

**BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC QUY NHƠN**

**BÁO CÁO ĐỀ TÀI NCKH SINH VIÊN
PHÂN TÍCH CHUỖI THỜI GIAN
BẰNG MÔ HÌNH ARIMA VỚI PHẦN MỀM R**

Sinh viên thực hiện

**Cao Thị Ái Loan, Phùng Thị Hồng Diễm
Nguyễn Quốc Dương, Lê Phương Thảo, Đinh Thị Quỳnh Như**

Người hướng dẫn

TS. LÊ THANH BÌNH.

MỞ ĐẦU

Ngày nay, dự báo đóng một vai trò rất quan trọng trong việc hoạch định nhằm đưa ra các chính sách và chiến lược tối ưu nhất cho công việc. Có rất nhiều mô hình dự báo khác nhau như mô hình hồi quy đơn, mô hình hồi quy bội, Mỗi mô hình đều có ưu nhược điểm riêng và tùy thuộc vào đặc điểm của bộ dữ liệu đầu vào. Đối với dữ liệu có xu hướng tuyến tính và có tính dừng, mô hình ARIMA cho khả năng dự báo rất tốt. Nhằm nâng cao hiệu quả của dự báo, chúng tôi kết hợp lý thuyết mô hình ARIMA với phần mềm R để dự báo độc lập trên các tập dữ liệu thực tế. Từ kết quả dự báo, các nhà quản trị có thể tham khảo để đưa ra phương hướng tối ưu cho công việc của mình.

Nội dung báo cáo được trình bày thành 2 chương:

- **Chương 1. Tổng quan về lý thuyết chuỗi thời gian và mô hình ARIMA.**
- **Chương 2. Phân tích chuỗi thời gian bằng mô hình ARIMA với phần mềm R.**

CHƯƠNG 1. TỔNG QUAN VỀ LÝ THUYẾT CHUỖI THỜI GIAN VÀ MÔ HÌNH ARIMA

1.1 Tổng quan lý thuyết chuỗi thời gian

1.1.1 Các kiến thức mở đầu

1.1.2 Chuỗi thời gian có tính dừng

1.1.3 Hồi quy cổ điển trong chuỗi thời gian

1.2 Mô hình ARIMA

1.2.1 Mô hình ARMA

1.2.2 Phương trình sai phân

1.2.3 Hàm ACF và PACF

1.2.4 Phương trình dự báo

1.2.5 Mô hình ARIMA

1.2.6 Mô hình ARIMA theo mùa

1.1.1 Các kiến thức mở đầu

Khái niệm: Chuỗi thời gian

Mô tả đầy đủ của một chuỗi thời gian được quan sát như một tập hợp của n biến ngẫu nhiên tại các thời điểm tùy ý t_1, t_2, \dots, t_n (với n là số nguyên dương bất kỳ), được cho bởi hàm phân phối đồng thời, là xác suất mà mọi biến ngẫu nhiên trong chuỗi đều nhận giá trị nhỏ hơn các hằng số c_1, c_2, \dots, c_n , tức là

$$F(c_1, c_2, \dots, c_n) = P(x_{t_1} \leq c_1, x_{t_2} \leq c_2, \dots, x_{t_n} \leq c_n). \quad (1)$$

1.1.2 Chuỗi thời gian có tính dừng

Chuỗi có tính dừng là một khái niệm rất quan trọng trong phân tích chuỗi thời gian. Nó được chia làm 2 loại:

- Chuỗi có tính dừng ngặt
- Chuỗi có tính dừng yếu

Trong bài báo cáo này, chúng tôi khai thác đặc điểm của chuỗi có tính dừng yếu.

1.1.2 Chuỗi thời gian có tính dừng

Định nghĩa: Chuỗi có tính dừng yếu [1]

Chuỗi thời gian x_t có tính dừng yếu (*tính dừng*) là một quá trình phương sai hữu hạn sao cho

- (i) hàm giá trị trung bình μ_t là hằng số và không phụ thuộc vào thời gian t ,
- (ii) hàm hiệp phương sai $\gamma(s, t)$ chỉ phụ thuộc vào độ sai khác $|s - t|$.

Chuỗi thời gian được gọi là *không có tính dừng* nếu nó không thỏa mãn được một trong các điều kiện trên.

1.1.2 Chuỗi thời gian có tính dừng

Tính chất: Phân phối mẫu lớn của ACF [1]

Nếu x_t là nhiễu trắng thì ACF mẫu $\hat{\rho}_x(h)$ sẽ có phân phối xấp xỉ phân phối chuẩn với giá trị trung bình bằng 0 khi n đủ lớn và độ lệch chuẩn được cho bởi

$$\sigma_{\hat{\rho}_x(h)} = \frac{1}{\sqrt{n}}, \quad (2)$$

trong đó $h = 1, 2, \dots, H$ với H cố định nhưng tùy ý.

1.1.2 Chuỗi thời gian có tính dừng

Dựa vào kết quả trên, chúng ta có được một phương pháp để đánh giá sơ lược về ý nghĩa của các đỉnh trong $\hat{\rho}(h)$ bằng cách xác định xem vị trí của các đỉnh được quan sát với khoảng $\pm \frac{2}{\sqrt{n}}$ (hoặc cộng / trừ hai lần sai số tiêu chuẩn); với một quá trình nhiễu trắng có khoảng 95% các đỉnh của ACF mẫu nằm trong giới hạn này.

1.1.2 Chuỗi thời gian có tính dừng

Cách xác định tính dừng của một chuỗi thời gian:

- Bằng cách **quan sát đồ thị của chuỗi thời gian**, chúng ta có thể khẳng định chuỗi x_t có tính dừng nếu đồ thị của nó có *xu hướng tăng hoặc giảm trong thời gian dài*.
- **Quan sát đồ thị ACF của chuỗi**: nếu chuỗi có tính dừng thì ACF giảm nhanh, ngẫu nhiên và không theo xu hướng; nếu chuỗi không có tính dừng thì đồ thị giảm chậm tương đối đều đặn theo độ trễ.
- Đối với dữ liệu không có tính dừng, **giá trị của $\hat{\rho}(h)$ (tự tương quan mẫu) tại lag 1 lớn và dương**.
- **Kiểm định ADF (Augmented Dickey–Fuller)**: nếu giá trị của p nhỏ hơn 0.05 thì chuỗi có tính dừng và ngược lại.

1.1.2 Chuỗi thời gian có tính dừng

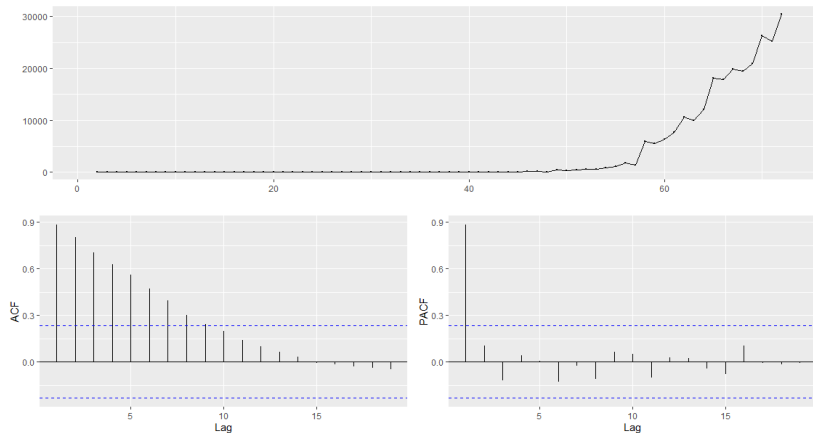


Figure: Chuỗi dữ liệu số ca nhiễm mới COVID-19 theo ngày ở Mỹ

1.1.2 Chuỗi thời gian có tính dừng

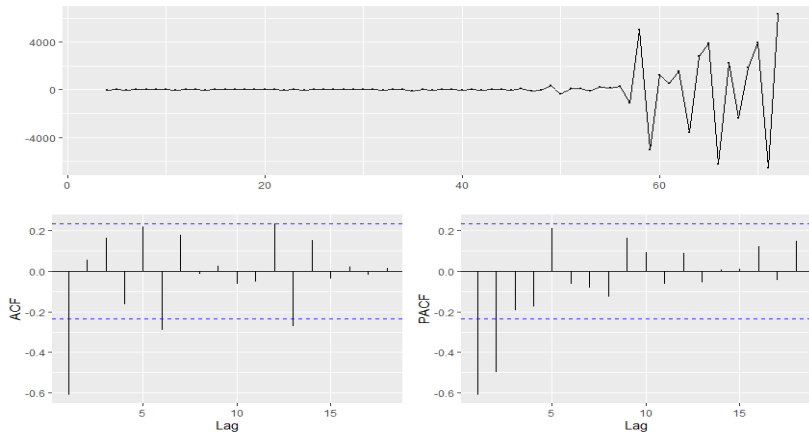


Figure: Chuỗi dữ liệu số ca nhiễm mới COVID-19 theo ngày ở Mỹ

1.1.2 Chuỗi thời gian có tính dừng

- Trong thực tế, phần lớn các chuỗi thời gian là chuỗi không có tính dừng. Đối với mô hình dự báo, chúng tôi cần một chuỗi ổn định. Do vậy, nếu một chuỗi thời gian không có tính dừng thì buộc chúng ta phải biến đổi chuỗi đó thành chuỗi dừng, từ đó mới có thể xây dựng được mô hình và tiến hành dự báo.

1.1.2 Chuỗi thời gian có tính dừng

- Trong thực tế, phần lớn các chuỗi thời gian là chuỗi không có tính dừng. Đối với mô hình dự báo, chúng tôi cần một chuỗi ổn định. Do vậy, nếu một chuỗi thời gian không có tính dừng thì buộc chúng ta phải biến đổi chuỗi đó thành chuỗi dừng, từ đó mới có thể xây dựng được mô hình và tiến hành dự báo.
- Có nhiều cách biến đổi chuỗi không dừng thành chuỗi dừng như dùng phép biến đổi *log*, lấy sai phân, Trong bài báo cáo này, chúng tôi sẽ đề cập đến *phương pháp lấy sai phân giúp loại bỏ hoặc giảm tính không dừng của chuỗi thời gian*.

1.1.2 Chuỗi thời gian có tính dừng

Định nghĩa: [1]

Sai phân bậc 1 được ký hiệu là

$$\nabla x_t = x_t - x_{t-1}. \quad (3)$$

1.1.2 Chuỗi thời gian có tính dừng

Định nghĩa: toán tử Backshift [1]

$$Bx_t = x_{t-1}$$

và mở rộng nó thành lũy thừa

$$B^2x_t = B(Bx_t) = Bx_{t-1} = Bx_{t-2}.$$

Tổng quát lên, ta được

$$B^kx_t = x_{t-k}.$$

1.1.2 Chuỗi thời gian có tính dừng

- Phương trình sai phân bậc 1:

$$\begin{aligned}\nabla x_t &= x_t - x_{t-1} \\ &= x_t - Bx_t = (1 - B)x_t\end{aligned}$$

- Phương trình sai phân bậc 2:

$$\begin{aligned}\nabla^2 x_t &= (1 - B)^2 x_t = (1 - 2B + B^2)x_t \\ &= x_t - 2x_{t-1} + x_{t-2}\end{aligned}$$

- Phương trình sai phân bậc d:

$$\nabla^d = (1 - B)^d$$

1.1.2 Chuỗi thời gian có tính dừng

Sai phân theo mùa

Sai phân theo mùa là sai phân giữa quan sát và quan sát tương ứng theo chu kì năm.

$$\nabla x_t = x_t - x_{t-m}$$

trong đó m = số các mùa vụ. Ví dụ, dữ liệu theo tháng thì $m = 12$, dữ liệu theo quý thì $m = 4$.

1.1.2 Chuỗi thời gian có tính dừng

Phân biệt các quá trình sai phân

- ❶ Sai phân bậc một làm thay đổi giữa một quan sát và các quan sát tiếp theo;
- ❷ Sai phân theo mùa làm thay đổi một năm đến một năm tiếp theo.

1.1.2 Chuỗi thời gian có tính dừng

Chú ý [3]

Đối với chuỗi thời gian không có tính dừng và cần phải lấy sai phân và sai phân theo mùa thì

- Thứ tự lấy sai phân là không quan trọng (lấy sai phân trước hay lấy sai phân theo mùa trước thì đều thu được kết quả như nhau).
- Nếu chuỗi dữ liệu có tính mùa vụ mạnh, chúng tôi sẽ ưu tiên lấy sai phân theo mùa trước bởi vì đôi khi kết quả của chuỗi sẽ có tính dừng ngay mà không cần phải lấy tiếp sai phân bậc 1 nữa.

1.1.3 Hồi quy cổ điển trong chuỗi thời gian

Các tiêu chuẩn thông tin [1]

$$AIC = \ln(\hat{\sigma}_k^2) + \frac{n + 2k}{n} \quad (4)$$

$$AICc = \ln(\hat{\sigma}_k^2) + \frac{n + k}{n - k - 2} \quad (5)$$

$$SIC = \ln(\hat{\sigma}_k^2) + \frac{k \ln(n)}{n} \quad (6)$$

1.1.3 Hồi quy cổ điển trong chuỗi thời gian

- Căn của sai số bình phương trung bình (RMSE - Root Mean Squared Error):

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n} \sum_{t=1}^n (Y_t - \hat{Y}_t)^2}.$$

- Sai số tuyệt đối trung bình (MAE - Mean Absolute Error):

$$MAE = \frac{\sum_{t=1}^n |Y_t - \hat{Y}_t|}{n}.$$

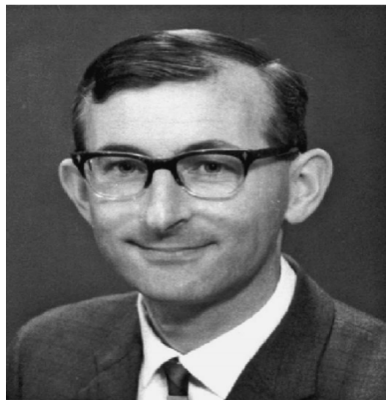
- Sai số phần trăm tuyệt đối trung bình (MAPE - Mean Absolute Percentage Error):

$$MAPE = \frac{1}{n} \sum_{t=1}^n \left| \frac{Y_t - \hat{Y}_t}{Y_t} \right|.$$

1.2 Mô hình ARIMA



George E. P. Box



Gwilym M. Jenkins

1.2.1 Mô hình ARMA

Định nghĩa [1]

Mô hình tự hồi quy bậc p , ký hiệu là $AR(p)$, có dạng:

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + w_t, \quad (7)$$

trong đó x_t là có tính dừng, $\phi_1, \phi_2, \dots, \phi_p$ là các hằng số ($\phi_p \neq 0$). Trừ khi có phát biểu khác, chúng ta giả sử rằng w_t là một chuỗi nhiễu trắng Gaussian có giá trị trung bình bằng 0 và phương sai σ_w^2 . Khi đó giá trị trung bình của x_t trong (7) bằng 0.

1.2.1 Mô hình ARMA

Định nghĩa [1]

Nếu giá trị trung bình μ của x_t khác 0 thì thế $x_t = x_t - \mu$ vào (7), ta được,

$$x_t - \mu = \phi_1(x_{t-1} - \mu) + \phi_2(x_{t-2} - \mu) + \cdots + \phi_p(x_{t-p} - \mu) + w_t, \quad (8)$$

hoặc

$$x_t = \alpha + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + w_t, \quad (9)$$

trong đó $\alpha = \mu(1 - \phi_1 - \cdots - \phi_p)$.

1.2.1 Mô hình ARMA

Chúng ta cũng có thể sử dụng toán tử backshift để viết mô hình $AR(p)$ (7) dưới dạng

$$\phi(B)x_t = w_t, \quad (10)$$

trong đó $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p$ là *toán tử tự hồi quy*.

1.2.1 Mô hình ARMA

Định nghĩa: [1]

Mô hình trung bình trượt bậc q , ký hiệu $MA(q)$, có dạng

$$x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdots + \theta_q w_{t-q}, \quad (11)$$

với q độ trễ theo trung bình trượt, $\theta_1, \theta_2, \dots, \theta_q$ ($\theta_q \neq 0$) là các tham số, w_t là nhiễu trắng Gaussian.

1.2.1 Mô hình ARMA

Chúng ta cũng có thể viết quá trình MA(q) ở dạng tương đương

$$x_t = \theta(B)w_t, \quad (12)$$

trong đó $\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$ là *toán tử trung bình trượt*.

1.2.1 Mô hình ARMA

Định nghĩa: Mô hình tự hồi quy trung bình trượt ARMA(p,q)

Một chuỗi thời gian $\{x_t; t = 0, \pm 1, \pm 2, \dots\}$ là **ARMA(p, q)** nếu nó có tính dừng và có dạng

$$x_t = \phi_1 x_{t-1} + \dots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \dots + \theta_q w_{t-q}, \quad (13)$$

với $\phi_p \neq 0, \theta_q \neq 0$ và $\sigma_w^2 > 0$; p và q lần lượt là bậc của tự hồi quy và trung bình trượt; $\{w_t; t = 0, \pm 1, \pm 2, \dots\}$ là dãy nhiễu trắng Gaussian. Nếu x_t có giá trị trung bình $\mu \neq 0$, chúng ta đặt $\alpha = \mu(1 - \phi_1 - \dots - \phi_p)$ và mô hình có dạng

$$x_t = \alpha + \phi_1 x_{t-1} + \dots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \dots + \theta_q w_{t-q}. \quad (14)$$

1.2.1 Mô hình ARMA

Mô hình ARMA(p, q) trong (13) còn có thể được viết dưới dạng ngắn gọn:

$$\phi(B)x_t = \theta(B)w_t, \quad (15)$$

trong đó $\phi(B)$ là toán tử tự hồi quy và $\theta(B)$ là toán tử trung bình trượt.

1.2.2 Phương trình sai phân

Phương trình sai phân thuần nhất bậc p

Cho dãy số u_0, u_1, u_2, \dots , khi đó *phương trình sai phân thuần nhất bậc p* có dạng

$$u_n - \alpha_1 u_{n-1} - \dots - \alpha_p u_{n-p} = 0, \quad \alpha_p \neq 0, \quad n = p, p+1, \dots \quad (16)$$

Đa thức liên kết là

$$\alpha(z) = 1 - \alpha_1 z - \dots - \alpha_p z^p.$$

1.2.2 Phương trình sai phân

Phương trình sai phân thuần nhất bậc p

Cho dãy số u_0, u_1, u_2, \dots , khi đó *phương trình sai phân thuần nhất bậc p* có dạng

$$u_n - \alpha_1 u_{n-1} - \dots - \alpha_p u_{n-p} = 0, \quad (17)$$

với $\alpha_p \neq 0$, $n = p, p+1, \dots$

Đa thức liên kết là

$$\alpha(z) = 1 - \alpha_1 z - \dots - \alpha_p z^p.$$

1.2.2 Phương trình sai phân

Giả sử $\alpha(z)$ có r nghiệm phân biệt, z_1 có bội m_1 , z_2 có bội m_2, \dots , và z_r có bội m_r , khi đó $m_1 + m_2 + \dots + m_r = p$. Nghiệm tổng quát của phương trình sai phân (17) là

$$u_n = z_1^{-n} P_1(n) + z_2^{-n} P_2(n) + \dots + z_r^{-n} P_r(n), \quad (18)$$

trong đó $P_j(n)$ là một đa thức theo n có bậc $m_j - 1$, với $j = 1, 2, \dots, r$. Với p điều kiện ban đầu u_0, \dots, u_{p-1} , chúng ta có thể tìm được $P_j(n)$.

1.2.3 Hàm ACF và PACF

ACF của một MA(q):

$$\rho(h) = \begin{cases} \frac{\sum_{j=0}^{q-h} \theta_j \theta_{j+h}}{1 + \theta_1^2 + \dots + \theta_q^2} & , 1 \leq h \leq q \\ 0 & , h > q. \end{cases} \quad (19)$$

1.2.3 Hàm ACF và PACF

Để xác định PACF cho chuỗi thời gian có tính dừng có giá trị trung bình bằng 0, đặt x_h^{h-1} là biểu diễn hồi quy của x_h trên $\{x_{h-1}, x_{h-2}, \dots, x_1\}$, được viết dưới dạng

$$x_h^{h-1} = \beta_1 x_{h-1} + \beta_2 x_{h-2} + \dots + \beta_{h-1} x_1. \quad (20)$$

Ngoài ra, đặt x_0^{h-1} là biểu diễn hồi quy của x_0 trên $\{x_1, x_2, \dots, x_{h-1}\}$, khi đó

$$x_0^{h-1} = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{h-1} x_{h-1}. \quad (21)$$

1.2.3 Hàm ACF và PACF

Định nghĩa: Hàm PACF [1]

Hàm tự tương quan từng phần (*PACF* - *Partial Auto-Correlation Function*) của một quá trình dừng x_t , ký hiệu là ϕ_{hh} , với $h = 1, 2, \dots$ có dạng

$$\phi_{11} = \text{corr}(x_1, x_0) = \rho(1) \quad (22)$$

và

$$\phi_{hh} = \text{corr}(x_h - x_h^{h-1}, x_0 - x_0^{h-1}), \quad h \geq 2. \quad (23)$$

1.2.4 Phương trình dự báo

Tính chất [1]

Dự báo tuyến tính tốt nhất (*BLP - Best Linear Prediction*) cho các quá trình có tính dừng.

Cho dữ liệu x_1, \dots, x_n , công cụ dự báo tuyến tính tốt nhất, $x_{n+m}^n = \alpha_0 + \sum_{k=1}^n \alpha_k x_k$ của x_{n+m} , với $m \geq 1$ được tìm thấy bằng cách giải

$$E[(x_{n+m} - x_{n+m}^n)x_k] = 0, \quad k = 0, 1, \dots, n. \quad (24)$$

trong đó $x_0 = 1$.

1.2.4 Phương trình dự báo

Tính chất [1]

Dự báo tuyến tính tốt nhất (*BLP - Best Linear Prediction*) cho các quá trình có tính dừng.

Cho dữ liệu x_1, \dots, x_n , công cụ dự báo tuyến tính tốt nhất, $x_{n+m}^n = \alpha_0 + \sum_{k=1}^n \alpha_k x_k$ của x_{n+m} , với $m \geq 1$ được tìm thấy bằng cách giải

$$E[(x_{n+m} - x_{n+m}^n)x_k] = 0, \quad k = 0, 1, \dots, n. \quad (24)$$

trong đó $x_0 = 1$.

Giải phương trình (24), ta được

$$x_{n+m}^n = \mu + \sum_{k=1}^n \alpha_k (x_k - \mu). \quad (25)$$

1.2.4 Phương trình dự báo

Dự báo trước 1 bước [1]

BLP của x_{n+1} có dạng

$$x_{n+1}^n = \phi_{n1}x_n + \phi_{n2}x_{n-1} + \cdots + \phi_{nn}x_1, \quad (26)$$

Sử dụng tính chất dự báo tuyến tính tốt nhất cho các quá trình có tính dừng, các hệ số $\{\phi_{n1}, \phi_{n2}, \dots, \phi_{nn}\}$ thỏa mãn

$$E[(x_{n+1} - \sum_{j=1}^n \phi_{nj}x_{n+1-j})x_{n+1-k}] = 0, \quad k = 1, \dots, n,$$

1.2.4 Phương trình dự báo

Dự báo trước m bước [1]

Cho dữ liệu $\{x_1, \dots, x_n\}$, dự báo trước m bước là

$$x_{n+m}^n = \phi_{n1}^{(m)} x_n + \phi_{n2}^{(m)} x_{n-1} + \dots + \phi_{nn}^{(m)} x_1, \quad (27)$$

trong đó $\{\phi_{n1}^{(m)}, \phi_{n2}^{(m)}, \dots, \phi_{nn}^{(m)}\}$ thỏa mãn phương trình dự báo

$$\sum_{j=1}^n \phi_{nj}^{(m)} E(x_{n+1-j} x_{n+1-k}) = E(x_{n+m} x_{n+1-k}), \quad k = 1, \dots, n,$$

1.2.5 Mô hình ARIMA

Định nghĩa: [1]

Một quá trình x_t được gọi là **ARIMA**(p, d, q) nếu

$$\nabla^d x_t = (1 - B)^d x_t$$

là *ARMA*(p, q). Mô hình *ARIMA*(p, d, q) có dạng tổng quát là

$$\phi(B)(1 - B)^d x_t = \theta(B)w_t. \quad (28)$$

Nếu $E(\nabla^d x_t) = \mu$, chúng ta viết mô hình có dạng

$$\phi(B)(1 - B)^d x_t = \alpha + \theta(B)w_t,$$

trong đó $\alpha = \mu(1 - \phi_1 - \dots - \phi_p)$.

1.2.5 Mô hình ARIMA

Mô hình ARIMA(1,1,1):

$$(1 - \phi_1 B)(1 - B)x_t = \alpha + (1 + \theta_1 B)w_t$$

↑

↑

↑

$AR(1)$ Sai phân bậc 1 $MA(1)$

Ta có:

$$x_t = \alpha + \phi_1 x_{t-1} - \phi_1 x_{t-2} + \theta_1 w_{t-1} + w_t$$

1.2.5 Mô hình ARIMA

Các bước để xây dựng mô hình ARIMA

- 1 Kiểm tra tính dừng của chuỗi thời gian.
- 2 Chuyển một chuỗi không có tính dừng về chuỗi có tính dừng bằng cách lấy sai phân.
- 3 Xác định bậc p và q .
- 4 Kiểm định độ chính xác của mô hình.
- 5 Dự báo.

1.2.5 Mô hình ARIMA

Định nghĩa [1]

Mô hình tự hồi quy trung bình trượt theo mùa $\text{ARMA}(P, Q)_s$ có dạng:

$$\Phi_P(B^s)x_t = \Theta_Q(B^s)w_t, \quad (29)$$

trong đó s là chu kỳ;

$\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \dots - \Phi_P B^{Ps}$ là toán tử tự hồi quy theo mùa bậc P ; $\Theta_Q(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \dots + \Theta_Q B^{Qs}$ là toán tử trung bình trượt theo mùa bậc Q .

1.2.5 Mô hình ARIMA

Tổng quát, chúng ta có thể kết hợp các toán tử theo mùa và không theo mùa thành mô hình tự hồi quy trung bình trượt theo mùa, được kí hiệu là $\text{ARMA}(p, q) \times (P, Q)_s$ và mô hình tổng quát được viết dưới dạng

$$\Phi_P(B^s)\phi(B)x_t = \Theta_Q(B^s)\theta(B)w_t. \quad (30)$$

1.2.5 Mô hình ARIMA

Định nghĩa [1]

Sai phân theo mùa bậc D được định nghĩa là

$$\nabla_s^D x_t = (1 - B^s)^D x_t, \quad (31)$$

trong đó $D = 1, 2, \dots$

1.2.5 Mô hình ARIMA

Định nghĩa [1]

Mô hình tự hồi quy tích hợp trung bình trượt theo mùa của Box và Jenkins (1970) có dạng tổng quát là

$$\Phi_P(B^s)\phi(B)\nabla_s^D\nabla^d x_t = \alpha + \Theta_Q(B^s)\theta(B)w_t, \quad (32)$$

trong đó w_t là chuỗi nhiễu trắng Gaussian. Mô hình tổng quát được ký hiệu là $ARIMA(p, d, q) \times (P, D, Q)_s$. Các toán tử tự hồi quy bậc p và trung bình trượt q thường được biểu thị bằng đa thức $\phi(B)$ và $\theta(B)$; toán tử tự hồi quy theo mùa bậc P và toán tử trung bình trượt theo mùa bậc Q lần lượt là $\Phi_P(B^s)$ và $\Theta_Q(B^s)$, và các thành phần sai phân không theo mùa và theo mùa là $\nabla^d = (1 - B)^d$ và $\nabla_s^D = (1 - B^s)^D$.

CHƯƠNG 2. PHÂN TÍCH CHUỖI THỜI GIAN BẰNG MÔ HÌNH ARIMA VỚI PHẪM MỀM R.

2.1 Tổng quan về R và các gói lệnh

2.2 Phân tích dữ liệu COVID-19

2.2.1 Xử lý số liệu thô.

2.2.2 Phân tích tổng quan COVID-19 trên toàn thế giới.

2.2.3 Dự báo số ca nhiễm mới COVID-19 tại Mỹ.

2.2.4 Dự báo số ca tử vong mới COVID-19 tại Italy.

2.3 Phân tích và dự báo lượng mưa tại trạm quan trắc Quy Nhơn

2.4 Website dashboard COVID-19 với Shinyapps

2.1 Tổng quan về R và các gói lệnh

Tổng quan về R

- Phân tích thống kê và đồ thị.
- Mã nguồn mở, miễn phí.
- Có đầy đủ tính năng của các phần mềm thương mại đắt tiền như SPSS, AMOS, STATA hay EViews và nó có tính năng vượt trội hơn hẳn.
- Công cụ cho Data Mining, Big Data, Data Visualization và Machine Learning.
- Năm 1996, hai nhà thống kê học Ross Ihaka và Robert Gentleman thuộc Trường đại học Auckland, New Zealand phát hoạ ra ngôn ngữ R.

2.1 Tổng quan về R và các gói lệnh

Các gói lệnh được sử dụng trong bài báo cáo

- Gói lệnh "forecast" của Rob Hyndman và các cộng sự (Version 8.12 - 2020).
- Gói lệnh "ggplot2" dùng để vẽ biểu đồ với đa dạng lựa chọn với nhiều tùy biến.
- Gói lệnh "magrittr" chứa các toán tử.
- Gói lệnh "gridExtra" dùng để chứa các ảnh dưới dạng lưới.
- Gói lệnh "kableExtra" dùng để chuyển dữ liệu bảng từ R sang code Latex với nhiều tùy chỉnh.

2.2 Phân tích dữ liệu COVID-19



2.2.1 Xử lý dữ liệu thô

- Chúng tôi lấy dữ liệu toàn cầu dưới dạng thô được thống kê hàng ngày của Johns Hopkins CSSE tại <https://github.com/CSSEGISandData/COVID-19>.
- Các nguồn dữ liệu được thống kê bao gồm Tổ chức Y tế Thế giới (WHO), Trung tâm Kiểm soát và Phòng ngừa Dịch bệnh Hoa Kỳ (CDC) và Ủy ban Y tế Quốc gia của Cộng hòa Nhân dân Trung Hoa (NHC).
- Dữ liệu này được thu thập bởi Đại học Johns Hopkins, được phát triển bởi nhóm ESRI Living Atlas.

2.2.2 Phân tích tổng quan COVID-19 trên toàn thế giới.

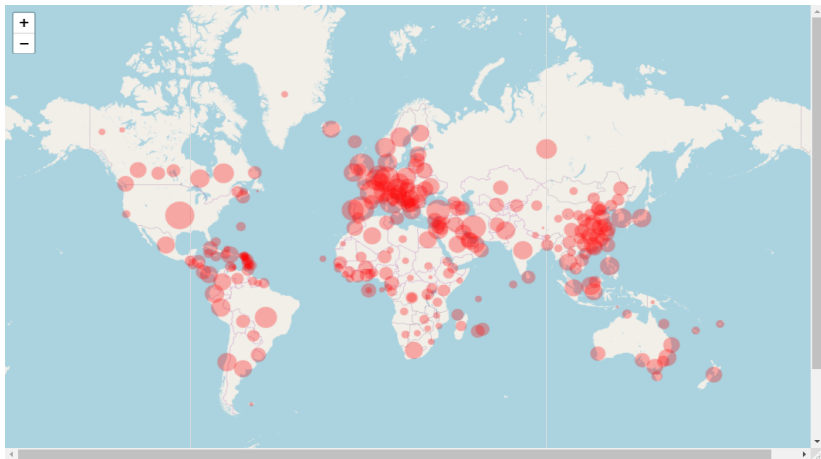


Figure: Bản đồ phân bố dịch COVID-19 trên toàn thế giới (7/4/2020)

2.2.2 Phân tích tổng quan COVID-19 trên toàn thế giới.

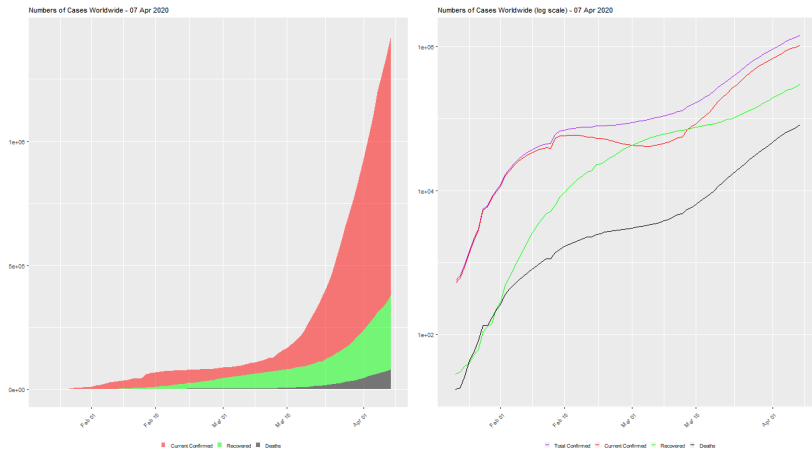


Figure: Đồ thị dữ liệu COVID-19 trên toàn thế giới (7/4/2020)

2.2.2 Phân tích tổng quan COVID-19 trên toàn thế giới.

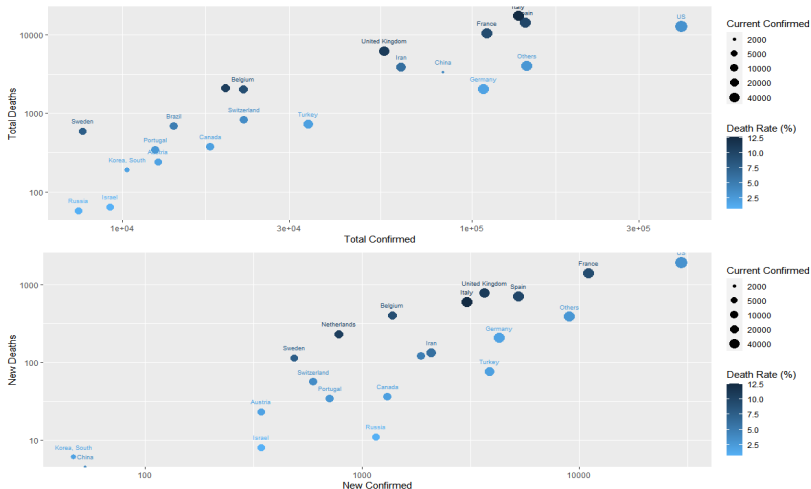


Figure: Top 20 quốc gia (7/4/2020)

2.2.2 Phân tích tổng quan COVID-19 trên toàn thế giới.

Các nghiên cứu sử dụng mô hình ARIMA như là một công cụ hữu ích trong việc theo dõi và dự báo xu hướng thay đổi trong các bệnh truyền nhiễm.

2.2.2 Phân tích tổng quan COVID-19 trên toàn thế giới.

Các nghiên cứu sử dụng mô hình ARIMA như là một công cụ hữu ích trong việc theo dõi và dự báo xu hướng thay đổi trong các bệnh truyền nhiễm.

- L.LIU và các cộng sự của mình đã sử dụng mô hình ARIMA để dự báo tỷ lệ mắc bệnh tay, chân, miệng ở tỉnh Tứ Xuyên, Trung Quốc [1].

2.2.2 Phân tích tổng quan COVID-19 trên toàn thế giới.

Các nghiên cứu sử dụng mô hình ARIMA như là một công cụ hữu ích trong việc theo dõi và dự báo xu hướng thay đổi trong các bệnh truyền nhiễm.

- L.LIU và các cộng sự của mình đã sử dụng mô hình ARIMA để dự báo tỷ lệ mắc bệnh tay, chân, miệng ở tỉnh Tứ Xuyên, Trung Quốc [1].
- Li và các cộng sự đã áp dụng mô hình ARIMA để dự báo tỷ lệ mắc bệnh sốt xuất huyết tại tỉnh Lâm Nghi, Trung Quốc [2].

2.2.2 Phân tích tổng quan COVID-19 trên toàn thế giới.

Các nghiên cứu sử dụng mô hình ARIMA như là một công cụ hữu ích trong việc theo dõi và dự báo xu hướng thay đổi trong các bệnh truyền nhiễm.

- L.LIU và các cộng sự của mình đã sử dụng mô hình ARIMA để dự báo tỷ lệ mắc bệnh tay, chân, miệng ở tỉnh Tứ Xuyên, Trung Quốc [1].
- Li và các cộng sự đã áp dụng mô hình ARIMA để dự báo tỷ lệ mắc bệnh sốt xuất huyết tại tỉnh Lâm Nghi, Trung Quốc [2].
- Earnest cùng các cộng sự đã dùng mô hình ARIMA như một công cụ hữu ích cho quản trị viên và các bác sỹ trong việc lập kế hoạch phân bố giường bệnh cho các bệnh nhân trong đợt dịch SARS bùng phát [3].

2.2.3 Dự báo số ca nhiễm mới COVID-19 tại Mỹ.

Tình hình nước Mỹ:

2.2.3 Dự báo số ca nhiễm mới COVID-19 tại Mỹ.

Tình hình nước Mỹ:

- hơn 350000 ca nhiễm và gần 1300 ca tử vong (tính tại thời điểm ngày 7/4/2020). Hoa Kỳ trở thành tâm điểm toàn cầu của đại dịch khi vượt qua số ca nhiễm ở Trung Quốc - nơi mầm bệnh của COVID-19.

2.2.3 Dự báo số ca nhiễm mới COVID-19 tại Mỹ.

Tình hình nước Mỹ:

- hơn 350000 ca nhiễm và gần 1300 ca tử vong (tính tại thời điểm ngày 7/4/2020). Hoa Kỳ trở thành tâm điểm toàn cầu của đại dịch khi vượt qua số ca nhiễm ở Trung Quốc - nơi mầm bệnh của COVID-19.
- Hoa Kỳ là một trong những quốc gia có nền y tế phát triển bậc nhất thế giới tuy nhiên vẫn tồn tại một số lỗ hổng nhất định.

2.2.3 Dự báo số ca nhiễm mới COVID-19 tại Mỹ.

Tình hình nước Mỹ:

- hơn 350000 ca nhiễm và gần 1300 ca tử vong (tính tại thời điểm ngày 7/4/2020). Hoa Kỳ trở thành tâm điểm toàn cầu của đại dịch khi vượt qua số ca nhiễm ở Trung Quốc - nơi mầm bệnh của COVID-19.
- Hoa Kỳ là một trong những quốc gia có nền y tế phát triển bậc nhất thế giới tuy nhiên vẫn tồn tại một số lỗ hổng nhất định.
- COVID-19 xuất hiện vào thời điểm trọng yếu của chính trị Mỹ nên động thái xử lý dịch bệnh chậm trễ và thiếu dứt khoát ngay từ đầu.

⇒ Dự báo diễn biến tình hình dịch bệnh ở Mỹ là việc hết sức quan trọng và cần thiết.

2.2.3 Dự báo số ca nhiễm mới COVID-19 tại Mỹ.

```
data %<>% filter(country=="US")  
data.US <- data$new.confirmed[1:74] %>% ts()  
train <- data$new.confirmed[1:70] %>% ts()  
test <- data$new.confirmed[71:74] %>% ts()
```


2.2.3 Dự báo số ca nhiễm mới COVID-19 tại Mỹ.

Bước 1: Kiểm tra tính dừng của tập "train"

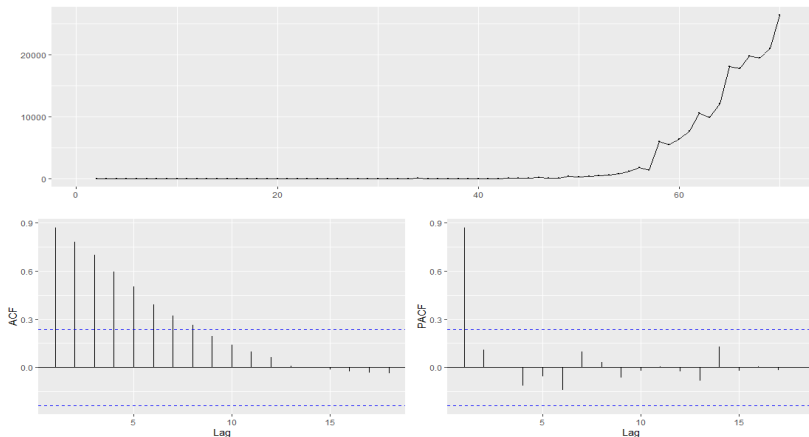


Figure: Đồ thị ACF và PACF của tập "train"

2.2.3 Dự báo số ca nhiễm mới COVID-19 tại Mỹ.

```
adf.test (train, alternative = "stationary")  
##    Augmented Dickey-Fuller Test  
## data: train  
## Dickey-Fuller = 3.3984, Lag order = 4, p-value = 0.99
```

2.2.3 Dự báo số ca nhiễm mới COVID-19 tại Mỹ.

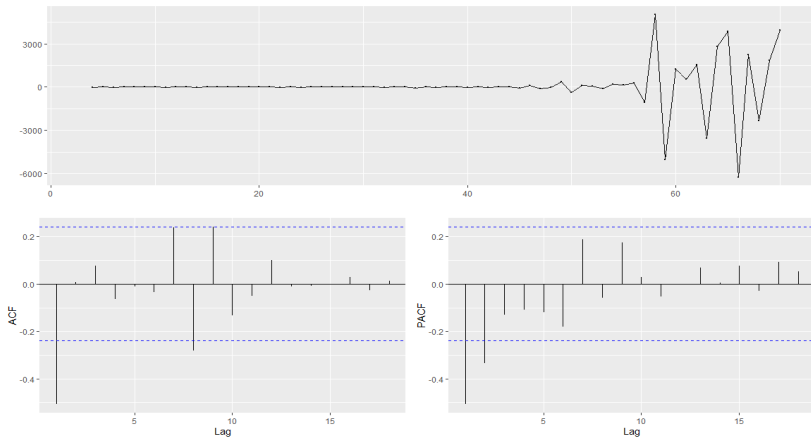
Bước 2: Chuyển đổi chuỗi không có tình dừng thành chuỗi có tính dừng

```
train %>% diff() -> diff1  
diff1 %>% ggtsdisplay()
```

2.2.3 Dự báo số ca nhiễm mới COVID-19 tại Mỹ.

```
adf.test (diff1, alternative = "stationary")  
##    Augmented Dickey-Fuller Test  
## data: train.diff  
## Dickey-Fuller = -1.6228, Lag order = 4, p-value = 0.7283  
## alternative hypothesis: stationary
```

2.2.3 Dự báo số ca nhiễm mới COVID-19 tại Mỹ.

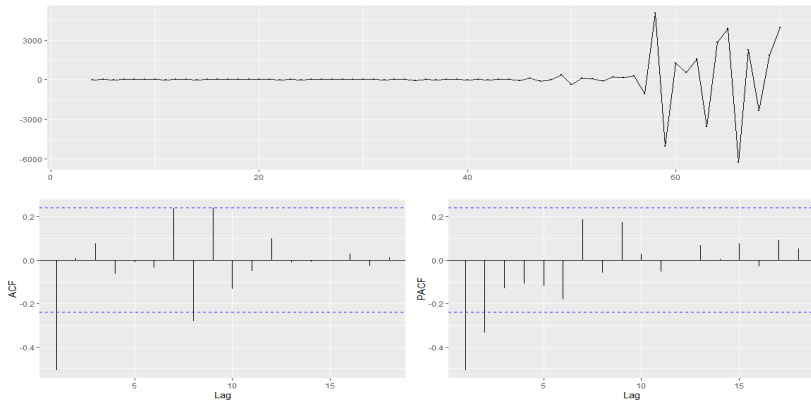


2.2.3 Dự báo số ca nhiễm mới COVID-19 tại Mỹ.

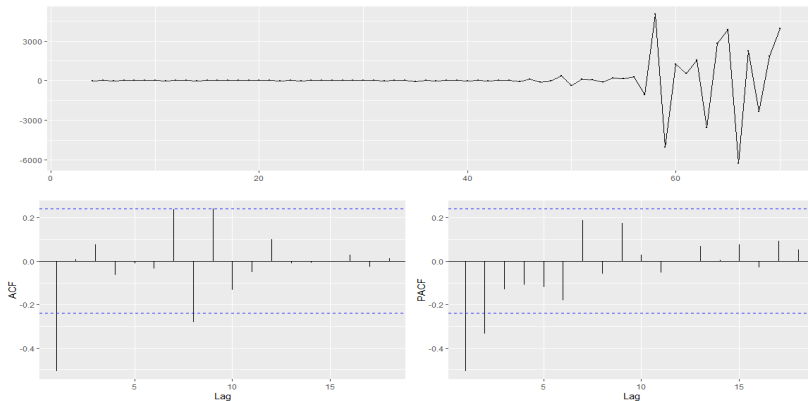
```
adf.test (train.diff, alternative = "stationary")  
##      Augmented Dickey-Fuller Test  
## data: train.diff  
## Dickey-Fuller = -6.4328, Lag order = 4, p-value = 0.01
```

Kết quả của kiểm tra ADF cho thấy $p\text{-value} = 0.01 < 0.05 \implies$
Chuỗi tập "train" có tính dừng sau khi lấy sai phân bậc 2.

2.2.3 Dự báo số ca nhiễm mới COVID-19 tại Mỹ.



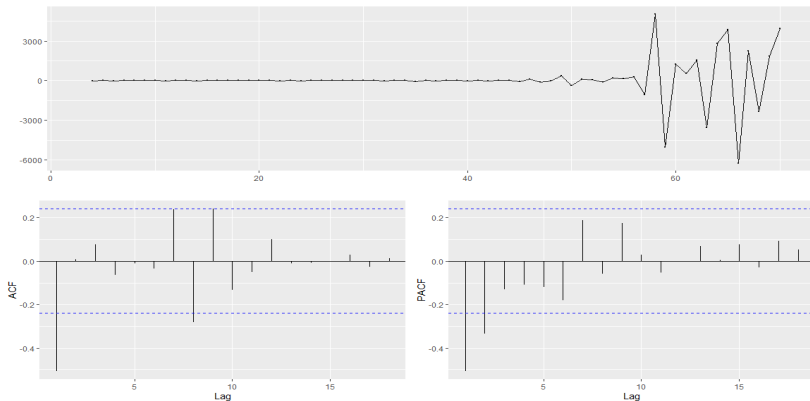
2.2.3 Dự báo số ca nhiễm mới COVID-19 tại Mỹ.



⇒ Mô hình đơn giản từ đồ thị ACF là $ARIMA(0, 2, 2)$

⇒

2.2.3 Dự báo số ca nhiễm mới COVID-19 tại Mỹ.



⇒ Mô hình đơn giản từ đồ thị ACF là $ARIMA(0, 2, 2)$

⇒ $AIC_c = 1116.71$

2.2.3 Dự báo số ca nhiễm mới COVID-19 tại Mỹ.

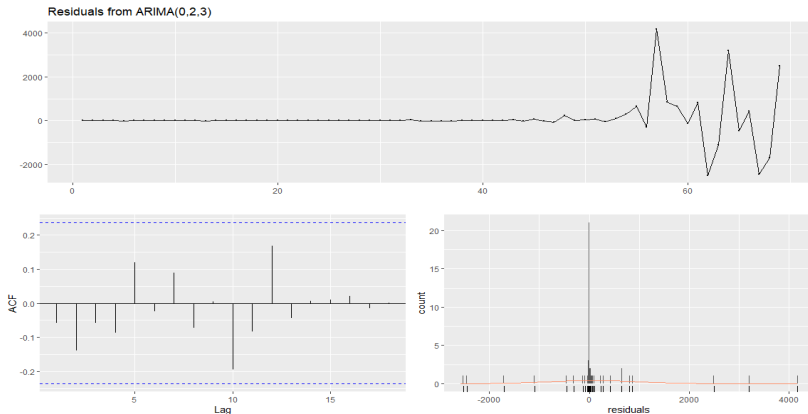
Bước 4: Khớp và chọn mô hình phù hợp nhất

STT	ARIMA Models	AICc	RMSE train
1	ARIMA(0,2,2)	1116.71	904.6045
2	ARIMA(1,2,2)	1116.61	887.8929
3	ARIMA(0,2,3)	1115.40	862.7566
4	ARIMA(1,2,3)	1115.48	878.6758
5	ARIMA(0,2,1)	1135.40	1095.719

⇒ Mô hình ARIMA(0, 2, 3) là mô hình phù hợp nhất (AICc và RMSE nhỏ nhất).

2.2.3 Dự báo số ca nhiễm mới COVID-19 tại Mỹ.

Bước 5: Kiểm tra phần dư từ mô hình được chọn



2.2.3 Dự báo số ca nhiễm mới COVID-19 tại Mỹ.

Bước 5: Đánh giá sai số của mô hình ARIMA(0,2,3) và dự báo

```
best.md <- Arima(train, order = c(0,2,3))  
test.fc <- forecast(best.md, h = 4)$mean %>% as.vector()  
df_md <- tibble(Actual = test %>% as.vector(),  
Forecast = test.fc %>% round(0),  
Error = Forecast - Actual,  
Error_Percent = round(Error / Actual, 2))
```

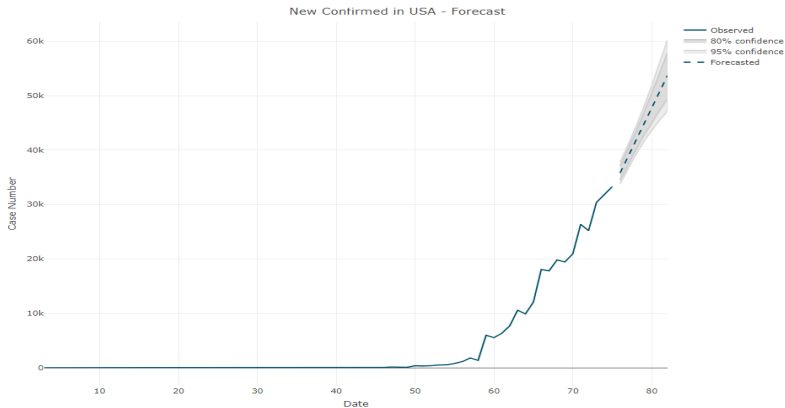
2.2.3 Dự báo số ca nhiễm mới COVID-19 tại Mỹ.

Date	Actual	Forecast	Error	Error_Percent
2020-04-01	25200	26711	1511	6%
2020-04-02	30390	27733	-2657	-9%
2020-04-03	31824	29476	-2348	-7%
2020-04-04	33267	31219	-2048	-6%

2.2.3 Dự báo số ca nhiễm mới COVID-19 tại Mỹ.

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
75	35795.23	34513.19	37077.28	33834.52	37755.95
76	38739.10	37311.61	40166.59	36555.94	40922.26
77	41727.88	40074.77	43380.99	39199.66	44256.10
78	44716.66	42636.16	46797.16	41534.81	47898.51
79	47705.44	45023.33	50387.55	43603.51	51807.37
80	50694.22	47274.09	54114.35	45463.59	55924.85
81	53683.00	49415.92	57950.08	47157.06	60208.94

2.2.3 Dự báo số ca nhiễm mới COVID-19 tại Mỹ.



2.2.4 Dự báo số ca tử vong mới COVID-19 tại Ý.

Tình hình COVID-19 tại Ý

- Cơ quan Bảo vệ Dân sự xác nhận thêm 3.039 ca dương tính với SARS-CoV-2 trong ngày 7/4. Mặc dù đây là số ca nhiễm mới thấp nhất kể từ ngày 17/3 tại quốc gia này, nhưng số ca tử vong lại tăng lên rất cao.
- Có thêm 604 ca tử vong do SARS-CoV-2 ở Italy, nâng tổng số bệnh nhân COVID-19 thiệt mạng lên đến 17.127 người - mức cao nhất trên thế giới.

2.2.4 Dự báo số ca tử vong mới COVID-19 tại Ý.

Tình hình COVID-19 tại Ý

- Cơ quan Bảo vệ Dân sự xác nhận thêm 3.039 ca dương tính với SARS-CoV-2 trong ngày 7/4. Mặc dù đây là số ca nhiễm mới thấp nhất kể từ ngày 17/3 tại quốc gia này, nhưng số ca tử vong lại tăng lên rất cao.
- Có thêm 604 ca tử vong do SARS-CoV-2 ở Italy, nâng tổng số bệnh nhân COVID-19 thiệt mạng lên đến 17.127 người - mức cao nhất trên thế giới.

Vậy nguyên nhân dẫn đến sự tăng vọt này là gì?

2.2.4 Dự báo số ca tử vong mới COVID-19 tại Ý.

Nhóm Ioannidis vừa có một bài báo được đăng trên tạp chí JAMA Int Med đưa ra những quan điểm trả lời cho câu hỏi trên [1].

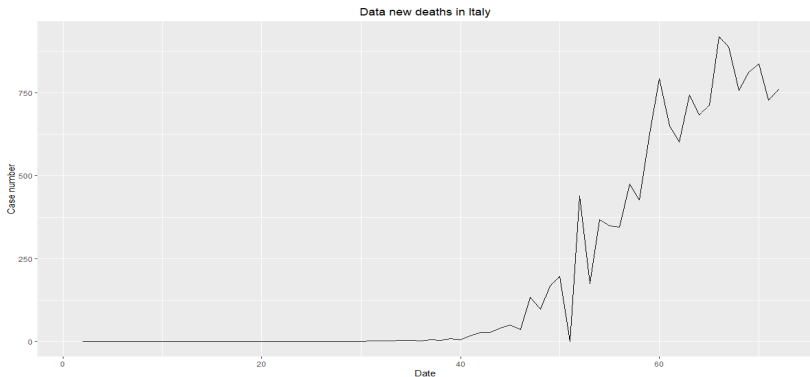
- Yếu tố lão hóa dân số: Ý có dân số già bậc nhất châu Âu. Gần một phần tư (23%) dân số của nước này có độ tuổi từ 65 trở lên. Hơn thế, những người này tiền sử bệnh như hô hấp, tim mạch, tiểu đường và ung thư. Do đó, gánh nặng dịch bệnh đã đè lên bệnh trạng sẵn có.
- Hệ thống y tế: mặc dù Ý là nước có hệ thống y tế Nhà nước rất tốt, nhưng số giường bệnh ICU thì rất khiêm tốn (toàn quốc chỉ có 5090 giường, tức 8.4 trên 100,000 dân).

2.2.4 Dự báo số ca tử vong mới COVID-19 tại Ý.

```
data %<>% filter(country=="Italy")  
data.Italy <- data$new.deaths[1:77] %>% ts()  
train <- data$new.deaths[1:72] %>% ts()  
test <- data$new.deaths[73:77] %>% ts()
```

2.2.4 Dự báo số ca tử vong mới COVID-19 tại Ý.

Bước 1: Vẽ đồ thị số ca tử vong mới tại Ý trên tập "train"



2.2.4 Dự báo số ca tử vong mới COVID-19 tại Ý.

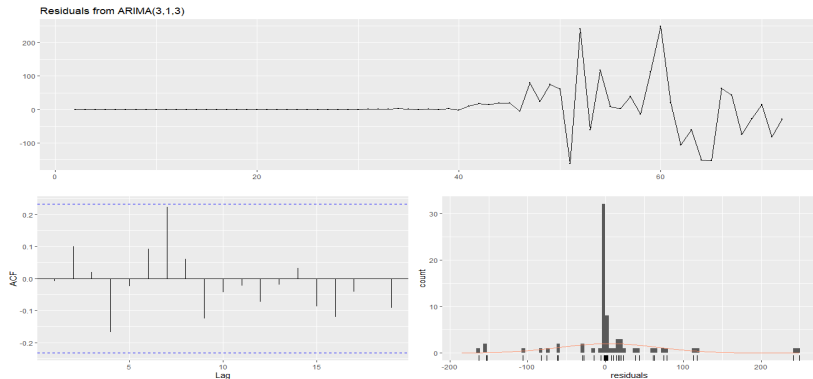
Bước 2: Sử dụng hàm "auto.arima" để tìm ra mô hình tốt nhất

```
arima.Italy <- auto.arima(train)

## Best model: ARIMA(3,1,3)
## Coefficients:
##          ar1      ar2      ar3      ma1      ma2      ma3
##          0.0056 0.7302 0.057 -0.9051 -0.6095 0.8668
## s.e. 0.2056 0.1923 0.132  0.2259  0.3003 0.2290
## sigma^2 estimated as 4356: log likelihood=-392.39
## AIC=798.77 AICc=800.58 BIC=814.51
```

2.2.4 Dự báo số ca tử vong mới COVID-19 tại Ý.

Bước 3: Kiểm tra phần dư từ mô hình ARIMA(3,1,3)



2.2.4 Dự báo số ca tử vong mới COVID-19 tại Ý.

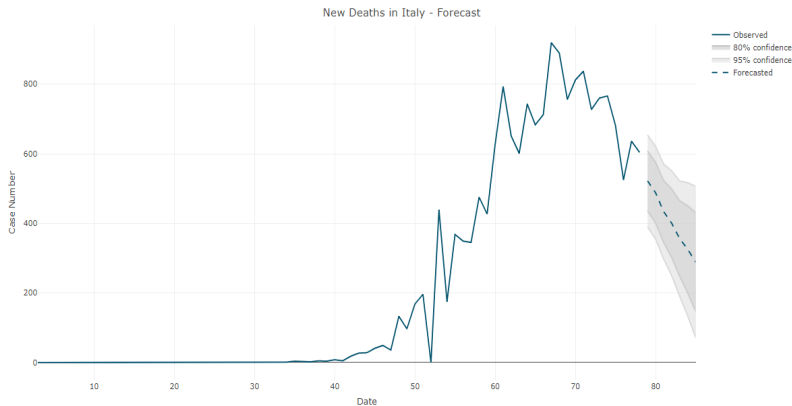
Bước 4: Đánh giá sai số từ mô hình ARIMA(3,1,3) và dự báo

Date	Actual	Forecast	Error	Error_Percent
2020-04-03	766	770	4	1%
2020-04-04	681	735	54	8%
2020-04-05	525	720	195	37%
2020-04-06	636	694	58	9%
2020-04-07	604	681	77	13%

2.2.4 Dự báo số ca tử vong mới COVID-19 tại Ý.

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
78	521.8161	435.4723	608.1600	389.76460	653.8677
79	487.9330	400.8170	575.0490	354.70053	621.1654
80	433.8354	344.6312	523.0397	297.40930	570.2616
81	400.5004	302.0105	498.9904	249.87302	551.1279
82	356.1975	247.7964	464.5986	190.41228	521.9827
83	325.4238	200.2443	450.6032	133.97834	516.8692
84	288.5974	146.3005	430.8943	70.97312	506.2218

2.2.4 Dự báo số ca tử vong mới COVID-19 tại Ý.



2.3 Phân tích và dự báo lượng mưa tại trạm quan trắc Quy Nhơn.

Lợi ích của việc dự báo lượng mưa?

- Lượng mưa là một trong những hiện tượng quan trọng nhất của hệ thống tự nhiên có ảnh hưởng chung đến biến đổi khí hậu.
- Do đó, việc mô hình hóa và dự báo nó là cần thiết để giải quyết một số vấn đề liên quan đến quy hoạch và quản lý hệ thống tài nguyên nước, công trình thủy lợi, quản lý nông nghiệp, ...

2.3 Phân tích và dự báo lượng mưa tại trạm quan trắc Quy Nhơn.

Các nghiên cứu sử dụng ARIMA theo mùa để dự báo lượng mưa

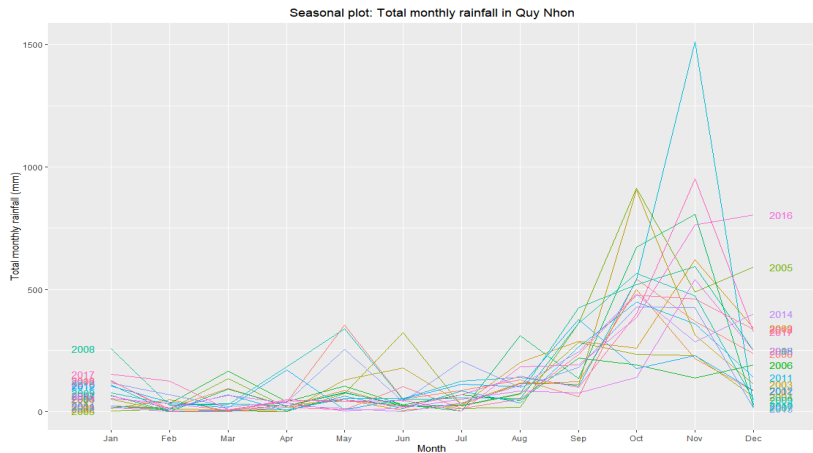
- Rahman và các cộng sự đã có một bài nghiên cứu đánh giá giữa 2 mô hình ARIMA và ANFIS để dự báo thời tiết cho thành phố Dhaka, kết quả cho thấy mô hình ARIMA thực hiện tốt hơn ANFIS [3].
- Dizon công bố kết quả nghiên cứu về ARIMA theo mùa là một mô hình rất tốt cho dự báo chuỗi thời gian có tính mùa vụ mạnh [4]. Momani sử dụng thành công mô hình ARIMA để dự báo xu hướng lượng mưa của Jordan [?].
- Tại Việt Nam, Nguyễn Hữu Quyền đã có một bài luận văn thạc sĩ khoa học về ứng dụng mô hình động thái ARIMAX để dự báo lượng mưa vụ đông xuân ở một số tỉnh vùng đồng bằng Bắc Bộ.

2.3 Phân tích và dự báo lượng mưa tại trạm quan trắc Quy Nhơn.

Chúng tôi lấy dữ liệu của trạm quan trắc Quy Nhơn từ *Trung tâm dữ liệu khí tượng thủy văn quốc gia* (<http://cmh.com.vn/>). Dữ liệu được thống kê từ tháng 1/2000 đến tháng 12/2018.

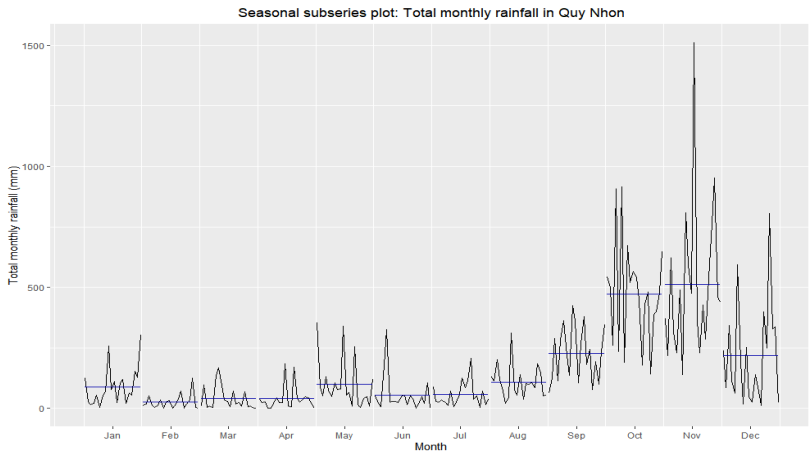
2.3 Phân tích và dự báo lượng mưa tại trạm quan trắc Quy Nhơn.

Đặc điểm lượng mưa tại trạm quan trắc Quy Nhơn

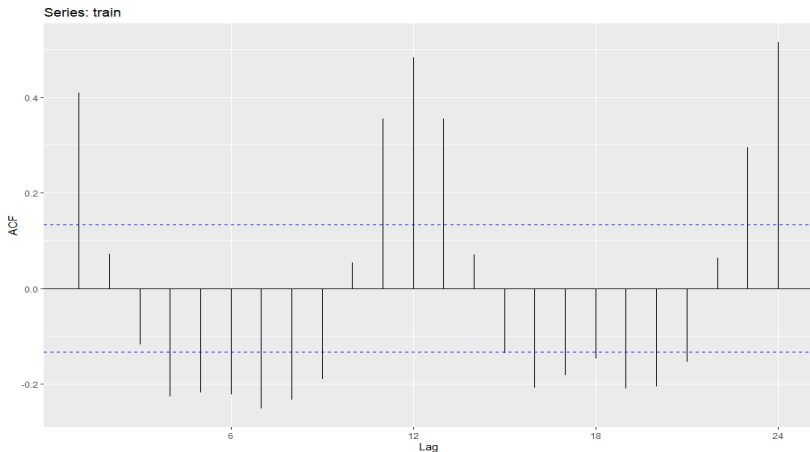


2.3 Phân tích và dự báo lượng mưa tại trạm quan trắc Quy Nhơn.

Bước 1: Kiểm tra tính dừng của tập "train"

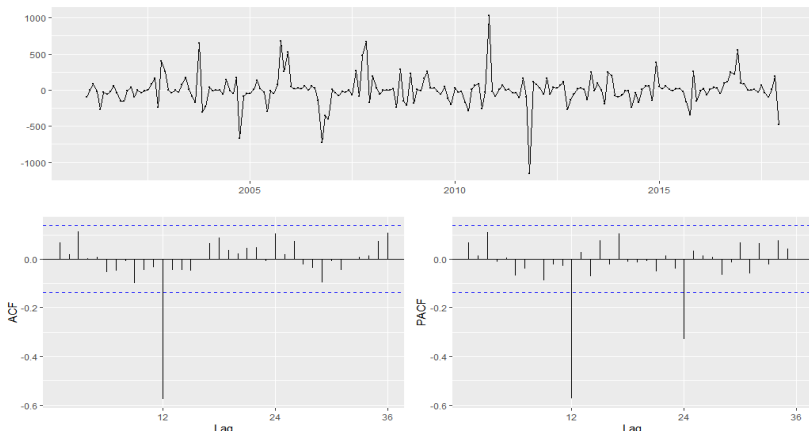


2.3 Phân tích và dự báo lượng mưa tại trạm quan trắc Quy Nhơn.



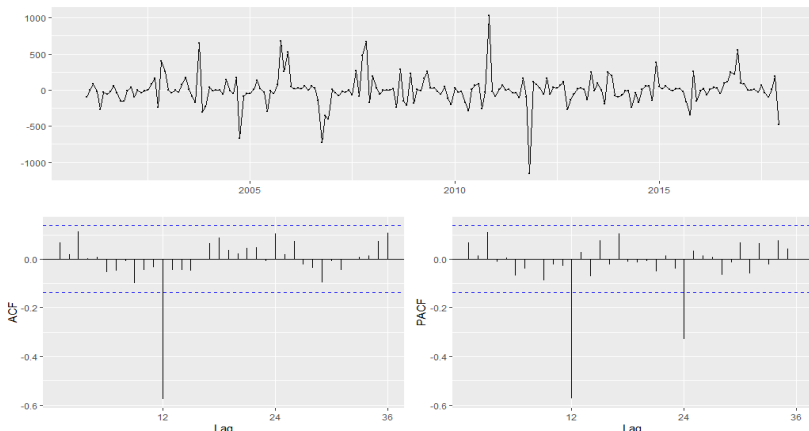
2.3 Phân tích và dự báo lượng mưa tại trạm quan trắc Quy Nhơn.

Bước 2: Chuyển chuỗi không có tính dừng thành chuỗi có tính dừng



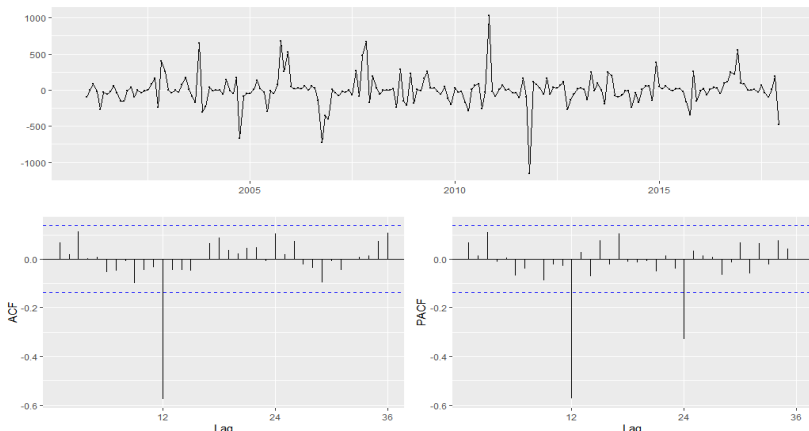
2.3 Phân tích và dự báo lượng mưa tại trạm quan trắc Quy Nhơn.

Bước 2: Chuyển chuỗi không có tính dừng thành chuỗi có tính dừng



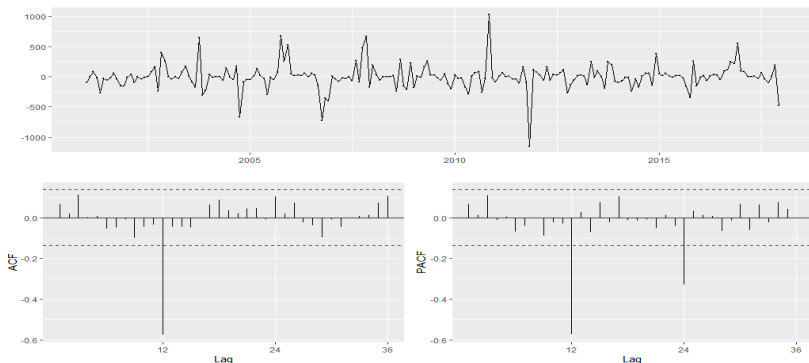
2.3 Phân tích và dự báo lượng mưa tại trạm quan trắc Quy Nhơn.

Bước 2: Chuyển chuỗi không có tính dừng thành chuỗi có tính dừng



2.3 Phân tích và dự báo lượng mưa tại trạm quan trắc Quy Nhơn.

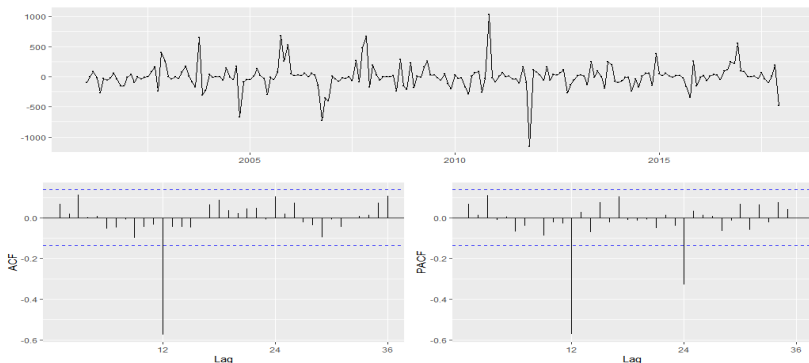
Bước 3: Chọn mô hình thích hợp từ đồ thị ACF và PACF



ACF \Rightarrow ARIMA(0,0,0)(0,1,1)₁₂ [AICc=2648.02]

2.3 Phân tích và dự báo lượng mưa tại trạm quan trắc Quy Nhơn.

Bước 3: Chọn mô hình thích hợp từ đồ thị ACF và PACF



ACF \Rightarrow ARIMA(0,0,0)(0,1,1)₁₂ [AICc=2648.02]

PACF \Rightarrow ARIMA(0,0,0)(0,1,2)₁₂ [AICc=2648.89]

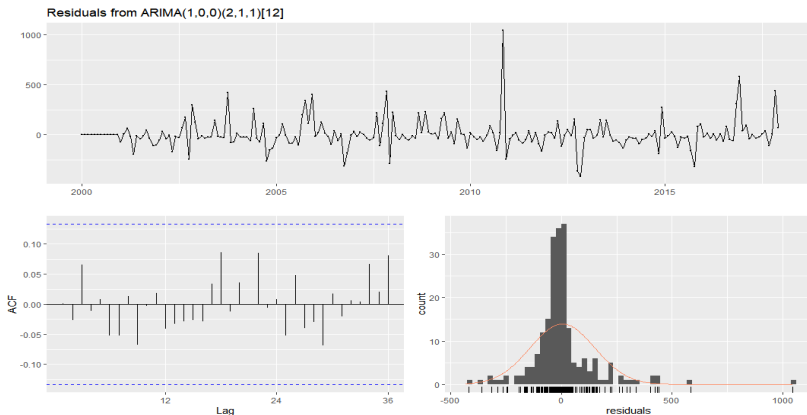
2.3 Phân tích và dự báo lượng mưa tại trạm quan trắc Quy Nhơn.

Bước 4: Khớp $ARIMA(0,0,0)(0,1,1)_{12}$ và chọn mô hình phù hợp nhất

STT	SARIMA Models	AICc	RMSE train
1	$ARIMA(0, 0, 0)(0, 1, 1)_{12}$	2650.64	146.5263
2	$ARIMA(1, 0, 0)(0, 1, 0)_{12}$	2649.27	146.3757
3	$ARIMA(0, 0, 1)(0, 1, 1)_{12}$	2649.26	146.3721
4	$ARIMA(0, 0, 0)(1, 1, 1)_{12}$	2648.60	146.8532
5	$ARIMA(0, 0, 0)(0, 1, 2)_{12}$	2648.89	146.8414
6	$ARIMA(1, 0, 0)(2, 1, 1)_{12}$	2648.02	145.6458
7	$ARIMA(1, 0, 1)(0, 1, 1)_{12}$	2651.02	146.3285
8	$ARIMA(0, 0, 0)(1, 1, 2)_{12}$	2650.42	146.7120

2.3 Phân tích và dự báo lượng mưa tại trạm quan trắc Quy Nhơn.

Bước 5: Kiểm tra phần dư từ mô hình $ARIMA(1,0,0)(2,1,1)_{12}$

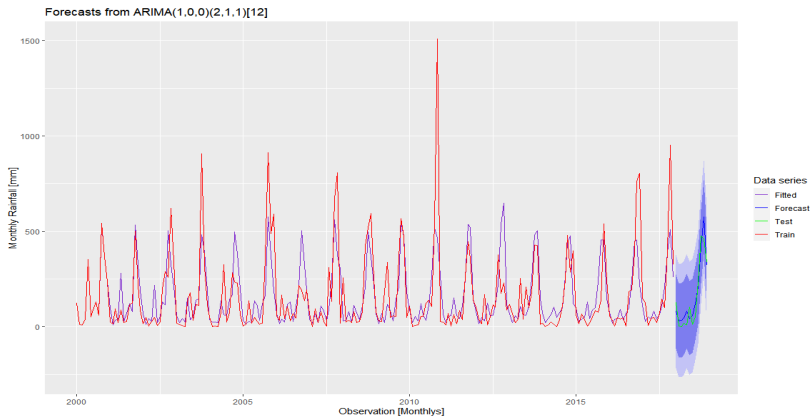


2.3 Phân tích và dự báo lượng mưa tại trạm quan trắc Quy Nhơn.

Bước 6: Đánh giá sai số từ mô hình $ARIMA(1,0,0)(2,1,1)_{12}$ và dự báo

	Point Actual(A)	Point Forecast (F)	Lo 80	Hi 80	Lo 95	Hi 95	$\frac{A - F}{A} \cdot 100\%$
Jan 2018	128.6	81.81485	-112.37714	276.0068	-215.17616	378.8059	36%
Feb 2018	2.8	32.32366	-162.30622	226.9535	-265.33706	329.9844	-1054%
Mar 2018	1.6	34.07819	-160.55367	228.7101	-263.58555	331.7419	-2029%
Apr 2018	20.0	46.95980	-147.67208	241.5917	-250.70397	344.6236	-134%
May 2018	9.4	82.40952	-112.22235	277.0414	-215.25424	380.0733	-776%
Jun 2018	103.7	44.63459	-149.99728	239.2665	-253.02917	342.2984	56%
Jul 2018	14.0	57.38414	-137.24773	252.0160	-240.27962	355.0479	-309%
Aug 2018	51.1	122.92861	-71.70327	317.5605	-174.73515	420.5924	-140%
Sep 2018	235.5	209.18078	14.54891	403.8127	-88.48298	506.8445	11%
Oct 2018	476.7	420.49951	225.86764	615.1314	122.83575	718.1633	11%
Nov 2018	462.0	574.92124	380.28937	769.5531	277.25748	872.5850	-24%
Dec 2018	337.9	321.89014	127.25849	516.5218	24.22672	619.5536	4%

2.3 Phân tích và dự báo lượng mưa tại trạm quan trắc Quy Nhơn.



2.3 Phân tích và dự báo lượng mưa tại trạm quan trắc Quy Nhơn.

Sự chênh lệch lớn lượng mưa hàng tháng ở mùa khô là do đâu?

2.3 Phân tích và dự báo lượng mưa tại trạm quan trắc Quy Nhơn.

Sự chênh lệch lớn lượng mưa hàng tháng ở mùa khô là do đâu?

- Một số quá trình xảy ra trong khoảng thời gian ngắn, như sự phát triển của hệ thống synop trong khí quyển là một trong những nguyên nhân dẫn đến sai số dự báo mùa.

2.3 Phân tích và dự báo lượng mưa tại trạm quan trắc Quy Nhơn.

Sự chênh lệch lớn lượng mưa hàng tháng ở mùa khô là do đâu?

- Một số quá trình xảy ra trong khoảng thời gian ngắn, như sự phát triển của hệ thống synop trong khí quyển là một trong những nguyên nhân dẫn đến sai số dự báo mùa.
- ENSO là nhân tố ảnh hưởng lớn nhất đến các dao động khí hậu hàng năm, chính sự kết hợp này là nguồn gốc chính sinh ra dị thường về nhiệt độ và lượng mưa trên phạm vi toàn cầu.

2.3 Phân tích và dự báo lượng mưa tại trạm quan trắc Quy Nhơn.

Bước 6: Đánh giá sai số từ mô hình $ARIMA(1,0,0)(2,1,1)_{12}$ và dự báo

	Point Actual(A)	Point Forecast (F)	Lo 80	Hi 80	Lo 95	Hi 95	$\frac{A - F}{A} \cdot 100\%$
Jan 2018	128.6	81.81485	-112.37714	276.0068	-215.17616	378.8059	36%
Feb 2018	2.8	32.32366	-162.30622	226.9535	-265.33706	329.9844	-1054%
Mar 2018	1.6	34.07819	-160.55367	228.7101	-263.58555	331.7419	-2029%
Apr 2018	20.0	46.95980	-147.67208	241.5917	-250.70397	344.6236	-134%
May 2018	9.4	82.40952	-112.22235	277.0414	-215.25424	380.0733	-776%
Jun 2018	103.7	44.63459	-149.99728	239.2665	-253.02917	342.2984	56%
Jul 2018	14.0	57.38414	-137.24773	252.0160	-240.27962	355.0479	-309%
Aug 2018	51.1	122.92861	-71.70327	317.5605	-174.73515	420.5924	-140%
Sep 2018	235.5	209.18078	14.54891	403.8127	-88.48298	506.8445	11%
Oct 2018	476.7	420.49951	225.86764	615.1314	122.83575	718.1633	11%
Nov 2018	462.0	574.92124	380.28937	769.5531	277.25748	872.5850	-24%
Dec 2018	337.9	321.89014	127.25849	516.5218	24.22672	619.5536	4%

2.3 Phân tích và dự báo lượng mưa tại trạm quan trắc Quy Nhơn.

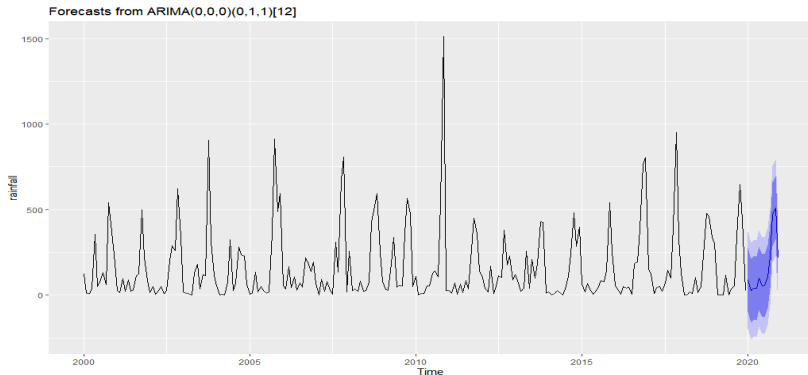


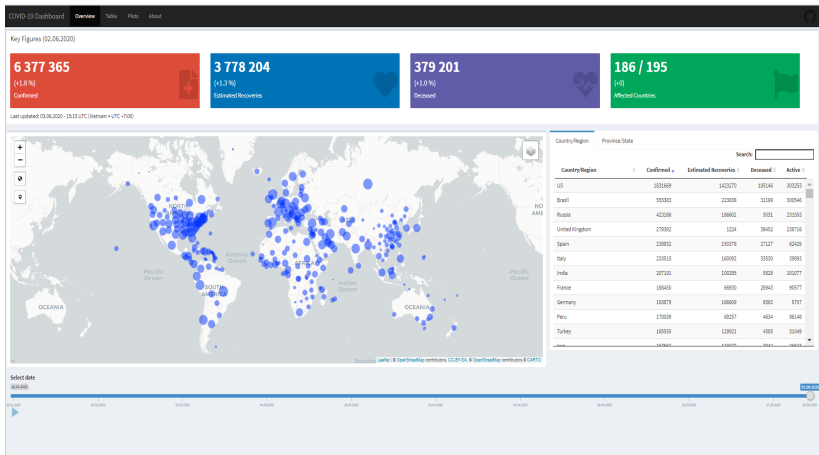
Figure: Đồ thị dự báo lượng mưa theo tháng cho năm 2019 và 2020

2.4 Website Dashboard COVID-19

`https://nguyenquocduong.shinyapps.io/NCKH/`

2.4 Website Dashboard COVID-19

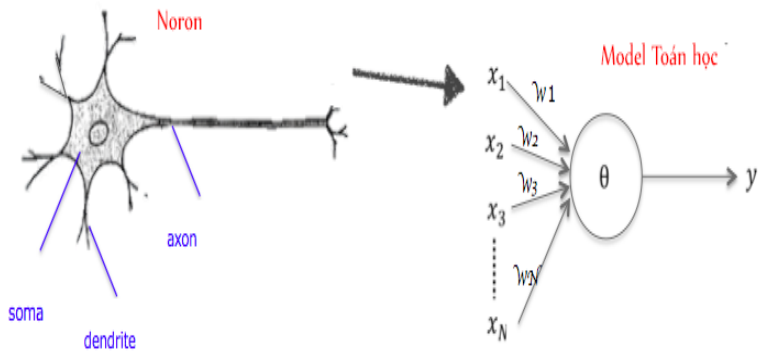
<https://nguyenquocduong.shinyapps.io/NCKH/>



2.5 So sánh mô hình ARIMA và NNAR

So sánh mô hình ARIMA và *mô hình mạng nơron tự hồi quy NNAR (Neural Network Auto-Regression)* trên tập dữ liệu lượng mưa hàng tháng tại trạm quan trắc Quy Nhơn dựa vào các thước đo sai số trên tập "test".

2.5 So sánh mô hình ARIMA và NNAR



2.5 So sánh mô hình ARIMA và NNAR

Hàm tổ hợp tuyến tính có dạng tổng quát

$$y_t = a_t + \sum_{i=1}^n w_{i,t} x_t.$$

2.5 So sánh mô hình ARIMA và NNAR

Trong tầng ẩn (hidden layer), khi đó y_t được điều chỉnh với việc sử dụng một hàm phi tuyến, chẳng hạn hàm sigmoid

$$f(y_t) = \frac{1}{1 + e^{-y_t}},$$

để nhận được giá trị đầu vào cho tầng kế tiếp.

2.5 So sánh mô hình ARIMA và NNAR

Đối với dữ liệu có tính mùa vụ, ta có mô hình $\text{NNAR}(p, P, k)_m$ tổng quát với các đầu vào là $(y_{t-1}, y_{t-2}, \dots, y_{t-p}, y_{t-m}, y_{t-2m}, y_{t-Pm})$ và k nơron tầng ẩn. Mô hình $\text{NNAR}(p, P, 0)_m$ là tương đương với mô hình $\text{ARIMA}(p, 0, 0)(P, 0, 0)_m$ nhưng không có các giới hạn về các tham số để đảm bảo tính dừng.

2.5 So sánh mô hình ARIMA và NNAR

```
#####  
# NNAR MODEL  
  
model.nnar <- nnetar(train)  
fit1 <- model.nnar %>% forecast(h = 12, PI = TRUE) %>%  
accuracy(rainfall)  
fit1[,c("RMSE", "MAE", "MASE")]  
  
## Model: NNAR(1,1,2)[12]  
  
## Average of 20 networks, each of which is a 2-2-1 network with 9 weights  
options were - linear output units, sigma^2 estimated as 26867  
  
##           RMSE      MAE      MASE  
## Training set 163.9116 105.6931 0.8598967  
## Test set     119.7967 103.0955 0.8387630
```

2.5 So sánh mô hình ARIMA và NNAR

```
# ARIMA MODEL

model.arima <- auto.arima(train)
fit2 <- model.arima %>% forecast(h = 12) %>%
accuracy(rainfall)
fit2[,c("RMSE", "MAE", "MASE")]

## ARIMA(1,0,0)(2,1,1)[12]

## Coefficients:
##          ar1      sar1      sar2      sma1
##      0.0672 -0.0526  0.1004 -0.8817
## s.e. 0.0682  0.0929  0.0876  0.0937

## sigma^2 estimated as 22910: log likelihood=-1320.17

##
##              RMSE              MAE              MASE
## Training set 145.64577    84.20232  0.6850522
## Test set      55.94754    49.54044  0.4030505
```

2.5 So sánh mô hình ARIMA và NNAR

NNAR MODEL

##		RMSE	MAE	MASE
## Training set		163.9116	105.6931	0.8598967
## Test set		119.7967	103.0955	0.8387630

ARIMA MODEL

##		RMSE	MAE	MASE
## Training set		145.64577	84.20232	0.6850522
## Test set		55.94754	49.54044	0.4030505

Những mặt thuận lợi khi sử dụng mô hình ARIMA

- ➊ ARIMA là một trong những mô hình tuyến tính phổ biến nhất trong dự báo chuỗi thời gian đã được áp dụng rộng rãi trong thập kỷ qua.
- ➋ ARIMA phát huy thế mạnh trong việc sử dụng để dự báo tài chính, chứng khoán, kinh tế lượng, khí tượng thủy văn, ...
- ➌ ARIMA thích hợp cho các bài toán dự báo ngắn hạn.
- ➍ Hơn nữa, ARIMA còn là nền tảng để xây dựng các mô hình lai phù hợp với từng loại dữ liệu cụ thể và cho kết quả chính xác hơn như ARIMA-ANN, ARIMA-LSTM, ARIMAX, ...

Những mặt hạn của mô hình ARIMA




- ➊ Trong ARIMA, một cấu trúc tương quan tuyến tính được giả định giữa các giá trị trong chuỗi thời gian. Do đó, ARIMA chỉ phát hiện được các khuôn mẫu tuyến tính có trong chuỗi dữ liệu còn khuôn mẫu không tuyến tính thì không được phát hiện.
- ➋ Các giá trị trong tương lai được dự báo phụ thuộc vào quá khứ nên đối với bài toán dự báo dài hạn, việc lựa chọn ARIMA là không phù hợp.






KẾT LUẬN VÀ KIẾN NGHỊ

Trong đề tài này, chúng tôi đã đạt được một số kết quả sau:

- ① Tìm hiểu lý thuyết căn bản, quan trọng về chuỗi thời gian, hồi quy cổ điển với chuỗi thời gian và mô hình ARIMA.
- ② Bước đầu đã nắm vững cách sử dụng các lệnh trong R để xây dựng mô hình ARIMA; hiểu và giải thích được kết quả đầu ra của mô hình ARIMA, cũng như nắm được kỹ năng về xử lý số liệu thô và kỹ năng vẽ hình, biểu đồ bằng phần mềm R.
- ③ Hơn nữa, chúng tôi đã xây dựng một website Dashboard COVID-19 nhằm theo dõi và kiểm soát tình hình dịch bệnh trên toàn cầu bằng phần mềm R.
- ④ Đồng thời, trong quá trình làm phần ví dụ thực hành với R, chúng tôi đã có điều kiện tìm hiểu thêm nhiều kiến thức mới ở các lĩnh vực khác nhau (ứng với từng ví dụ như: dịch tễ học, chăm sóc sức khỏe, khí tượng thủy văn, ...)

Tài liệu tham khảo

-  R. H. Shumway and D. S. Stoffer, *Time Series Analysis and Its Applications*, Springer Publisher, USA(2006).
-  G. E. P. Box, G. M. Jenkins, G. C. Reinsel and G. M. Liung, *Time Series Analysis: Forecasting and Control*, 5th edition, Publisher Wiley, Canada(2016).
-  R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 2nd edition, OTexts Publisher, USA(2018).
-  R. Krispin, *Hands-On Time Series Analysis with R*, Packt Publisher, UK(2019).

-  G. James, D. Witten, T. Hastie and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R, 1st edition*, Springer Texts in Statistics, USA(2007)
-  Nguyễn Văn Tuấn, *Phân tích dữ liệu với R*, NXB Tổng hợp T.P Hồ Chí Minh, 2014.
-  Nguyễn Chí Dũng, *Kinh tế lượng ứng dụng với R*, Ebook, 2017.
-  R. J. Hyndman, *Forecasting Functions for Time Series and Linear Models*, CRAN, 2020.
-  Nguyễn Duy Tiến và Vũ ngọc Yên, *Lý thuyết xác suất*, NXB Giáo dục, 2000.

-  Luz PM, Mendes BV, Codeco CT, Struchiner CJ and Galvani AP, *Time series analysis of dengue incidence in Rio de Janeiro*, , Am J Trop Med Hyg, Brazil(2008), 79 (6): 933-939.
-  Earnest A, Chen MI, Ng D and Leo YS, *Using autoregressive integrated moving average(ARIMA) models to predict and monitor the number of beds occupied during a SARS outbreak in a tertiary hospital in Singapore*, BMC Health Services Research, (2005), 5: 36-10.
-  Li XJ, Kang DM, Cao J and Wang JZ, *A time series model in incidence forecasting of hemorrhagic fever with renal syndrome*, Journal of Shandong University (Health Sciences), (2008), 46 (5): 547-549.







K. Mizumoto , K. Kagaya , A. Zarebski and G. Chowell,
*Estimating the asymptomatic proportion of
coronavirus disease 2019 (COVID-19) cases on
board the Diamond Princess cruise ship,
Yokohama, Japan, 2020*, Euro Surveill, 2020, 25(10).



C. Wang, L. Liu, X. Hao, H. Guo, Q. Wang, J. Huang, N.
He, H. Yu, X. Lin, A. Pan, S. Wei and T. Wu, *Evolving
Epidemiology and Impact of Non-pharmaceutical
Interventions on the Outbreak of Coronavirus
Disease 2019 in Wuhan, China*, The Preprint Server
for Healthy Sciences, 2020.



R. Li, S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang and J
Shaman, *Substantial undocumented infection
facilitates the rapid dissemination of novel
coronavirus (SARS-CoV2)*, Science, 2020.

-  S. Boccia, W. Ricciardi and J. P. A. Ioannidis, *What Other Countries Can Learn From Italy During the COVID-19 Pandemic*, JAMA Intern Med. Published online April 07, 2020.
-  E.M. Rasmusson and T.H. Carpenter, *The relationship between eastern equatorial Pacific SSTs and rainfall over India and Sri Lanka*, Mon. Wea. Rev, 1983, 517-528.
-  C.F. Ropelewski and M.S. Halpert, “*Global and Regional Scale Precipitation Patterns Associated with the El Niño/Southern Oscillation*”, Mon Wea Rev, 1987, 1606-1626.
-  Website:
<http://www.nguyenquocduong.github.com>.

KẾT THÚC BÁO CÁO

TRÂN TRỌNG CẢM ƠN!