ISSN 2354-0575

ỨNG DỤNG MÔ HÌNH CHUỖI THỜI GIAN DỰ BÁO BỆNH TRUYỀN NHIỄM TẠI TỈNH HƯNG YÊN

Lương Xuân Hồng[1], Phạm Thị Ánh Hương[1], Nguyễn Văn Chiến[2], Đàm Quang Thịnh[1], Trần Thị Thu Huyền[1], Nguyễn Văn Hậu[1,*]

[1] Trường Đại học Sư phạm Kỹ thuật Hưng Yên

[2] Trường Đại học Bách khoa Hà Nội

** Tác giả liên hệ: haunv@utehy.edu.vn*

Ngày tòa soạn nhận được bài báo: 05/12/2021
Ngày phản biện đánh giá và sửa chữa: 12/02/2022
Ngày bài báo được duyệt đăng: 05/03/2022

Tóm tắt:

Hưng Yên, một tỉnh nằm ở trung tâm đồng bằng sông Hồng, miền Bắc Việt Nam, bị ảnh hưởng nặng nề bởi các bệnh truyền nhiễm. Trong thời gian gần đây, rất nhiều các nhà khoa học đã nghiên cứu và áp dụng các mô hình chuỗi thời gian để dự báo tỉ lệ hay số các bệnh truyền nhiễm, để từ đó giúp bộ phận Y tế ngăn chặn sự lây lan của những đợt bùng phát dịch bệnh truyền nhiễm chết người (ví dụ: sốt xuất huyết, tiêu chảy, cúm, ...). Trong bài báo này, chúng tôi tìm hiểu và ứng dụng ba mô hình chuỗi thời gian dự báo bệnh truyền nhiễm tại Hưng Yên: mô hình tự hồi qui (Auto regression - AR), trung bình động (Moving average - MA), và mô hình tự hồi qui tích hợp trung bình động (Autoregressive Integrated Moving Average - ARIMA). Thực nghiệm cho thấy mô hình ARIMA cho kết quả tốt hơn.

Từ khóa: *Mô hình chuỗi thời gian, mô hình tự hồi qui, mô hình trung bình động, mô hình tự hồi qui tích hợp trung bình động, bệnh truyền nhiễm.*

1. Giới thiệu

Việt Nam nằm trong khu vực Đông Nam Á, có mức độ tiếp xúc với các hiểm họa liên quan đến khí hậu và các hiện tượng thời tiết cực đoan ở mức độ cao. Chỉ số rủi ro khí hậu toàn cầu 2020 xếp Việt Nam là quốc gia thứ sáu trên thế giới bị ảnh hưởng nặng nề nhất bởi biến đổi khí hậu và các hiện tượng thời tiết cực đoan trong giai đoạn 1999-2018 [3]. Biến đổi khí hậu được dự báo sẽ làm tăng nhiệt độ cũng như mức độ nghiêm trọng và tần suất của các hiện tượng thời tiết cực đoan, do đó sẽ làm tăng số người có nguy cơ mắc các bệnh nhạy cảm với khí hậu, bao gồm các bệnh truyền nhiễm.

Hưng Yên là một tỉnh nằm ở trung tâm đồng bằng sông Hồng, miền Bắc Việt Nam cũng là tỉnh bị ảnh hưởng nặng nề bởi các bệnh truyền nhiễm. Trong giai đoạn từ năm 1997 đến 2016 2.916 ca mắc sốt xuất huyết (trung bình khoảng 145 ca mỗi năm). Số liệu do Viện Vệ sinh Dịch tễ Trung ương (NIHE), đơn vị chịu trách nhiệm độ chính xác của thông tin trong cơ sở dữ liệu, cung cấp.

Sốt xuất huyết (dengue) là bệnh truyền nhiễm gây dịch do vi rút Dengue gây nên. Vi rút Dengue có 4 týp huyết thanh là DEN-1, DEN-2, DEN-3 và DEN-4. Vi rút truyền từ người bệnh sang người lành do muỗi đốt. Muỗi Aedes aegypti là côn trùng trung gian truyền bệnh chủ yếu. [2]

Bệnh xảy ra quanh năm, thường gia tăng vào mùa mưa. Bệnh gặp ở cả trẻ em và người lớn. Đặc điểm của sốt xuất huyết Dengue là sốt, xuất huyết và thoát huyết tương, có thể dẫn đến sốc giảm thể tích tuần hoàn, rối loạn đông máu, suy tạng, nếu không được chẩn đoán sớm và xử trí kịp thời dễ dẫn đến tử vong. [2]

Gánh nặng của bệnh truyền nhiễm đã gây rất nhiều khó khăn cho hệ thống y tế tỉnh Hưng Yên. Do đó, việc xây dựng một chương trình dự báo bệnh truyền nhiễm tại tỉnh Hưng Yên để giúp cung cấp thông tin cho y tế cộng đồng phòng chống bùng phát dịch bệnh đã trở thành một vấn đề vô cùng cấp thiết. [1].

Các phương pháp Máy học có thể được tận dụng để ngăn chặn sự lây lan của những đợt bùng phát dịch bệnh truyền nhiễm chết người (ví dụ: sốt xuất huyết, tiêu chảy, cúm, ...). Điều này có thể được thực hiện bằng cách áp dụng các thuật toán máy học trong việc dự đoán và phát hiện bệnh truyền nhiễm.

Để dự báo sự bùng phát dịch bệnh [4], [5], [6], các thuật toán máy học có thể được sử dụng để tìm hiểu các tập dữ liệu bao gồm thông tin về các loại virus đã biết, quần thể động vật, nhân khẩu học, sinh học và thông tin đa dạng sinh học, cơ sở hạ tầng vật lý sẵn có, các yếu tố khí hậu, cũng như xác định vị trí địa lý của các dịch bệnh để dự đoán bất kỳ đợt bùng phát nào. Ví dụ: dự báo bùng phát bệnh sốt rét có thể được thực hiện bằng cách sử dụng mô hình Máy vectơ hỗ trợ (SVM) và Mạng thần kinh nhân tạo (ANN); sử dụng Lượng mưa trung bình hàng tháng, Nhiệt độ, Độ ẩm, Tổng số ca mắc để đánh giá hiệu suất của các mô hình [4].

Các quan chức y tế cộng đồng cũng có thể sử dụng dữ liệu của Hệ thống Thông tin Địa lý (GIS) và các phương pháp phân tích không gian để thu thập thông tin hoặc chủ động dự đoán để ngăn chặn các đợt bùng phát trong tương lai [7]. Công nghệ thông tin địa lý có thể được sử dụng để trích xuất vị trí không gian của các ca bệnh và khám phá những

thay đổi không gian và thời gian của dịch bệnh cũng như mối quan hệ không gian của nó với các đối tượng khác được lưu trữ trong GIS [5].

Trong bài báo này chúng tôi tập trung nghiên cứu một số mô hình dự báo chuỗi thời gian.

2. Mô hình chuỗi thời gian

2.1 Dữ liệu chuỗi thời gian

Trong toán học, dữ liệu chuỗi thời gian được định nghĩa là những điểm dữ liệu đã được đánh chỉ số theo thời gian và có khoảng cách đều nhau giữa những quan sát liên tiếp. Đó có thể là dữ liệu về mức tiêu thụ điện năng theo giờ, giá chứng khoán hàng ngày, dự báo thời tiết hàng tháng, tổng thu nhập quốc dân của một quốc gia hàng năm.

Ưu điểm của chuỗi thời gian là nó có thể lưu trữ được trạng thái của một trường dữ liệu theo thời gian. Ta đã biết thế giới luôn vận động, các sự vật, hiện tượng hiếm khi dừng lại ở trạng thái tĩnh mà thường thay đổi. Do đó dữ liệu chuỗi thời gian có tính ứng dụng rất cao và được áp dụng trong rất nhiều lĩnh vực khác nhau như: thống kê, kinh tế lượng, toán tài chính, dự báo thời tiết, dự báo bệnh tật, ... Chính vì thế dữ liệu chuỗi thời gian đóng một vai trò cực kỳ quan trọng đối với sự phát triển của nhân loại. Dữ liệu chuỗi thời gian ở khắp mọi nơi, trong: tài chính, thương mại điện tử, kinh doanh, khoa học xã hội, y tế, v.v.

Tính xu hướng (trend) là yếu tố thể hiện xu hướng thay đổi của dữ liệu theo thời gian. Đây là đặc trưng thường thấy của rất nhiều dữ liệu chuỗi thời gian. Ví dụ như giá cả thị trường tăng do ảnh hưởng của lạm phát, dân số thế giới tăng qua các năm, nhiệt độ trung bình trái đất tăng theo thời gian do hiệu ứng nhà kính. Tính xu hướng ảnh hưởng không nhỏ tới việc đưa ra nhận định về mối quan hệ tương quan giữa các chuỗi số. Tức là về bản chất các chuỗi không tương quan nhưng do chúng cùng có chung xu hướng theo thời gian nên chúng ta nhận định chúng là tương quan. Ví dụ: số ca mắc sốt xuất huyết hàng năm và các yếu tố thời tiết như nhiệt độ, độ ẩm, lượng mưa có mối quan hệ cùng chiều (hay còn gọi là tương quan thuận). Không khó để chúng ta nhận định được bản chất của sự tương quan này là do chúng có cùng sự tương quan với sự phát triển của muỗi vằn, tác nhân trung gian truyền bệnh sốt xuất huyết. Vì muỗi vằn thường phát triển mạnh nhất vào mùa mưa, thời tiết nóng ẩm, khi nhiệt độ trung bình hàng tháng vượt trên 20 °C.

Tính chu kỳ (*seasonality*) là qui luật có tính chất lặp lại của dữ liệu theo thời gian. Sự thay đổi thời tiết, sự phát triển của loài người hay sự bùng phát của dịch bệnh đều bị ảnh hưởng của chu kỳ và lặp lại theo thời gian. Chính vì thế tìm ra được yếu tố chu kỳ sẽ giúp ích cho việc dự báo chính xác hơn. Một ví dụ về tầm quan trọng của chu kỳ đó là hệ thống y tế cộng đồng phải nắm được dịch bệnh sẽ xuất hiện vào thời điểm nào trong năm? Như sốt xuất huyết thường xuất hiện vào khoảng thời gian giao mùa, vào mùa mưa từ tháng 7 đến tháng 11 hàng năm. Từ đó sẽ có kế hoạch chuẩn bị cơ sở vật chất và điều động nguồn lực con người để phòng chống dịch hiệu quả. Nếu không hiểu được tính chu kỳ của chuỗi thời gian, hệ thống y tế cộng đồng có thể dự báo sai thời điểm bùng phát dịch bệnh dẫn đến bị động trong công tác phòng chống dịch, gây hậu quả nghiêm trọng về người và của.

Chúng ta có thể kể ra:

- Các hiện tượng tự nhiên: biến động thời tiết
- Các hoạt động kinh doanh: bắt đầu hoặc kết thúc năm tài chính
- Hành vi xã hội và văn hóa: ngày lễ hoặc các hoạt động tôn giáo

Tính dừng (*Stationarity*)

Chuỗi thời gian dừng là chuỗi mà các thuộc tính của nó không phụ thuộc vào thời gian mà chuỗi đó được quan sát. Do đó, chuỗi thời gian có xu hướng (trend) hoặc có tính chu kỳ (seasonal), không dừng sẽ ảnh hưởng đến giá trị của chuỗi thời gian tại thời gian khác nhau. Mặt khác, chúng ta coi sai số (white noise) là dừng - không quan trọng khi bạn quan sát nó, nó sẽ trông giống nhau tại bất kỳ thời điểm nào.

Nói chung, một chuỗi thời gian có tính dừng sẽ không có mô hình dự đoán được trong dài hạn. Biểu đồ thời gian sẽ cho thấy chuỗi gần như nằm ngang (mặc dù có thể xảy ra một số hành vi theo chu kỳ), với phương sai không đổi.

Tính dừng là một đặc tính quan trọng của chuỗi thời gian. Chuỗi thời gian được cho là ổn định nếu giá trị trung bình, phương sai và hiệp phương sai của nó không thay đổi đáng kể theo thời gian. Chuỗi thời gian ổn định rất dễ dự đoán vì chúng ta có thể giả định rằng các thuộc tính thống kê trong tương lai giống hoặc tỷ lệ với các thuộc tính thống kê hiện tại.

Kiểm tra tính dừng

Giả thuyết không (Null hypothesis) là một loại phỏng đoán được sử dụng trong thống kê cho rằng không có ý nghĩa thống kê tồn tại trong tập hợp các quan sát nhất định. Nó giả định rằng bất kì sự khác biệt hay ý nghĩa nào bạn quan sát được trong một tập hợp dữ liệu là do sự ngẫu nhiên. Giả thuyết không được cho là đúng cho đến khi có bằng chứng thống kê bác bỏ nó với một giả thuyết thay thế khác. Đối lập với giả thuyết không là giả thuyết thay thế (Alternative hypothesis).

Phép kiểm tra Dickey-Fuller tăng cường (ADF) kiểm tra giả thuyết Không (null hypothesis) rằng một gốc đơn vị có trong một mẫu chuỗi thời gian. Giả thuyết thay thế khác nhau tùy thuộc vào phiên bản thử nghiệm được sử dụng, nhưng thường là cố định hoặc ổn định theo xu hướng.

Quay trở lại bài toán này, chúng ta chứng minh:

- Giả thuyết không - H0 - cho rằng tập giá trị chuỗi thời gian có unit root dẫn tới tính không dừng (non-stationary)
- hay bác bỏ H0, tức là đi với Giả thuyết thay

thế (rằng chuỗi thời gian không có unit root dẫn đến tính dừng (stationary).

Cuối cùng, chúng ta quyết định điều này dựa trên giá trị p:

• Giá trị p nhỏ (thường là ≤ 0,05) chỉ ra bằng chứng mạnh mẽ chống lại giả thuyết H0, vì vậy bạn bác bỏ giả thuyết H0.

• Giá trị p lớn (> 0,05) cho thấy bằng chứng yếu chống lại giả thuyết H0, vì vậy bạn không thể bác bỏ giả thuyết H0.

2.1. Mô hình tự hồi qui (Auto Regression - AR)

Mô hình hình tự hồi qui (AR) dựa trên trực giác rằng quá khứ dự đoán tương lai và do đó giả định một quy trình chuỗi thời gian trong đó giá trị tại một thời điểm trong thời gian t là một hàm của chuỗi các giá trị tại các điểm trong thời gian trước đó. Sau đây, đề tài sẽ thảo luận về mô hình này sẽ chi tiết hơn để cho bạn biết cách các nhà thống kê xem xét những đặc trưng của mô hình đó. Vì lý do này, chúng tôi bắt đầu với một tổng quan lý thuyết.

Tự hồi qui giống như những nỗ lực đầu tiên của chúng ta để điều chỉnh một chuỗi thời gian, đặc biệt nếu nó không có thông tin nào khác ngoài chính chuỗi thời gian đó. Nó chính xác như những gì tên của nó ngụ ý: một hồi quy về các giá trị trong quá khứ để dự đoán các giá trị trong tương lai.

Mô hình AR đơn giản nhất, mô hình AR (1), mô tả một hệ thống như sau:

$$y_t = b_0 + b_1 \times y_{t-1} + e_t \quad (1)$$

Giá trị của chuỗi tại thời điểm t là một hàm của hằng số b_0, giá trị của nó ở bước thời gian trước đó (t-1) nhân với một hằng số khác $b_1 \times y_{t-1}$ và một số hạng– thường gọi là lỗi/độ nhiễu (trắng) - cũng thay đổi theo thời gian e_t. Độ nhiễu được giả định là có phương sai không đổi và giá trị trung bình bằng 0. Chúng ta biểu thị một thuật ngữ tự hồi qui chỉ quay lại thời điểm trước đó ngay lập tức dưới dạng mô hình AR(1) vì nó bao gồm việc xem lại một lần trễ (one lag).

Một cách tình cờ, mô hình AR(1) có dạng giống hệt với một mô hình hồi quy tuyến tính đơn giản với chỉ một biến giải thích. Đó là, nó ánh xạ tới:

$$Y = b_0 + b_1 \times x + e$$

Chúng ta có thể tính toán cả hai giá trị, y_t và phương sai của nó, nếu chúng ta biết b_0 và b_1:

$$E(y_t \mid y_{t-1}) = b_0 + b_1 \times y_{t-1} + e_t \quad (2)$$
$$Var(y_t \mid y_{t-1}) = Var(e_t) = Var(e)$$

Chúng ta có thể tổng quát của ký hiệu này cho phép giá trị hiện tại của một quá trình AR phụ thuộc vào p giá trị gần đây nhất, tạo ra một quá trình AR(p).

Bây giờ chúng tôi chuyển sang ký hiệu truyền thống hơn, sử dụng ϕ để biểu thị các hệ số tự hồi quy:

$$y_t = \phi_0 + \phi_1 \times y_{t-1} + \phi_2 \times y_{t-2} + ... + \phi_p \times y_{t-p} + e_t$$

Và khi biểu diễn AR(1) sẽ trở thành:

$$y_t = \phi_0 + \phi_1 \times y_{t-1} + e_t \quad (3)$$

Chúng ta giả sử quá trình này là tính dừng (stationary) và sau đó hoạt động "quay ngược" để xem điều đó ngụ ý gì về các hệ số. Đầu tiên, từ giả định về tính dừng, chúng ta biết rằng giá trị kỳ vọng của quá trình phải luôn như nhau tại mọi thời điểm.

Chúng ta có thể viết lại y_t cho mỗi phương trình cho một quy trình AR(1):

$$E(y_t) = \mu = E(y_{t-1})$$

Theo định nghĩa $E(e_t) = 0$. Ngoài ra, các giá trị ϕ là các hằng số, nên giá trị trung bình của nó cũng là các hằng số. Phương trình (3) có thể suy ra:

$$E(y_t) = E(\phi_0 + \phi_1 \times y_{t-1} + e_t)$$

Suy ra: $\mu = \phi_0 + \phi_1 \times \mu + 0$

Điều này đồng nghĩa với: $\mu = \dfrac{\phi_0}{1 - \phi_1}$ (4)

Chúng ta đã tìm được sự liên hệ giữa giá trị trung bình của mô hình tại thời điểm (t) và các hệ số của nó.

Từ phương trình (4) chúng ta cũng suy ra được

$$\phi_0 = \mu \times (1 - \phi_1) \quad (5)$$

Thay (5) vào phương trình (3) ta được:

$y_t = \phi_0 + \phi_1 \times y_{t-1} + e_t$

$y_t = (\mu - \mu \times \phi_1). + \phi_1 \times y_{t-1} + e_t$

$y_t - \mu = \phi_1 (y_{t-1} - \mu) + e_t$

Hoàn toàn tương tự, chúng ta cũng có được phương trình sau:

$$y_{t-1} - \mu = \phi_1 (y_{t-2} - \mu) + e_{t-1} \quad (6)$$

Thay (6) vào (5) ta được:

$$y_t - \mu = \phi_1(\phi_1 (y_{t-2} - \mu) + e_{t-1}) + e_t$$

Sắp xếp lại:

$$y_t - \mu = e_t + \phi_1(e_{t-1} + \phi_1 (y_{t-2} - \mu)) \quad (7)$$

Không có gì khó khăn cho chúng ta tiếp tục công thức:

$$y_t - \mu = e_t + \phi_1(e_{t-1} + \phi_1(e_{t-2} + \phi_1 (y_{t-3} - \mu)))$$
$$= e_t + \phi_1 e_{t-1} + \phi_1^2 e_{t-2} + \phi_1^3 e_{t-3} + ...$$

Và chúng ta đi tới một kết luận tổng quát là:

$$y_t - \mu = \sum_{t=0}^{\infty} \phi_1^i e_{t-i} \quad (8)$$

Như vậy là hiệu của y_t và trung bình giá trị của mô hình là một hàm tuyến tính của các nhiễu (error terms hoặc white noises).

Kết quả này có thể dẫn tới $E[(y_t - \mu) \times e_{t+k}]=0$ vì điều kiện là các giá trị của e_t tại các thời điểm t khác nhau là độc lập.

Hơn nữa, ta có từ phương trình:

$$y_t - \mu = \phi_1 (y_{t-1} - \mu) + e_t$$
$$var(y_t) = \phi_1^2 var(y_{t-1}) + var(e_t)$$

Vì chúng ta giả định mô hình có tính dừng ($var(y_t) = var(y_{t-1})$ nên từ phương trình trên ta lại có

$$var(y_t) = \phi_1^2 var(y_t) + var(e_t)$$
$$var(y_t) = \frac{var(e_t)}{1 - \phi_1^2}$$

Vì phương sai luôn lớn hơn hoặc bằng 0 nên $\phi_1^2 < 1$ để đảm bảo $var(y_t) > 0$. Điều này gợi ý cho chúng ta rằng quá trình dừng của mô hình sẽ cho chúng ta kết quả $-1 < \phi_1 < 1$. Đây cũng chính là điều kiện cần và đủ cho một mô hình có tính dừng yếu (weak staionarity).

2.3. Mô hình trung bình động (Moving Average - MA)

Mô hình trung bình động (Moving Average -

MA) dựa trên một quá trình tổng thể trong đó giá trị tại mỗi thời điểm là một hàm giá trị của các "nhiễu" trong quá khứ gần nhất, các nhiễu này độc lập với với nhau. Chúng tôi sẽ xem xét mô hình này cùng một loạt các bước mà chúng ta đã tìm hiểu mô hình AR.

Mô hình trung bình động có thể được biểu diễn tương tự như mô hình tự hồi quy (AR), ngoại trừ việc các phần tử trong phương trình tuyến tính là các "nhiễu" hiện tại và quá khứ hơn là các giá trị hiện tại và quá khứ của chính mô hình. Vì vậy, một mô hình MA bậc q được biểu diễn dưới dạng:

$$y_t = \mu + e_t + \theta_1 \times e_{t-1} + \theta_2 \times e_{t-2} + ... + \theta_p \times e_{t-q}$$

Các nhà kinh tế học nói về các "nhiễu" này như là "cú sốc" đối với hệ thống, trong khi một chuyên gia về kỹ thuật điện có thể nói về điều này như một chuỗi các xung và bản thân mô hình như một bộ lọc đáp ứng xung hữu hạn, có nghĩa là các tác động của bất kỳ xung cụ thể nào chỉ tồn tại trong một khoảng thời gian hữu hạn. Từ ngữ không quan trọng, nhưng khái niệm về nhiều sự kiện độc lập tại các thời điểm khác nhau trong quá khứ ảnh hưởng đến giá trị hiện tại của quá trình, mỗi người đều có đóng góp riêng, là ý tưởng chính.

Các mô hình MA, theo định nghĩa, là mô hình dừng yếu (weakly stationary) mà không cần thêm bất kỳ ràng buộc nào đối với các tham số. Điều này là do giá trị trung bình và phương sai của một quá trình MA đều hữu hạn và bất biến theo thời gian vì các điều khoản sai số được giả định là độc lập và phân phối giống nhau với giá trị trung bình bằng 0. Chúng ta có thể kiểm định:

$$E(y_t = \mu + e_t + \theta_1 \times e_{t-1} + \theta_2 \times e_{t-2} + ... + \theta_p \times e_{t-q}$$
$$= E(\mu) + \theta_1 \times 0 + \theta_2 \times 0 + ... + \theta_p \times 0) = \mu$$

2.4. Mô hình tự hồi qui tích hợp trung bình động (Autoregressive Integrated Moving Average - ARIMA)

Chúng ta đã thảo luận các mô hình AR và MA riêng lẻ, bây giờ chúng ta xem xét mô hình Trung bình động tích hợp tự hồi qui (ARIMA), mô hình chuỗi thời gian này kết hợp cả 2 mô hình AR và MA. Cảm nhận sẽ dẫn chúng ta đến một mô hình ARMA, nhưng chúng ta lại tìm hiểu mô hình mở rộng sang ARIMA, mô hình này bao hàm cho sự khác biệt (differencing), một cách loại bỏ các xu hướng (trends) và biện giải cho một chuỗi thời gian dừng. Sự khác biệt (differencing) là chuyển đổi các giá trị của một chuỗi thời gian thành các thay đổi của các giá trị theo một chuỗi thời gian. Thông thường, điều này được thực hiện bằng cách tính toán sự khác biệt theo từng cặp của các điểm liền kề trong thời gian, sao cho giá trị của chuỗi sai phân tại thời điểm t là giá trị tại thời điểm t trừ đi giá trị tại thời điểm t - 1. Tuy nhiên, việc sai lệch cũng có thể được thực hiện trên độ trễ (lag windows) khác nhau cho thuận tiện.

Các mô hình ARIMA đang là một trong số những mô hình cho kết quả tốt nhất, đặc biệt trong các trường hợp tập dữ liệu nhỏ, nơi mà tập dữ liệu không phù hợp cho nhiều mô hình học máy và học sâu (deep learning). Tuy nhiên, ngay cả các mô hình ARIMA cũng có nguy cơ bị vấn đề quá khớp (overfitting).

Bạn có thể đang vò đầu bứt tai tại thời điểm này nếu bạn đang chú ý vì chúng ta vừa đưa cùng một dữ liệu cho cả quy trình AR và MA mà không nhận xét về nó. Đây là một thói quen khó chịu mà đôi khi bạn có thể mắc phải trong sách giáo khoa phân tích chuỗi thời gian. Một số tác giả sẽ đối phó với sự lười biếng dữ liệu này, trong khi những người khác sẽ bỏ qua nó một cách nhạt nhẽo. Chúng ta chưa tìm hiểu kĩ xem một trong hai mô hình trước đây có phải là mô hình thực sự tốt hay không, nhưng có vẻ như từ quá trình điều chỉnh mà chúng ta đã sử dụng rằng có những lập luận có thể bảo vệ được để mô tả dữ liệu bằng mô hình AR hoặc MA. Điều này đặt ra câu hỏi: có hữu ích không nếu kết hợp cả hai các hành vi vào cùng một mô hình?

Intergrated: là quá trình lấy sai phân. Yêu cầu chung của các thuật toán trong chuỗi thời gian là chuỗi phải đảm bảo tính dừng (stationarity). Hầu hết các chuỗi đều tăng hoặc giảm theo thời gian. Do đó yếu tố tương quan giữa chúng chưa chắc là thực sự mà là do chúng cùng tương quan theo thời gian. Khi biến đổi sang chuỗi dừng, các nhân tố ảnh hưởng thời gian được loại bỏ và chuỗi sẽ dễ dự báo hơn. Để tạo thành chuỗi dừng, một phương pháp đơn giản nhất là lấy sai phân. Bậc của sai phân để tạo thành chuỗi dừng còn gọi là bậc của quá trình đồng tích hợp (order of intergration). Qúa trình sai phân bậc d của chuỗi được thực hiện như sau:

- Sai phân bậc 1: $I(1) = \Delta(y_t) = y_t - y_{t-1}$
- Sai phân bậc d: $I(d) = \Delta^d(y_t) = \Delta(\Delta(...\Delta(y_t)))$

Mô hình tự hồi qui tích hợp với trung bình trượt (ARIMA) dựa trên giả thuyết chuỗi dừng và phương sai của cá "nhiễu" (white noise) không đổi. Mô hình sử dụng đầu vào chính là những tín hiệu quá khứ của chuỗi được dự báo để dự báo nó. Các tín hiệu đó bao gồm: chuỗi tự hồi qui AR (auto regression) và chuỗi trung bình trượt MA (moving average). Hầu hết các chuỗi thời gian sẽ có xu hướng tăng hoặc giảm theo thời gian, do đó yếu tố chuỗi dừng thường không đạt được. Trong trường hợp chuỗi không dừng thì ta sẽ cần biến đổi sang chuỗi dừng bằng sai phân. Khi đó tham số đặc trưng của mô hình sẽ có thêm thành phần bậc của sai phân d và mô hình được đặc tả bởi 3 tham số ARIMA(p, d, q).

Phương trình tổng quát của ARIMA là:

$$y_t = \phi_0 + \phi_1 \times y_{t-1} + \phi_2 \times y_{t-2} \dots \phi_p \times y_{t-p}$$
$$+ \theta_1 \times e_{t-1} + \theta_2 \times e_{t-2} + ... + \theta_p \times e_{t-q}$$

Như vậy, ARIMA là một mô hình hồi quy tuyến tính với p giá trị trước và q giá trị sai số là các đặc tính (features) và các giá trị ϕ_i và θ_j là các hệ số hồi qui. Như vậy mô hình ARIMA(p,d,q) gồm:

- p,d,q là các siêu tham số (hyper-parameters)
- ARIMA(p,d,q) là mô hình hồi qui tuyến tính

có p giá trị quá khứ và q sai số với sai phân bậc d.

2.5. Lựa chọn tham số

Dữ liệu có thể tuân theo mô hình ARIMA (p, d, 0) nếu các đồ thị ACF và PACF của dữ liệu đã phân biệt hiển thị các mẫu sau:

• ACF đang phân rã theo hàm mũ hoặc hình sin;

• Có một đỉnh khác biệt ở độ trễ p trong PACF, nhưng không có gì vượt quá độ trễ p.

Dữ liệu có thể tuân theo mô hình ARIMA (0, d, q) nếu các đồ thị ACF và PACF của dữ liệu đã phân biệt hiển thị các mẫu sau:

• PACF đang phân rã theo hàm mũ hoặc hình sin;

• Có một đỉnh khác biệt ở độ trễ q trong ACF, nhưng không có gì vượt quá độ trễ q.

Cách xác định sai phân d trong mô hình

Bậc sai phân (differencing) phù hợp là độ sai phân tối thiểu cần thiết để có được một chuỗi gần dừng biến đổi xung quanh một giá trị trung bình xác định và đồ thị ACF đạt đến 0 khá nhanh.

Nếu tự tương quan là dương với nhiều số độ trễ (10 hoặc nhiều hơn), thì chuỗi đó cần phải phân biệt thêm. Mặt khác, nếu bản thân tự tương quan trễ bằng 1 quá âm, thì chuỗi có thể bị sai lệch quá mức rồi.

Tóm lại, chúng ta có thể hiểu rằng: Nếu mô hình chưa dừng, ta sẽ thực hiện sai phân (differencing) và kiểm tra tính dừng. Nếu nó dừng, hãy lấy biểu đồ tương quan và phù hợp với mô hình ARMA (p, q) đến sự khác biệt trong đó p là điểm cắt đối với PACF và q là điểm cắt đối với ACF. Đây là mô hình ARIMA (p, 1, q) cho dữ liệu ban đầu. Tuy nhiên, nếu sau sai phân (differencing) vẫn không dừng, hãy lấy thêm sai phân (differencing) và để quá trình tiếp tục.

3. Dữ liệu và Thực nghiệm

3.1. Dữ liệu

Trong đề tài này, chúng tôi sử dụng dữ liệu về bệnh cúm ở tỉnh Hưng Yên được thu thập từ năm 1997 đến năm 2016 với hai trường quan trọng nhất là Influenza_cases (Số ca mắc bệnh cúm) và Influenza_rates (Tỷ lệ mắc cúm trên 10.000 dân). Số liệu được thu thập vào ngày mồng 1 hàng tháng do Viện Vệ sinh Dịch tễ Trung ương (NIHE), đơn vị chịu trách nhiệm độ chính xác của thông tin trong cơ sở dữ liệu, cung cấp.

Ngoài ra 12 yếu tố khí hậu mà chúng tôi tin rằng có liên quan mật thiết đến việc lây truyền bệnh cúm cũng đã được thu thập bao gồm:

• Total_Evaporation: Tổng lượng nước bốc hơi;

• Total_Rainfall: Tổng lượng mưa trong tháng;

• Max_Daily_Rainfall: Lượng mưa hàng ngày lớn nhất;

• n_rainigdays: Số ngày mưa trong tháng;

• Average_temperature: Nhiệt độ trung bình;

• Max_Average_Temperature: Nhiệt độ trung bình lớn nhất;

• Min_Average_Temperature: Nhiệt độ trung bình nhỏ nhất;

• Max_Absolute_Temperature: Nhiệt độ lớn nhất tuyệt đối;

• Min_Asolute_Temperature: Nhiệt độ nhỏ nhất tuyệt đối;

• Average_Humidity: Độ ẩm trung bình;

• Min_Humudity: Độ ẩm nhỏ nhất;

• n_hours_sunshine: Số giờ nắng.

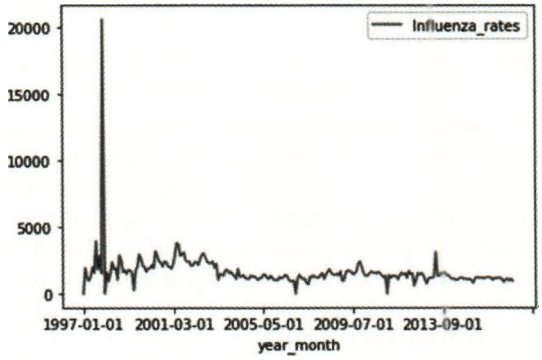
Dữ liệu do Viện Khoa học Khí Tượng Thủy Văn và Biến đổi khí hậu (IMHEN) cung cấp.

Do nhu cầu của mô hình dự đoán chuỗi thời gian chỉ cần thời gian và số ca nhiễm tương ứng, nên chúng tôi không sử dụng 12 đặc tính trên.

Để có thông tin ban đầu về số ca nhiễm cúm của tỉnh Hưng Yên trong giai đoạn 1997-2016, chúng tôi đã tiến hành vẽ biểu đồ dựa trên dữ liệu đã được cung cấp.

3.2. Xử lý dữ liệu

Hình 1 mô tả dữ liệu của tỉnh Hưng Yên về số ca nhiễm (trên 10000 dân) trong 20 năm (từ 1997-2016).



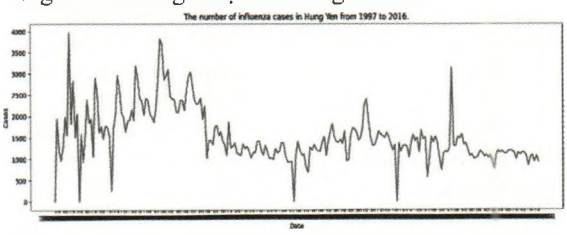
Hình 1. Vẽ biểu đồ về dữ liệu để có thông tin ban đầu về số ca nhiễm của Hưng Yên từ 1997-2016.

Sau khi quan sát và thống kê, chúng tôi thấy có một số dữ liệu bất thường hoặc bị thiếu. Đây có thể là do sai sót khi nhập liệu hoặc là thiếu (trong trường hợp ghi là 0).

Dữ liệu thiếu của một số tháng. Một số dữ liệu của các tháng (rất ít) không có dữ liệu nên chúng tôi xử lý bằng cách gán giá trị của tháng đó bằng giá trị trung bình trong năm.

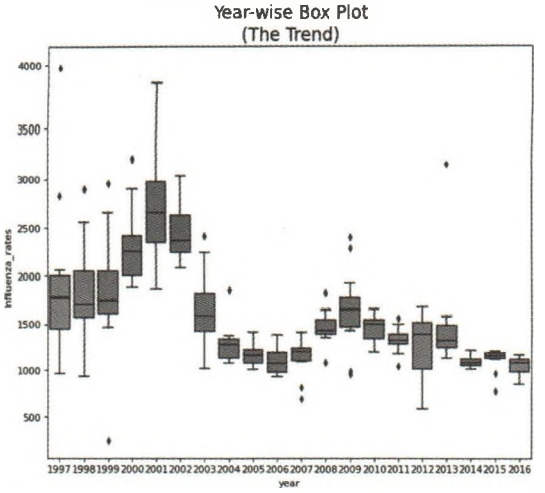
Qua biểu đồ ta thấy có dữ liệu bất thường ở tháng 12.1997 với 20598 ca nhiễm cúm. Cao gấp 10 lần bình thường. Chúng tôi tin rằng đây là sự nhầm lẫn về mặt nhập liệu. Do vậy, chúng tôi đã giảm đi 10 lần giá trị của tháng 12.1997.

Ngược lại, tháng 5.1999, với 252 ca, đây là con số nhỏ bất thường. Chúng tôi tin rằng đây là sự nhầm lẫn về mặt nhập liệu. Do vậy, chúng tôi đã tăng lên 10 lần giá trị của tháng 5.1999.



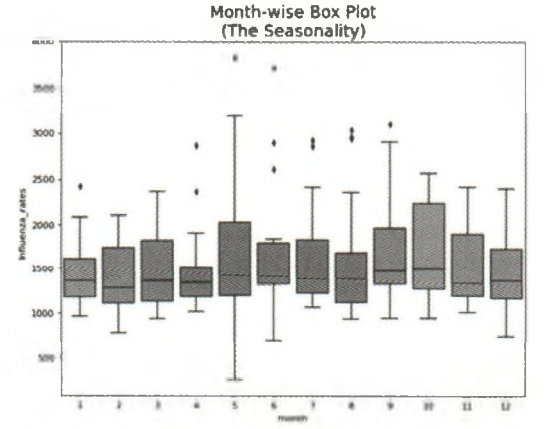
Hình 2. Biểu đồ về dữ liệu để có thông tin ban đầu về số ca nhiễm của Hưng Yên từ 1997-2016 sau khi xử lý.

Hình 2 là dữ liệu sau khi qua xử lý. Qua biểu đồ, chúng ta thấy dữ liệu đã loại bỏ được những điểm bất thường (ouliers).



Hình 3. Biểu đồ phân phối dữ liệu ca nhiễm của Hưng Yên theo năm từ 1997-2016.

Hình 3 cho chúng ta thấy những năm đầu có sự biến động rất lớn với số ca nhiễm cao, đặc biệt là năm 2001 có số ca nhiễm tăng bất thường. Sau đó số ca nhiễm giảm dần từ 2001 xuống năm 2006. Từ năm 2004 số ca nhiễm đã ổn định theo năm. Ở Hưng Yên, số ca nhiễm cũng tăng từ 2006 tới 2009, sao đó lại giảm dần tới năm 2016, mặc dù năm 2012 có sự biến động khá lớn.

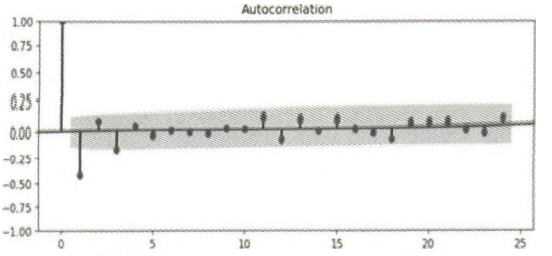


Hình 4. Biểu đồ phân phối dữ liệu ca nhiễm của Hưng Yên theo tháng từ 1997-2016.

Hình 4 cho chúng ta thấy số ca nhiễm theo tháng của Hưng Yên trong những năm từ 1997 tới 2016. Chúng ta thấy rằng số ca nhiễm của tháng 5 và tháng 10 thay đổi nhiều nhất. Có thể đây là thời điểm giao mùa trong năm. Tính trung bình thì số ca nhiễm cúm của Hưng Yên không có nhiều biến động.

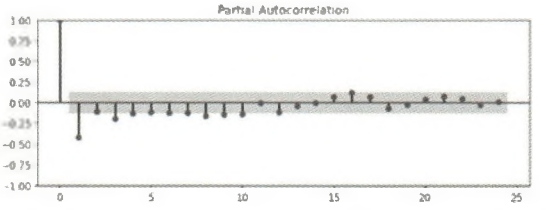
Trong bài toán này, khi kiểm tra dữ liệu gốc cho kết quả p-value = 0.736. Giá trị này không đủ bằng chứng bác bỏ giả thuyết H0. Tức là tập giá trị chuỗi thời gian có unit root dẫn tới tính không dừng (non-stationary).

Sau khi tính sai phân lần 1 (d=1) chúng ta tính được giá trị p-value = 1.843 e-10. Giá trị này là đủ cơ sở để bác bỏ giả thuyết H0. Tức là tập giá trị chuỗi thời gian không có unit root và có dừng (stationary). *Chúng ta sẽ không tính tiếp sai phân nữa!*



Hình 5. Biểu đồ tự tương quan thành phần của dữ liệu sau khi lấy sai phân d=1.

Hình 5 cho thấy ACF đang phân rã theo hàm mũ hoặc hình sin. Quan sát hình trên, chúng ta thấy có mứng tăng đột biến trong ACF q = 1.

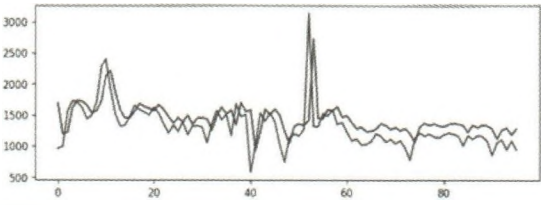


Hình 6. Biểu đồ tự tương quan của dữ liệu sau khi lấy sai phân d=1.

Hình 6 cho thấy PACF đang phân rã theo hàm mũ hoặc hình sin. Trong PACF cho độ trễ d = 1, có 1 mức tăng đột biến và sau đó không có mức tăng đột biến nào. Do vậy p = 1.

Phần sau sẽ trình bày hai mô hình AR và ARIMA trong quá trình dự đoán. Trong phần này, dữ liệu được chia làm 2 phần: tập huấn luyện (training) chiếm 2/3 tổng số dữ liệu (160 tháng), và tập kiểm tra (test) chiếm 1/3 (80 tháng). Với tập kiểm tra, chúng tôi thể hiện trên đồ thị đường màu đỏ là mô hình dự đoán, đường màu xanh là giá trị thực.

3.3. Mô hình AR



Hình 7. Biểu đồ so sánh kết quả dự đoán số ca nhiễm bằng mô hình AR(1) của Hưng Yên theo thời gian từ 1997-2016.

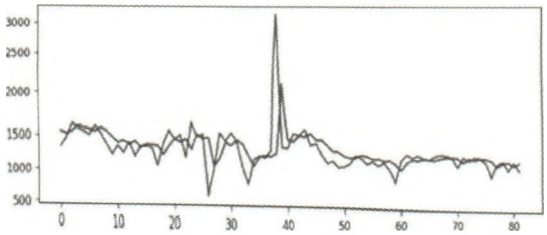
Hình 7 cho kết quả dự đoán của mô hình AR(1) về số ca nhiễm cúm của tỉnh Hưng Yên trong thời gian từ 1997-2016.

Kết quả sai số trung bình bình phương (root-mean-square error - RMSE) trong mô hình với độ trễ lag = 1 là AR(1) là 339.504.

3.4. Mô hình ARIMA

Kết quả sai số trung bình bình phương (root-mean-square - RMSE) là 290.937. Kết quả này nhỏ

hơn so với mô hình AR. Điều này chứng tỏ mô hình ARIMA hiệu quả hơn trong dự đoán.



Hình 8. Biểu đồ so sánh kết quả dự đoán số ca nhiễm bằng mô hình ARIMA(1,1,1) của Hưng Yên theo thời gian từ 1997-2016.

4. Kết luận

Trong bài báo này, chúng tôi đã thực hiện ba công việc sau. Thứ nhất, chúng tôi đã tìm hiểu và trình bày một số khái niệm và kĩ thuật quan trọng cùng mô hình dự báo chuỗi thời gian (mô hình AR, MA, ARIMA). Thứ hai, chúng tôi thu thập và xử lý dữ liệu cho bệnh truyền nhiễm của Hưng Yên trong 20 năm, từ 1997-2016. Thứ ba, chúng tôi dự đoán cho số ca nhiễm bằng hai mô hình, AR(1) và ARIMA(1,1,1). Trong khi mô hình AR cho kết quả sai số trung bình bình phương (root-mean-square - RMSE) là 313.590 thì mô hình ARIMA cho kết quả là 290.937. Điều này khẳng định mô hình ARIMA cho kết quả tốt hơn (sai số nhỏ hơn).

Trong tương lai, Chúng tôi muốn mở rộng hướng nghiên cứu theo hai hướng như sau. Thứ nhất, chúng tôi tiếp tục nghiên cứu và tìm hiểu những mô hình thống kê và học máy truyền thống khác ngoài AR và ARIMA. Ví dụ như: SARIMA (Seasonal ARIMA), ARCH (Autoregressive Conditional Heteroskedasticity). Thứ hai, chúng tôi cũng tìm hiểu và áp dụng phương pháp học máy hiện đại vào phân tích chuỗi thời gian. Đây là một hướng đi và tiếp cận phân tích chuỗi thời gian mới nhưng đã cho thấy nhiều hứa hẹn. Đặc biệt, chúng tôi sẽ sử dụng các mô hình học sâu (deep learning models) để dự báo chuỗi thời gian đã khắc phục được những hạn chế của học máy truyền thống với nhiều cách tiếp cận khác nhau. Hiện nay, Mạng thần kinh hồi quy (RNN) là kiến trúc cổ điển và được sử dụng nhiều nhất cho các bài toán Dự báo chuỗi thời gian.

Lời cảm ơn

Nghiên cứu này được tài trợ bởi trường Đại học Sư phạm Kỹ thuật Hưng Yên trong đề tài mã số UTEHY.L.2021.51.

Tài liệu tham khảo

[1]. Bộ Y tế (2011), *Hướng dẫn chẩn đoán và điều trị cúm mùa*, Ban hành kèm theo Quyết định số 2078/QĐ-BYT ngày 23 tháng 6 năm 2011.

[2]. Bộ Y tế (2019), *Hướng dẫn chẩn đoán, điều trị sốt xuất huyết Dengue*, Ban hành kèm theo Quyết định số 3705/QĐ-BYT ngày 22 tháng 8 năm 2019.

[3]. Eckstein D, Künzel V, Schäfer L, Winges M. GLOBAL CLIMATE RISK INDEX 2020. *Who Suffers Most from Extreme Weather Events? Weather-Related Loss Events in 2018 and 1999 to 2018. 2019.*

[4]. V. Sharma, Malaria, Outbreak prediction model using machine learning. *International Journal of Advanced Research in Computer Engineering and Technology.*

[5]. Sirisena P., Noordeen F., Kurukulasuriya H., Romesh T.A., Fernando L. Effect of climatic factors and population density on the distribution of Dengue in Sri Lanka: a gis based evaluation for prediction of outbreaks. *PLoS ONE*, 2017, **12(1)**, pp. 1–14.

[6]. Heinrichs B., Eickhoff S.B. Your evidence? Machine learning algorithms for medical diagnosis and prediction. *Hum. Brain Mapp*, 2020, **41(6)**, pp. 1435–1444.

[7]. Li Q., Cao W., Ren H., Ji Z., Jiang H. Spatiotemporal responses of Dengue fever transmission to the road network in an urban area. *Acta Trop*, 2018, **183**, pp. 8–13. http://www.sciencedirect.com/science/article/pii/S0001706X17311294.

APPLYING TIME SERIES MODELS FOR PREDICTING THE RATE OF INFLUENZA FLU IN HUNG YEN PROVINCE

Abstract:

Hung Yen is a province located in the center of the Red River Delta, North Vietnam is also a province heavily affected by infectious diseases. In recent times, many scientists have studied and applied time series models to predict the rate and the number of infectious diseases, thereby helping the health-care organizations to prevent the spread of deadly infectious disease outbreaks (e.g. dengue, diarrhea, influenza, etc.). In this paper, we investigate and apply three time series models: Auto regression (AR), moving average (MA), and Integrated Autoregression model. Autoregressive Integrated Moving Average (ARIMA). Experiments show that the ARIMA model gives better results.

Keywords: *time series models, Auto regression model, moving average model, Integrated Autoregression model, influenza flu.*