S + 1
}
if (D$low[l] <= v[i,4] & v[i,4] <= D$up[l]){
R = R + v[i,4]
S = S + 1
}
if (D$low[l] <= v[i,5] & v[i,5] <= D$up[l]) {
R = R + v[i,5]
S = S + 1
```

```
}
if (D$low[l] <= v[i,6] & v[i,6] <= D$up[l]) {
R = R + v[i,6] S = S + 1
}
if (D$low[l] <= v[i,7] & v[i,7] <= D$up[l]) {
R = R + v[i,7]
S = S + 1
}
if (D$low[l] <= v[i,8] & v[i,8] <= D$up[l]) {
R = R + v[i,8]
S = S + 1
}
if (D$low[l] <= v[i,9] & v[i,9] <= D$up[l]) {
R = R + v[i,9]
S = S + 1
}
if (D$low[l] <= v[i,10] & v[i,10] <= D$up[l]) {
R = R + v[i,10]
S = S + 1
}
if (D$low[l] <= v[i,11] & v[i,11] <= D$up[l]) {
R = R + v[i,11]
S = S + 1
}
if (D$low[l] <= v[i,12] & v[i,12] <= D$up[l]) {
R = R + v[i,12]
```

```
S = S + 1
}
}
f[i+1,l]=(0+R + D$Bw[l])/(0+S + 1)
f[i+1,l+n]=k[as.numeric(substr(ts$loai[i],2,2)),l]
}}
fc <- vector(mode="numeric", length=0)
for(i in 1:dim(ts)[1]+1)
{fc[i]<-sum(f[i,1:n]*f[i,(n+1):(n*2)])/sum(f[i,(n+1):(2*n)])}
vl=data.frame(f,fc)
print(vl)
tsluu=ts
dubao <- vector(mode="numeric", length=0)
dubaovariation=c(NA,fc)
for(i in 4:length(dubaovariation)){if(i<(dim(ts)[1]+1))
dubao[i]=dubaovariation[i]+st[i-1]
else dubao[i]=dubao[i-1]+dubaovariation[i]}
print(dubao)
dubao=dubao[1:(length(dubao)-1)]
Y=data.frame(st,dubao)
X=data.frame(Y[5:dim(Y)[1],1],dubao[5:dim(Y)[1]])
print(X)
MAE=sum(abs(X[,1]-X[,2]))/dim(X)[1]
MSE=sum((abs(X[,1]-X[,2]))2̂)/dim(X)[1]
MAPE=(100/dim(X)[1])*sum(abs((X[,1]-X[,2])/X[,1]))
Error=matrix(c("MAE","MSE","MAPE",round(MAE,3),round(MSE,3),
```

round(MAPE,3)),nrow = 2,ncol = 3,byrow = T)

print(Error)

} singpublic(ts,7,0,0)