Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

# DỰ BÁO DÂN SỐ VIỆT NAM BẰNG CÁC MÔ HÌNH THỐNG KÊ

Võ Văn Tài[1] và Phạm Minh Trực[2]

[1] *Khoa Khoa học Tự nhiên, Trường Đại học Cần Thơ*
[2] *Học viên Cao học, Khoa Khoa học Tự nhiên, Trường Đại học Cần Thơ*

***Thông tin chung:***
*Ngày nhận: 19/03/2014*
*Ngày chấp nhận: 28/08/2014*

***Title:***
*Forecasting Vietnam's population by statistical models*

***Từ khóa:***
*Hồi quy, chuỗi thời gian, chuỗi thời gian mờ, dự báo, tiêu chuẩn AIC*

***Keywords:***
*Regression, time series, fuzzy time series, forecast, AIC criterion*

**ABSTRACT**

*This study uses different models of regression, time series and fuzzy time series to forecast Vietnam's population from historical data. By using statistical criterions, the most appropriate model can be found for forecasting Vietnam's population to 2020.*

**TÓM TẮT**

*Nghiên cứu này sử dụng các mô hình khác nhau của hồi quy, chuỗi thời gian và chuỗi thời gian mờ để dự báo dân số nước ta dựa trên các số liệu của quá khứ. Sử dụng các tiêu chuẩn thống kê để tìm mô hình thích hợp nhất cho mỗi trường hợp, từ đó tiến hành dự báo dân số nước ta đến năm 2020.*

## 1 GIỚI THIỆU

Dân số là một vấn đề lớn mà mỗi chính phủ đều phải có sự quan tâm đặc biệt bởi vì nó ảnh hưởng trực tiếp đến sự phát triển kinh tế xã hội của quốc gia mình. Dự báo dân số là một công việc phải thực hiện đầu tiên, không thể thiếu được trước khi hoạch định các chính sách vĩ mô ngắn hạn cũng như dài hạn của một địa phương, một quốc gia. Các chính sách cho tất cả các lĩnh vực cần phải dựa trên thông tin về dân số. Dự báo dân số tốt, không những tận dụng được nguồn nhân lực hợp lý nhất trong phát triển kinh tế xã hội mà còn tiết kiệm cũng như chủ động trong xây dựng cơ sở vật chất, đội ngũ cán bộ,… của tất cả các lĩnh vực.

Trong thống kê, hai mô hình chính đang được sử dụng rộng rãi trong dự báo là mô hình hồi quy và mô hình chuỗi thời gian. Trong hai mô hình này, chuỗi thời gian được xem có nhiều ưu điểm hơn. Chuỗi thời gian đang được sử dụng phổ biến và hiệu quả trong nghiên cứu khoa học bởi vì rất nhiều số liệu cần dự báo được thu thập theo thời gian. Các mô hình chuỗi thời gian như tự hồi qui (AR), trung bình trượt (MA), tự hồi qui trung bình trượt (ARMA), tự hồi qui tích hợp trung bình trượt (ARIMA),… đã được áp dụng rất phổ biến trong các dự báo của kinh tế xã hội,… Tuy nhiên, dự báo bằng mô hình chuỗi thời gian sẽ không có hiệu quả nếu chuỗi dữ liệu không dừng và không tuyến tính. Với sự kết hợp của lý thuyết tập mờ, những số liệu thu được của quá khứ có sự liên kết xác suất theo một quy tắc nhất định. Chuỗi thời gian mờ tận dụng sự liên kết số liệu này đã được chứng minh có nhiều ưu việt hơn trong dự báo so với chuỗi thời gian không mờ. Nhiều mô hình chuỗi thời gian mờ đã được đề nghị như mô hình của S.M.Chen (1996), K.Huarng (2001), A.M. Abasov *et al.* (2002), S.R.Singh (2009),… Theo tìm hiểu của chúng tôi, chuỗi thời gian mờ chưa được quan tâm đúng mức ở nước ta nên những dự báo cụ thể trong các lĩnh vực chưa được xem xét nhiều.

Hiện nay, ngành thống kê trên thế giới, đặc biệt là lĩnh vực dự báo đã có sự phát triển vượt bậc. Trong lĩnh vực dự báo dân số, có những mô hình, công cụ tính toán và sự đánh giá mới trong những năm gần đây. Với số liệu đã có, mô hình thống kê đã được nghiên cứu, cùng với các phần mềm thống kê hiện tại chúng ta hoàn toàn có thể xây dựng được các mô hình để dự báo tốt cho dân số nước ta. Kết quả dự báo sẽ là thông tin quan trọng để hoạch định các chính sách vĩ mô trong phát triển kinh tế xã hội của đất nước. Bài viết này khảo sát các mô hình hồi quy, chuỗi thời gian mờ và không mờ để tìm các mô hình thích hợp nhất trong dự báo dân số nước ta. Cách làm trong bài viết này có thể được áp dụng để dự báo dân số cho các tỉnh, huyện và nhiều lĩnh vực khác ở nước ta.

## 2 CÁC MÔ HÌNH DỰ BÁO

### 2.1 Mô hình hồi quy

Gọi $t$ là năm ứng với dân số dự báo $y_t$, các mô hình hồi quy được sử dụng trong nghiên cứu là

*Tuyến tính đơn*: $y_t = a + bt$ (1)

*Lũy* thừa: $y_t = e^{b\ln(t)+a}$ (2)

*Mũ biến* dạng: $y_t = a + bc^{\hat{t}}$ (3)

*Cấp số cộng*: $y_t = y_{2011}(1 + r_1\Delta t)$ (4)

*Cấp số* nhân: $y_t = y_{2011}(1 + r_2)^{\Delta t}$ (5)

Trong mô hình (1), (2) và (3), $a$ và $b$ là các hệ số của mô hình, $\hat{t} = \frac{1}{5}(t - 2011)$ với $t$ là thời gian cần dự báo; trong mô hình (4) và (5), $\Delta t$ là khoảng thời gian từ năm dự báo đến năm được chọn làm gốc; $r_2, r_2$ là tốc độ tăng dân số hằng năm được tính bởi

$$r_1 = \frac{\ln(y_2) - \ln(y_1)}{t_2 - t_1}, \quad r_2 = t_2 - t_1\sqrt{\frac{y_2}{y_1}}.$$

với $t_1$, $t_2$ là điểm thời gian đầu và cuối trong dãy số liệu được sử dụng để tính tốc độ gia tăng dân số tương ứng với số dân $y_1$ và $y_2$.

### 2.2 Mô hình chuỗi thời gian

*Mô hình tự hồi qui bậc p* (*AR(p)*):

$$y_t = \varphi_0 + \sum_{i=1}^{p} \varphi_t y_{t-i} + u_t \quad (6)$$

trong đó $\phi_i$ là các hệ số ước lượng của mô hình, $u_t$ là số hạng đảm bảo tính ồn trắng.

*Mô hình trung bình di động bậc q* (*MA*(*q*)):

$$y_t = \beta_0 + \sum_{i=1}^{q} \beta_i u_{t-i} \quad (7)$$

trong đó $\beta_i$ cũng là các hệ số ước lượng của mô hình và $u_i$ giống như trong (6).

Mô hình tự hồi qui và trung bình di động (ARMA(p,q)):

$$y_t = \varphi_0 + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + ... + \varphi_p y_{t-p} + u_t + \beta_1 u_{t-1} + \beta_2 u_{t-2} + ... + \beta_q u_{1-q} \quad (8)$$

Một quá trình *ARMA*(*p,q*) sẽ có quá trình tự hồi quy bậc $p$ và quá trình trung bình di động bậc $q$.

Mô hình trung bình di động tổng hợp với tự hồi qui ARIMA(p,d,q):

$$y_t = \delta + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + ... + \alpha_p y_{t-p} + \beta_1 + \varepsilon_{t-1} + ... + \beta_q \varepsilon_{t-q} + e_t \quad (9)$$

trong đó $\alpha_i, i = 1,, 2, ..., p$ là tham số tự hồi quy; $\varepsilon_{t-j}, j = 1, 2, ..., q$ là tham số trung bình di động; $\delta = \mu(\beta_1 + \beta_2 + ... + \beta_q)$; $\mu$ là giá trị trung bình của chuỗi thời gian; $e_t$ là sai số dự báo ($e_t = \hat{y}_t - y_t$ = số liệu dự báo - số liệu thực tế).

### 2.3 Mô hình hình chuỗi thời gian mờ

Hiện tại có nhiều mô hình chuỗi thời gian mờ khác nhau được đề nghị. Trong ứng dụng, người ta thường sử dụng các mô hình của Chen (1996), Singh (2008), Huarng (2001), Abbasov – Mamedova (2003), và của Chen-Hsu (2004). Ngoại trừ mô hình của Abbasov –Mamedova, các mô hình còn lại đều được đề nghị gồm 4 bước, trong đó có 3 bước đầu giống nhau chỉ khác nhau ở bước cuối cùng: *mờ hóa dữ liệu*. Ba bước chung của các mô hình được đề nghị như sau:

*Bước* 1: Xác định tập nền $U$ trên các giá trị lịch sử của chuỗi thời gian: $U = [D_{\min} - D_1; D_{\max} + D_2]$, trong đó $D_{\min}$, $D_{max}$ lần lượt là giá trị lớn nhất và nhỏ nhất của chuỗi dữ liệu, $D_1$, $D_2$ là các số dương thích hợp được chọn.

*Bước* 2: Chia tập $U$ thành từng đoạn thích hợp và đều nhau $U_1, U_2, ..., U_n$. Xác định các tập mờ $A_i$ tương ứng với $U_i$. Nếu $A_i$ là giá trị mờ hóa tại thời điểm $t$ và $A_j$ là giá trị mờ hóa tại thời điểm $t$ +1 thì ta có mối quan hệ mờ $A_i \to A_j$ $(i, j = 1, 2, ....)$.

*Bước* 3: Xác định các nhóm quan hệ mờ.

Bước cuối cùng của từng mô hình được đề nghị cụ thể sau:

*2.3.1 Mô hình của Chen*

*Nguyên tắc* 1: Nếu $A_i$ là giá trị mờ hóa tại thời điểm $t$ và chỉ có mối quan hệ mờ duy nhất là $A_i \to A_j$ thì giá trị dự báo tại thời điểm $t$ +1 là $m_j$ ($m_j$ là điểm giữa của đoạn $U_j$).

*Nguyên tắc* 2: Nếu $A_i$ là giá trị mờ hóa tại thời điểm $t$ và có nhóm mối quan hệ mờ là $A_i \to A_j, A_k, A_l, ...$ thì giá trị dự báo tại thời điểm $t$ +1 là trung bình cộng của $m_j, m_k, m_l, ...$ ($m_j, m_k, m_l, ...$ là điểm giữa của đoạn $U_j, U_k, U_l, ...$).

*Nguyên tắc 3*: Nếu $A_i$ là giá trị mờ hóa tại thời điểm $t$ và không tồn tại mối quan hệ mờ nào thì giá trị dự báo tại thời điểm $t$ +1 là $m_i$ ($m_i$ là điểm giữa của đoạn $U_i$).

*2.3.2 Mô hình của Singh*

Với $k = \overline{3, n}$, mối quan hệ mờ của phần tử $k$ và $k + 1$ là $A_i \to A_j$.

Với $R = 0, S = 0$, tính các giá trị sau:

$$D_i = \left\| E_i - E_{i-1} \right| - \left| E_{i-1} - E_{i-2} \right\|, X_i = E_i + \frac{D_i}{2}, XX_i = E_i - \frac{D_i}{2}, Y_i = E_i + D_i, YY_i = E_i - D_i, P_i = E_i + \frac{D_i}{4}, PP_i = E_i - \frac{D_i}{4}$$

$$Q_i = E_i + 2 \times D_i,\ QQ_i = E_i - 2 \times D_i, G_i = E_i + \frac{D_i}{6}, GG_i = E_i - \frac{D_i}{6}, H_i = E_i + 3 \times D_i, HH_i = E_i - 3 \times D_i; F_j = \frac{R + M(A_j)}{S + 1}$$

Trong đó $E_i, E_{i-1}, E_{i-2}$ lần lượt là giá trị tại thời điểm $t$, $t$ -1, $t$ - 2; $A_i$, $A_j$ lần lượt là giá trị mờ tại thời điểm $t$, $t$ +1 ; $F_j$ là giá trị dự báo tại thời điểm $t$ + 1.

Khi đó ta có các nguyên tắc mờ hóa dữ liệu như sau:

$$X_i \in U_j \Rightarrow R = R + X_i,\ S = S + 1; XX_i \in U_j \Rightarrow R = R + XX_i,\ S = S + 1; Y_i \in U_j \Rightarrow R = R + Y_i, S = S + 1;$$

$$YY_i \in U_j \Rightarrow R = R + YY_i, S = S + 1; P_i \in U_j \Rightarrow R = R + P_i, S = S + 1; PP_i \in U_j \Rightarrow R = R + PP_i, S = S + 1;$$

$$Q_i \in U_j \Rightarrow R = R + Q_i, S = S + 1; QQ_i \in U_j \Rightarrow R = R + QQ_i, S = S + 1; G_i \in U_j \Rightarrow R = R + G_i, S = S + 1;$$

$$GG_i \in U_j \Rightarrow R = R + GG_i, S = S + 1; H_i \in U_j \Rightarrow R = R + H_i, S = S + 1; HH_i \in U_j \Rightarrow R = R + HH_i, S = S + 1$$

*2.3.3 Mô hình Heuristic*

Ta có giá trị mờ $F(t)$ có nhóm quan hệ mờ $A_j \to A_p, A_q, A_r, A_s, ...$ và hàm *Heuristic* $h(x; A_p, A_q, A_r, A_s, ...) = A_{p1}, A_{p2}, ..., A_{pk}$ với $x = X(t) - X(t-1)$. Nếu $x > 0$ thì $p_1, p_2, ..., p_k \geq j$, ngược lại nếu $x < 0$ thì $p_1, p_2, ..., p_k \leq j$. Khi đó, nếu $x > 0$ thì $A_j \to A_{p1}, A_{p2}, ..., A_{pk}$ với

$p_1, p_2, ..., p_k \geq j$, nếu $x < 0$ thì

$A_j \to A_{p1}, A_{p2}, ..., A_{pk}$ với $p_1, p_2, ..., p_k \leq j$.

Nguyên tắc mờ hóa dữ liệu tương tự mô hình của Chen.

*2.3.4 Mô hình của Abbasov -Mamedova*

Mô hình chuỗi thời gian mờ này gồm 6 bước như sau:

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

*Bước* 2: Chia tập $U$ thành $n$ đoạn thời gian có độ dài bằng nhau chứa các giá trị biến đổi tương ứng với tỷ lệ tăng trưởng khác nhau của dân số. Đồng thời tính các giá trị trung bình của từng đoạn $\left(u_m^i, i = 1, ..., n\right)$.

*Bước* 3: Mô tả chất lượng của các giá trị biến đổi dân số như là một biến ngôn ngữ, xác định các giá trị tương ứng của biến ngôn ngữ hoặc thiết lập các tập mờ $F(t)$:

$$A^t = \left[\mu_{A_i}\left(u_i\right)/u_i\right], u_i \in U, \mu_{A_i} \in [0,1], \quad \mu_{A_i}\left(u_i\right) = \frac{1}{1+\left[C \times \left(U - u_m^i\right)\right]^2}$$

Trong đó $A^t$ là mờ hóa các biến của năm *t; C* là hằng số tự chọn sao cho $\mu_{A_i}\left(u_i\right) \in [0,1]$; $U$ là các biến đổi của từng năm, hoặc là giá trị trung bình; $u_m^i$ là giá trị trung bình của từng đoạn thứ *i*.

*Bước* 4*:* Mờ hóa các dữ liệu đầu vào hoặc chuyển đổi các giá trị số vào các giá trị mờ. Hoạt động này cho phép phản ánh sự tương ứng giá trị định lượng hay định tính của tỷ lệ phát triển dân số tiêu biểu trong giá trị của hàm quan hệ.

*Bước* 5*:* Lựa chọn tham số $w(1 < w < n)$; $n$ 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 $P^w(T)$.

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

Hay

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

$$F(t) = \left[max\left(R_{11}, R_{21}, ..., R_{i1}\right) max\left(R_{12}, R_{22}, ..., R_{i2}\right) ... max\left(R_{1j}, R_{2j}, ..., R_{ij}\right)\right]$$

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

*Bước 6:* Giải mờ kết quả thu được hoặc chuyển đổi các giá trị mờ vào các giá trị định tính. Dự báo cho năm tới $V(t)$:

$$V(t) = \frac{\sum_{i=1}^{w} \mu_t\left(u_i\right) \times u_m^i}{\sum_{i=1}^{w} \mu_t\left(u_i\right)}$$

Kết quả dự báo cho năm thứ *t* được tính theo công thức $N(t) = N(t-1) + V(t)$. trong đó $N(t)$ là dân số của năm *t*, $V(t)$ là số dân thay đổi từ năm *t* - 1 đến năm *t*.

## 3 TỔNG QUAN VIỆC THỰC HIỆN

### 3.1 Nguồn số liệu

Bài viết sử dụng số liệu của quá khứ từ trang web của Tổng cục thống kê 12/2012. Cụ thể số liệu được cho bởi các bảng sau:

**Bảng 1: Dân số cả nước giai đoạn 1975 – 2011**

| Năm | Số dân | Năm | Số dân | Năm | Số dân |
|---|---|---|---|---|---|
| 1975 | 47.6 | 1987 | 62.5 | 1999 | 76.6 |
| 1976 | 49.2 | 1988 | 63.7 | 2000 | 77.9 |
| 1977 | 50.4 | 1989 | 64.8 | 2001 | 78.9 |
| 1978 | 51.4 | 1990 | 66.2 | 2002 | 79.7 |
| 1979 | 52.5 | 1991 | 67.8 | 2003 | 80.9 |
| 1980 | 53.8 | 1992 | 69.4 | 2004 | 82.0 |
| 1981 | 54.9 | 1993 | 71.0 | 2005 | 83.1 |
| 1982 | 56.2 | 1994 | 72.5 | 2006 | 84.1 |
| 1983 | 57.4 | 1995 | 74.0 | 2007 | 84.221 |
| 1984 | 58.8 | 1996 | 73.2 | 2008 | 85.122 |
| 1985 | 59.9 | 1997 | 74.3 | 2009 | 86.024 |
| 1986 | 61.1 | 1998 | 75.5 | 2010 | 86.928 |
| | | | | 2011 | 87.840 |

*(Số liệu Bảng 1, cũng như trong bài báo này được tính đơn vị là triệu người.)*

### 3.2 Phương pháp thực hiện

Sử dụng các mô hình hồi quy, chuỗi thời gian trên dữ liệu gốc và dữ liệu mờ hóa để dự báo tổng số dân. Cụ thể:

i) Sử dụng số liệu Bảng 1, các mô hình hồi quy trong phần 2.1, xây dựng các mô hình hồi quy cụ thể. Dùng các tiêu chuẩn đánh giá khác nhau để lựa chọn mô hình hồi quy phù hợp nhất.

ii) Sử dụng dữ liệu gốc, phương pháp Box-Jenkins xác định các mô hình chuỗi thời gian không mờ *AR(p)*, *MA*(*q*), *ARIMA*(*p,d,q*) có thể có. Dựa vào tiêu chuẩn *AIC* để lựa chọn mô hình chuỗi thời gian tốt nhất.

iii) Mờ hóa dữ liệu gốc bằng các mô hình của Chen, Singh, Huarng, Chen-Hsu. Sau khi lựa chọn được mô hình có chỉ số *MSE* nhỏ nhất, chúng ta cũng sử dụng phương pháp như đã làm trong ii) để tìm mô hình phù hợp nhất.

iv) Sử dụng mô hình chuỗi thời gian mờ Abbasov-Mamedova cho việc dự báo từ dữ liệu gốc.

v) Lựa chọn mô hình có chỉ số *AIC* nhỏ nhất từ i), ii), iii) và iv) để làm mô hình tối ưu nhất.

Với mô hình đã chọn, tiến hành dự báo dân số Việt Nam đến năm 2020. Việc xử lý được thực hiện bằng phần mềm thống kê R.

## 4 KẾT QUẢ DỰ BÁO TỔNG DÂN SỐ CỦA CẢ NƯỚC

### 4.1 Sử dụng các mô hình hồi quy

#### *4.1.1 Đường hồi quy tìm được*

Từ số liệu Bảng 1, các mô hình (1), (2), (3), (4) và (5) được thiết lập cụ thể như sau:

*Hồi quy tuyến tính đơn*:

$$y_t = 1.141t - 2204.674.$$

*Hồi quy lũy thừa*:

$$y_t = \exp(33.84 \ln t - 252.85).$$

*Cấp số cộng*:

$$y_t = 87.84[1 + 0.01092(t - 2011)].$$

*Cấp số nhân*:

$$y_t = 87.84(1 + 0.010977)^{t-2011}.$$

*Hàm mũ biến dạng*:

$$y_t = 97.53824 - 19.91821.(0.9397109)^{\frac{1}{5}(t-2000)}.$$

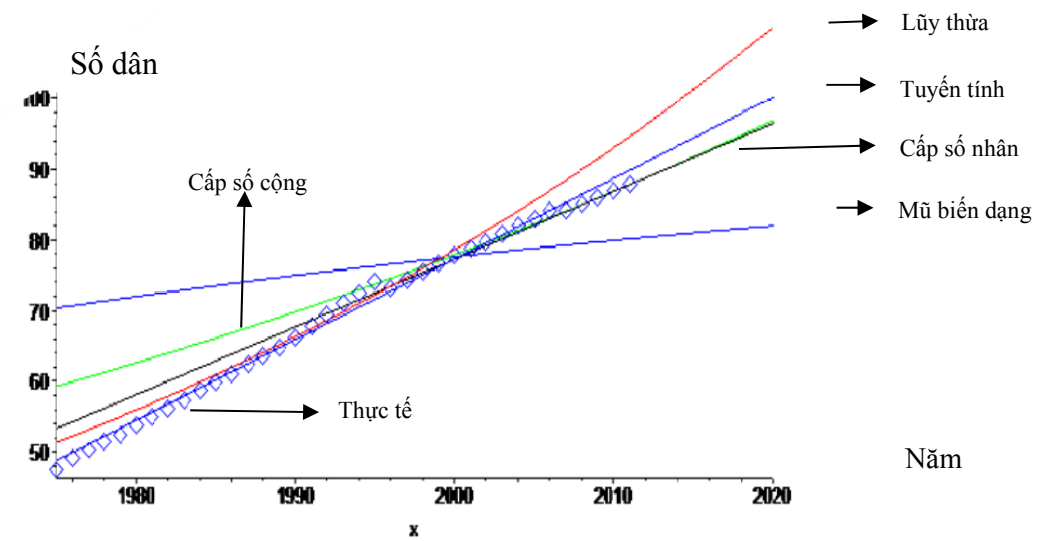
#### *4.1.2 Lựa chọn đường hồi quy*

Từ các mô hình đã được xây dựng trong mục *4.1.1*, ta có bảng tóm tắt kết quả tính các tiêu chuẩn đánh như sau:

**Bảng 2: Các tiêu chuẩn đánh giá mô hình hồi quy đã xây dựng**

| Hàm dự báo | $R^2$ | *AIC* | *SIC* | *MSE* |
|---|---|---|---|---|
| Tuyến tính đơn | 0.994 | 134.97 | 138.19 | 0.725 |
| Lũy thừa | 0.970 | 191.78 | 195.00 | 1.747 |
| Cấp số cộng | 0.950 | 25.29 | 26.26 | 0.576 |
| Cấp số nhân | 0.980 | 18.41 | 19.38 | 0.361 |
| Mũ biến dạng | 0.886 | 73.49 | 74.95 | 4.262 |

Chúng ta cũng có đồ thị cho các đường hồi quy xây dựng và số liệu thực tế như sau:




**Hình 1: Đồ thị các mô hình dự báo giai đoạn 1975-2011 và số liệu thực tế**

Nhận xét:

i) Hệ số xác định của các mô hình hồi quy xây dựng tăng dần theo thứ tự: Mũ biến dạng → cấp số cộng → cấp số nhân → lũy thừa → tuyến tính đơn. Trong đó, ngoại trừ mô hình mũ biến dạng có hệ số xác định thấp, còn lại các mô hình khác có hệ số xác định cao và không có sự sai lệch nhiều, chứng tỏ các mô hình hồi quy xây dựng có mức phù hợp khá tốt.

ii) Chỉ số AIC và SIC của các mô hình xây dựng tăng dần theo thứ tự: Cấp số nhân → cấp số cộng → mũ biến dạng → tuyến tính đơn → lũy thừa. Trong đó, mô hình cấp số cộng có chỉ số AIC và SIC nhỏ nhất nên đây là mô hình phù hợp hơn những mô hình còn lại.

iii) Sai số tuyệt đối trung bình MSE của các mô hình xây dựng tăng dần theo thứ tự: Cấp số nhân → cấp số cộng → tuyến tính đơn → lũy thừa → mũ biến dạng. Như vậy, chỉ số này cũng cho ta thấy mô hình cấp số nhân là phù hợp nhất.

iv) Đồ thị phân tán cho dữ liệu thực tế, các đường hồi quy đã thiết lập từ Hình 1 cho ta thấy mô hình cấp số nhân khá gần với giá trị thực tế.

Từ các nhận xét trên, ta thấy rằng trong các mô hình hồi quy xây dựng, mô hình cấp số nhân là phù hợp nhất.

### 4.2 Phương pháp chuỗi thời gian với dữ liệu không mờ

*4.2.1 Các mô hình dự báo theo dãy số thời gian*

Từ số liệu Bảng 1, kiểm tra tính dừng, đồ thị tự tương quan (ACF) và tự tương quan riêng (PACF), ta có các mô hình dự báo có thể như sau:

Mô hình MA: Kết quả phân tích cho ta thấy có một MA(1).

Mô hình AR: Sự phân tích cho ta thấy không tồn tại.

Mô hình ARIMA: Các mô hình có thể có là ARIMA(0,2,1);ARIMA(0,2,2);

ARIMA(0,2,3);ARIMA(1,2,0);ARIMA(2,2,0); ARIMA(3,2,0);ARIMA(1,2,1);ARIMA(1,2,2);ARIMA(1,2,3);ARIMA(2,2,1);ARIMA(3,2,1);ARIMA (2,2,2);ARIMA(2,2,3);ARIMA(3,2,2); ARIMA(3,2,3).

*4.2.2 Lựa chọn mô hình*

Dùng chỉ số AIC để tìm mô hình thích hợp nhất từ các mô hình có thể trên, ta có bảng tổng hợp sau:

**Bảng 3: Chỉ số AIC cho các mô hình chuỗi thời gian**

| Mô hình | *AIC* |
|---|---|
| ARIMA(0,2,1) | 46.83 |
| ARIMA(0,2,2) | 48.71 |
| ARIMA(0,2,3) | 50.67 |
| ARIMA(1,2,0) | 56.42 |
| ARIMA(2,2,0) | 55.11 |
| ARIMA(3,2,0) | 56.22 |
| ARIMA(1,2,1) | 48.72 |
| ARIMA(1,2,2) | 50.00 |
| ARIMA(1,2,3) | 52.67 |
| ARIMA(2,2,1) | 50.67 |
| ARIMA(3,2,1) | 52.63 |
| ARIMA(2,2,2) | 51.98 |
| ARIMA(2,2,3) | 53.89 |
| ARIMA(3,2,2) | 53.98 |
| ARIMA(3,2,3) | 55.11 |

So sánh các mô hình Bảng 3, ta thấy mô hình ARIMA (0,2,1) (hay MA(1)) có chỉ số *AIC* nhỏ nhất. Đồ thị Standardized Residuals có sai số chuẩn tập trung gần giá trị 0, đồ thị ACF of Residuals cho thấy tính phù hợp của mô hình. Như vậy, mô hình ARIMA (0,2,1) phù hợp dự báo là

$$X_t^{**} = \varepsilon_t - 0.8750\varepsilon_{t-1}; X_t^{**} = X_t - 2X_{t-1} + X_{t-2}$$

### 4.3 Phương pháp chuỗi thời gian với dữ liệu mờ hóa

*4.3.1 Mờ hóa dữ liệu*

Từ các nguyên tắc mờ hóa mô hình Chen, Singh, Huarng và Chen-Hsu đã trình bày trong phần 2.3, tính toán cho dữ liệu của Bảng 1 ta có kết quả sau:

**Bảng 4: Kết quả mờ hóa dữ liệu mô hình của Chen, Singh, Huarng và Chen-Hsu giai đoạn 1989-2011**

| Năm | Thực tế | Chen | Singh | Huarng | Chen-Hsu |
|---|---|---|---|---|---|
| 1989 | 64.8 | - | - | - | - |
| 1990 | 66.2 | 67 | - | 67 | - |
| 1991 | 67.8 | 67 | - | 67 | 67.75 |
| 1992 | 69.4 | 70 | 67.86 | 70 | 69.25 |
| 1993 | 71.0 | 70 | 71.50 | 70 | 71.50 |
| 1994 | 72.5 | 73 | 71.04 | 73 | 71.50 |
| 1995 | 74.0 | 73 | 74.50 | 73 | 74.50 |
| 1996 | 73.2 | 76 | 74.04 | 74.5 | 73.50 |
| 1997 | 74.3 | 76 | 73.85 | 76 | 74.25 |
| 1998 | 75.5 | 76 | 74.32 | 76 | 75.25 |
| 1999 | 76.6 | 76 | 77.50 | 76 | 76.75 |
| 2000 | 77.9 | 79 | 76.67 | 79 | 78.25 |
| 2001 | 78.9 | 79 | 77.87 | 79 | 78.25 |
| 2002 | 79.7 | 79 | 79.61 | 79 | 80.50 |
| 2003 | 80.9 | 82 | 79.76 | 82 | 80.50 |
| 2004 | 82.0 | 82 | 82.80 | 82 | 82.75 |
| 2005 | 83.1 | 85 | 82.31 | 85 | 82.75 |
| 2006 | 84.1 | 85 | 83.13 | 85 | 84.25 |
| 2007 | 85.171 | 85 | 86.50 | 85 | 84.25 |
| 2008 | 85.122 | 85 | 85.30 | 85 | 85.38 |
| 2009 | 86.024 | 86.5 | 86.02 | 86.5 | 86.13 |
| 2010 | 86.928 | 86.5 | 86.24 | 86.5 | 86.69 |
| 2011 | 87.840 | 86.5 | 86.90 | 86.5 | 87.63 |
| | MSE | 1.214 | 0.868 | 0.934 | 0.19 |

*4.3.2 Lựa chọn mô hình từ số liệu mờ hóa*

Trong các mô hình ở trên, mô hình Chen-Hsu có chỉ số MSE nhỏ nhất. Lấy dữ liệu mờ hóa theo mô hình này, thực hiện việc dự báo bằng mô hình chuỗi thời gian như dữ liệu không mờ của 4.2, ta có bảng tóm tắt chỉ số AIC như sau:

**Bảng 5: Các mô hình ARIMACH với dữ liệu mờ hóa theo Chen-Hsu**

| Mô hình | *AIC* |
|---|---|
| $ARIMA_{CH}(0,2,1)$ | 60.94 |
| $ARIMA_{CH}(0,2,2)$ | 55.90 |
| $ARIMA_{CH}(0,2,3)$ | 57.02 |
| $ARIMA_{CH}(1,2,0)$ | 56.91 |
| $ARIMA_{CH}(2,2,0)$ | 55.73 |
| $ARIMA_{CH}(3,2,0)$ | 57.54 |
| $ARIMA_{CH}(1,2,1)$ | 53.40 |
| $ARIMA_{CH}(1,2,2)$ | 55.35 |
| $ARIMA_{CH}(1,2,3)$ | 57.20 |
| $ARIMA_{CH}(2,2,1)$ | 55.34 |
| $ARIMA_{CH}(3,2,1)$ | 57.66 |
| $ARIMA_{CH}(2,2,2)$ | 57.33 |
| $ARIMA_{CH}(2,2,3)$ | 54.89 |
| $ARIMA_{CH}(3,2,2)$ | 56.88 |
| $ARIMA_{CH}(3,2,3)$ | 60.50 |

So sánh các mô hình trên ta thấy mô hình $ARIMA_{CH}(1,2,1)$ có chỉ số *AIC* nhỏ nhất. Vậy mô hình thích hợp để dự báo là $ARIMA_{CH}(1,2,1)$:

$$X_t^{**} = -0.6388X_{t-1}^{**} + \varepsilon_t - 0.8466\varepsilon_{t-1};$$

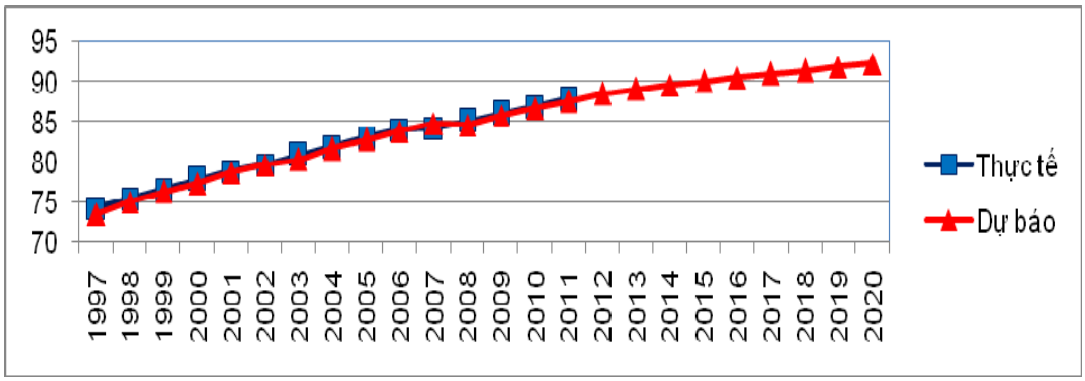
$$X_t^{**} = X_t - 2X_{t-1} + X_{t-2}$$

Phương pháp chuỗi thời gian mờ Abbasov-Mamedova

Sử dụng các bước thực hiện của mô hình chuỗi thời gian mờ Abbasov-Mamedova với dữ liệu Bảng 1, ta có bảng tính toán sau cùng như sau:

**Bảng 6: Kết quả dự báo dân số cả nước giai đoạn 1997 – 2011**

| Năm | Thực tế | | Dự báo | |
|---|---|---|---|---|
| | Số dân | Biến đổi | Số dân | Biến đổi |
| 1997 | 74.300 | 1.100 | 73.552 | 0.352 |
| 1998 | 75.500 | 1.200 | 75.091 | 0.791 |
| 1999 | 76.600 | 1.100 | 76.328 | 0.828 |
| 2000 | 77.900 | 1.300 | 77.387 | 0.787 |
| 2001 | 78.900 | 1.000 | 78.763 | 0.863 |
| 2002 | 79.700 | 0.800 | 79.649 | 0.749 |
| 2003 | 80.900 | 1.200 | 80.351 | 0.615 |
| 2004 | 82.000 | 1.100 | 81.728 | 0.828 |
| 2005 | 83.100 | 1.100 | 82.787 | 0.787 |
| 2006 | 84.100 | 1.000 | 83.887 | 0.787 |
| 2007 | 84.221 | 0.121 | 84.840 | 0.740 |
| 2008 | 85.122 | 0.901 | 84.657 | 0.436 |
| 2009 | 86.024 | 0.902 | 85.812 | 0.690 |
| 2010 | 86.928 | 0.904 | 86.714 | 0.690 |
| 2011 | 87.840 | 0.912 | 87.619 | 0.619 |
| MSE | | | | 0.347 |
| AIC | | | | 25.13 |




**Hình 2: Dân số thực tế và dự báo bằng mô hình Abbasov-Mamedova giai đoạn 1997-2020**

Chỉ số AIC và hình vẽ trực quan (Hình 2) cho ta thấy mô hình Abbasov-Mamedova có kết quả dự báo rất tốt dân số Việt Nam.

### 4.4 Dự báo

Từ các mô hình tối ưu đã lựa chọn trong 4.1, 4.2, 4.3 và 4.4, tiến hành dự báo dân số nước ta đến năm 2020, ta có bảng tổng hợp sau:

**Bảng 7: Dân số nước ta giai đoạn 2012-2020 từ các mô hình dự báo**

| Năm | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 |
|---|---|---|---|---|---|---|---|---|---|
| Abbasov-Mamedova | 88.540 | 89.110 | 89.620 | 90.080 | 90.970 | 91.390 | 91.810 | 92.230 | 92.230 |
| $ARIMA_{CH}(1,2,1)$ | 88.447 | 89.343 | 90.191 | 91.070 | 91.930 | 92.802 | 93.666 | 94.534 | 95.400 |
| ARIMA(0,2,1) | 88.782 | 89.723 | 90.665 | 91.607 | 92.548 | 93.490 | 94.431 | 95.373 | 96.315 |
| Cấp số nhân | 88.800 | 89.780 | 90.86 | 91.760 | 93.790 | 94.82 | 95.860 | 96.910 | 96.910 |

Trong các dự báo của Bảng 7, dựa vào chỉ số AIC, ta thấy mô hình Abbasov-Mamedova cho một kết quả dự báo tốt nhất dân số Việt Nam.

## 5 KẾT LUẬN

Bài báo đã khảo sát các mô hình khác nhau của hồi quy, chuỗi thời gian mờ và không mờ trong dự báo dân số nước ta, dựa vào tiêu chuẩn thống kê, kết luận được mô hình hoàn toàn dựa trên sự mờ hóa dữ liệu Abbasov-Mamedova cho một kết quả dự báo rất tốt. Đây là một kết quả dự báo tốt mà thực tế ứng dụng không nhiều bộ số liệu có được.

Mặc dù, sự phát triển dân số của nước ta phụ thuộc vào các chính sách về dân số của nhà nước trong tương lai, phụ thuộc vào sự phát triển kinh tế xã của đất nước, tuy nhiên với đặc điểm đối tượng dự báo không đòi hỏi quá chính xác, theo chúng tôi kết quả dự báo trên có thể được sử dụng trong hoạch định chính sách kinh tế xã hội vĩ mô cho các cấp quản lí.

Các mô hình và cách làm như đã thực hiện cho dự báo dân số cả nước trong bài viết này, có thể được thực hiện tương tự cho dự báo dân số của một huyện, tỉnh hoặc thành phố cũng như cho nhiều ứng dụng khác của thực tế.

## TÀI LIỆU THAM KHẢO

1. A.M. Abbasov et al, 2002. Fuzzy relational model for knowledge processing and decision making. Advances in Mathematics. 1: 1991-223.
2. A.M. Abbasov and M.H. Mamedova, 2003. Application of fuzzy time series to population forecasting, Vienna University of Technology. 12: 545-552.
3. H. Bozdogan, 2000. Akaike's information criterion and recent developments in information complexity. Journal of mathematical psychology. 44: 62-91.
4. K. Huarng, 2001. Huarng models of fuzzy time series for forecasting. Fuzzy Sets and Systems. 123: 369–386.
5. Q. Song and B.S. Chisom, 1993. Forecasting enrollments with fuzzy time series (Part I), Fuzzy Sets and Systems. 54: 1-9.
6. 6. Q. Song and B.S. Chisom, 1994. Forecasting enrollments with fuzzy time series (Part II), Fuzzy Sets and Systems. 62: 1-8.orecasting enrollments wi.
7. S.M.Chen, 1996. Forecasting enrollments based on fuzzy time series. Fuzzy Sets and Systems. 81: 311-319.

8. S.M. Chen and C.C.Hsu, 2004. A New method to forecast enrollments using fuzzy time series. International Journal of Applied Science and Engineering, 12: 234-244.
9. S.R. Singh, 2008. A computational method of forecasting based on fuzzy time series. Mathematics and Computers in Simulation. 79: 539–554.
10. 10.S.R. Singh, 2009. A computational method of forecasting based on high-order fuzzy time series. Expert Systems with Applications. 36:10551–10559.