Dự đoán số ca nhiễm COVID-19 ngắn hạn bằng phương pháp Double Exponential Smoothing

Lê Minh Tú *, **Phan Lộc Sơn** †, **Nguyễn Xuân Vy** ‡
Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia TP. HCM
Thành phố Hồ Chí Minh, Việt Nam
{19120734,19120033†,19120732‡}@student.hcmus.edu.vn*

Tóm tắt nội dung: Bài báo trình bày phương pháp double exponential smoothing để ứng dụng dự đoán số ca nhiễm COVID-19 ngắn hạn của một vài tỉnh và thành phố của Việt Nam. Mục đích của nghiên cứu là nếu dự đoán được số ca trong ngắn hạn, sẽ có thể làm chỉ báo cho công tác phòng chống dịch hiệu quả hơn. Bài báo sử dụng phương pháp DES vì thích hợp cho tập dữ liệu với số lượng ít và chất lượng không cao hiện tại. Thông qua bài báo nhóm muốn giới thiệu mô hình và thử nghiệm mô hình với vài tỉnh thành Việt Nam, cho thấy kết quả dự đoán của mô hình với các tỉnh thành ấy là chấp nhận được. Đồng thời giải thích đơn giản các tham số của mô hình thử nghiệm mà thông qua đó giải thích đơn giản tình hình dịch bệnh của các tỉnh thành đang thử nghiệm.

1. Giới thiệu chung

Mặc dù tình hịnh dịch bệnh Việt Nam ở thời điểm hiện tại đang trong giai đoạn ổn định, nhưng biến chủng mới Omicron đang bùng phát mạnh ở các nước phương Tây, thậm chí các nhà khoa học còn tìm thấy biến chủng khác kết hợp từ hai biến chủng Delta và Omicron có mức độ nguy hiểm cao hơn. Vì thế nhóm thử nghiệm xem việc dự báo số ca nhiễm trong vòng 5 ngày bằng phương pháp Double Exponential Smoothing (DES) có phù hợp hay không, nếu phù hợp thì có thể sử dụng nhanh trong tương lai khi cần thiết. Nhóm chọn dự báo bằng phương pháp DES bởi vì phương pháp này phù hợp với dữ liệu chuỗi thời gian, thích hợp với các diễn biến giai đoạn dịch khác nhau.

Thách thức của nhóm trong nghiên cứu này chính là tập dữ liệu thu thập được. Vì tập dữ liệu không đủ nhiều và có ít thuộc tính do đặc trưng mỗi tỉnh thành khác nhau, nhiều yếu tố tác động như cách chống dịch của địa phương, kiến thức của người dân về dịch bệnh ở các địa phương có sự chênh lệch. Dữ liệu thu thập ở Việt Nam không được sạch, ví dụ những ngày cuối tuần không thu thập dữ liệu số ca, tâm lý của người dân sợ phải bị cách ly, khi người dân bị nhiễm thì không báo cho địa phương mà chủ động điều trị, cách ly tại nhà nên có rất nhiều ca nhiễm ẩn.

Trong nghiên cứu này, nhóm sử dụng phương pháp double exponential smoothing để dự đoán số ca nhiễm COVID-19 ngắn hạn của một vài tỉnh và thành phố của Việt Nam. Dữ liệu mà nhóm dùng để thực nghiệm được lấy từ Trung tâm Giám sát an toàn không gian mạng Quốc gia (NCSC) của Việt Nam.

2. Phương pháp

Phương pháp DES hay phương pháp của Holt là một trong những phương pháp dự đoán chuỗi thời gian (time series forecasting). Phương pháp đánh giá trọng số lớn cho những giá trị trong dữ liệu gần với thời điểm dự đoán hơn so với các giá trị ở xa. Mô hình gồm 2 thành phần là level và trend, khi các giá trị quan sát thay đổi theo thời gian, mô hình có thể thích nghi với các giá trị dữ liệu mới và cho ra dự đoán chính xác hơn. Điểm mạnh của mô hình so với các mô hình dự đoán khác khi dự đoán số ca nhiễm covid là có thể thích nghi nhanh với các thay đổi trong tình hình dịch (biến thể mới, thay đổi chính sách phòng chống dịch, tăng độ phủ của vaccine,...) mà vốn sự thay đổi diễn ra nhanh chóng.

Các thuật toán trong cùng lớp Exponential Smoothing như Simple Exponential Smoothing không phù hợp vì không có thành phần trend, vốn là thành phần giúp mô hình thích nghi nhanh với các thay đổi tình hình dịch, mô hình Holt Winter có thêm thành phần seasonality, vốn không cần thiết vì dữ liệu không có đặc tính này.

Mô hình DES có dạng như sau:

$$\begin{aligned} s_0 &= x_0 \\ b_0 &= x_1 - x_0 \end{aligned} \tag{1}$$

với mọi $t > 0$:

$$\begin{aligned} s_t &= \alpha * x_t + (1 - \alpha) * (s_{t-1} + b_{t-1}) \\ b_t &= \beta * (s_t - s_{t-1}) + (1 - \beta) * b_{t-1} \end{aligned} \tag{2}$$

Với α là thành phần smoothing dữ liệu, β là thành phần smoothing trend. Hai giá trị này nhận giá trị từ 0 tới 1. Để dự đoán từ ngày thứ t làm mốc và sau m ngày, ta sử dụng công thức xấp xỉ sau:

$$F_{t+m} = s_t + m * b_t \tag{3}$$

Ta thay đổi mô hình với tối ưu trend và level ban đầu so với mô hình giới thiệu ở trên, tuy nhiên sẽ không ảnh hưởng đến kết quả nhiều. Để dự đoán được thì trước tiên tối ưu bốn tham số là α, β, s_0 và b_0 phù hợp nhất tập dữ liệu mỗi tỉnh thành, bằng cách tối ưu sao cho log-likelihood là lớn nhất với thư viện của python cho mỗi tập dữ liệu tỉnh thành. Để hiểu rõ hơn cách tối ưu tham số, kham khảo "Estimation and Prediction for a Class of Dynamic Nonlinear Statistical Models" [1]

3. Thử nghiệm

Ví dụ vài dòng dữ liệu của tập thử nghiệm như bảng I, dữ liệu thử nghiệm nhóm xét từ ngày 20/02/2021 đến ngày 15/01/2022.

Thử nghiệm sử dụng ngôn ngữ python cùng thư viện statsmodels để tự động fit 4 tham số để dự đoán và mô tả tình hình dịch của 2 tỉnh thành khác nhau là Bình Định (37_BĐ) và Hà Nội (01_HN).

Ngày	Tổng ca nhiễm	Ca nhiễm mới	Tử vong	Hồi phục	Tiêm mũi 1	Tiêm mũi 2	Mã vùng
28/11/2020	169	5	0	165	0	0	01_HN
22/10/2021	4639	0	38	518	2573106	3438187	01_HN
17/12/2021	21760	0	64	0	1134615	5708905	01_HN
01/01/2022	47940	0	140	0	553393	6262272	01_HN
17/11/2021	2487	0	19	0	476778	323240	37_BĐ
06/12/2021	5505	0	25	0	307987	638134	37_BĐ
13/09/2021	878	0	10	1	24389	63147	37_BĐ
03/01/2022	17645	0	61	0	253012	887022	37_BĐ

Bảng I
VÍ DỤ VỀ MỘT SỐ MẪU DỮ LIỆU THỬ NGHIỆM

3.1. Bình Định

Tham khảo bảng II, kết quả các tham số sau khi fit trên dữ liệu của tỉnh Bình Định. Biểu đồ 3.1 biểu diễn kết quả dự đoán 5 ngày cuối cùng của tập dữ liệu, giá trị dự đoán cụ thể nằm trong bảng III.

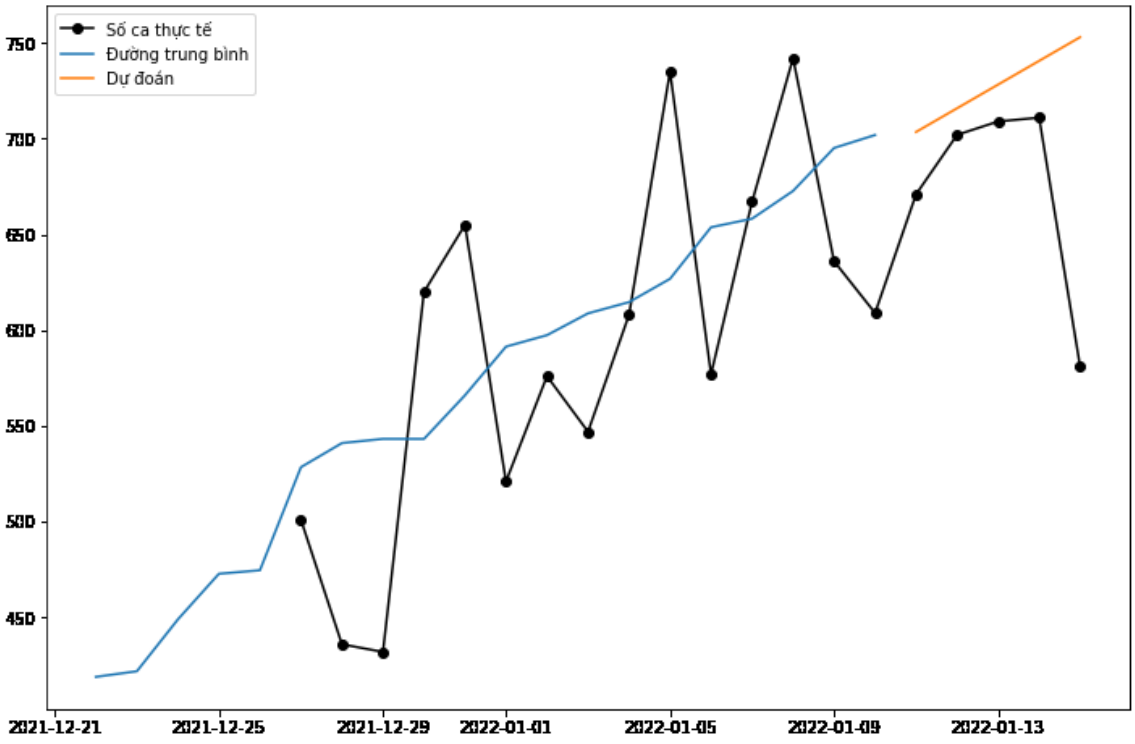
	Tham số	Giá trị	Tối ưu
smoothing_level	α	0.119831	Có
smoothing_trend	β	0.119831	Có
initial_level	s_0	-0.000818	Có
initial_trend	b_0	-0.000012	Có

Bảng II
CÁC THAM SỐ ĐƯỢC TỐI ƯU TRÊN DỮ LIỆU TỈNH BÌNH ĐỊNH

Ngày	Dự đoán	Thực tế
11/01/2022	703.53	671.0
12/01/2022	715.90	702.0
13/01/2022	728.27	709.0
14/01/2022	730.64	711.0
15/01/2022	753.01	581.0

Bảng III
DỰ ĐOÁN 5 NGÀY CUỐI CÙNG CỦA TẬP SỐ LIỆU BÌNH ĐỊNH

Kết quả dự đoán của Bình Định khá chính xác 3 ngày đầu và có độ lệch ở ngày 15/1. Điều này có thể do mô hình không dự đoán được trường hợp này hoặc chỉ báo số ca ghi nhận được có thể ít hơn so với số ca thực tế (15/1/2022 ứng với thứ 7, thường ghi nhận sụt giảm số ca vào ngày cuối tuần trong tập dữ liệu).



Hình 1. Dự báo 5 ngày tiếp theo tại Bình Định

3.2. Hà Nội

Tham khảo bảng IV, kết quả các tham số sau khi fit trên dữ liệu của thủ đô Hà Nội. Biểu đồ 3.2 biểu diễn kết quả dự đoán 5 ngày cuối cùng của tập dữ liệu, giá trị dự đoán cụ thể nằm trong bảng V.

	Tham số	Giá trị	Tối ưu
smoothing_level	α	0.679631	Có
smoothing_trend	β	0.037991	Có
initial_level	s_0	0.262133	Có
initial_trend	b_0	0.026290	Có

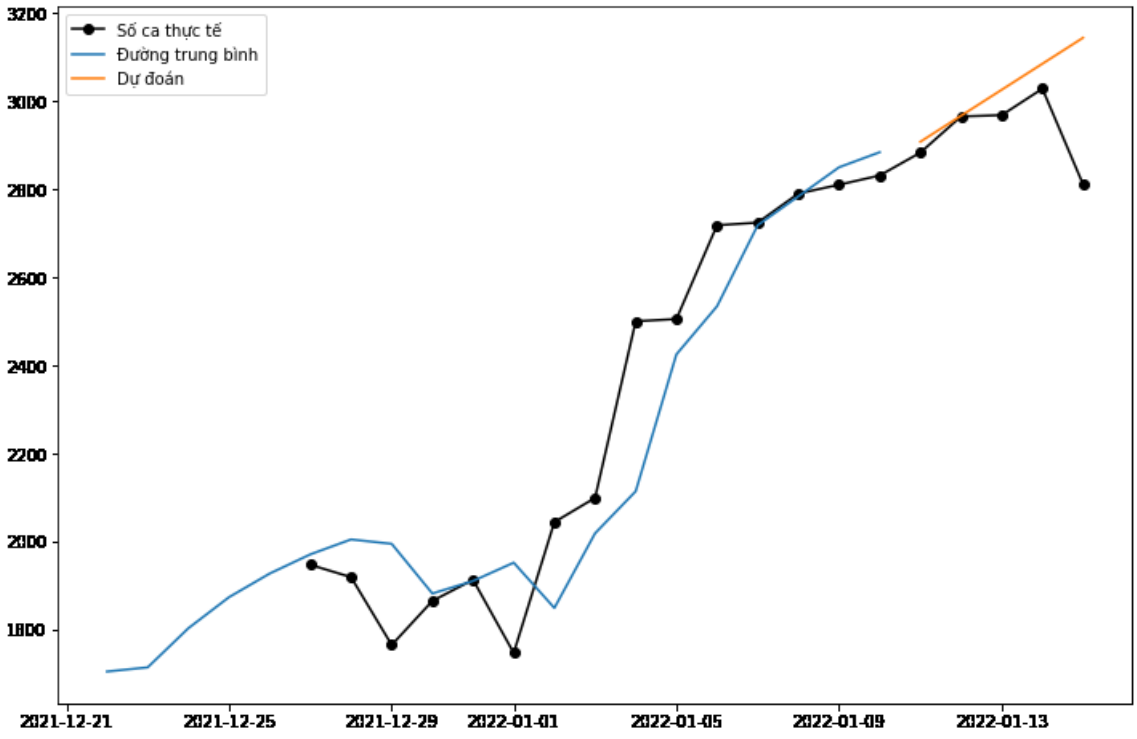
Bảng IV
CÁC THAM SỐ ĐƯỢC TỐI ƯU TRÊN DỮ LIỆU THỦ ĐÔ HÀ NỘI

Nhận xét tham số α khi so với Bình Định: lớn hơn nhiều so với Bình Định, tức số ca các ngày gần tác động lớn tới số ca của Hà Nội. Điều này là do Hà Nội đang trong đợt bùng phát dịch với số ca liên tục tăng cao kỷ lục, do đó số ca hôm nay phụ thuộc nhiều vào số ca các ngày gần kề.

Nhận xét kết quả dự đoán Hà Nội: dự đoán đúng hơn so với Bình Định vì đang có xu hướng tăng số ca cao. Tuy nhiên dự đoán ngày thứ 5 có sự chênh lệch và sẽ tiếp tục lệch nếu dự đoán rộng hơn, cần thêm tập dữ liệu liên tục.

Ngày	Dự đoán	Thực tế
11/01/2022	2908.64	2884.0
12/01/2022	2967.64	2966.0
13/01/2022	3026.65	2969.0
14/01/2022	3085.66	3029.0
15/01/2022	3144.67	2812.0

Bảng V
DỰ ĐOÁN 5 NGÀY CUỐI CÙNG CỦA TẬP SỐ LIỆU THỦ ĐÔ HÀ NỘI



Hình 2. Dự báo 5 ngày tiếp theo tại Hà Nội

4. Kết luận

Với tình dịch vẫn bất ổn với hiện tại (01/2022), đòi hỏi cần phải đưa ra biện pháp để kiểm soát ổn định. Thử nghiệm cũng cho thấy kết quả khả quan với tập dữ liệu. Thông qua nghiên cứu sử dụng phương pháp DES, nhóm mong muốn nghiên cứu được xem như là chỉ báo ngắn hạn về số ca nhiễm tương lai gần, góp phần trong phòng chống dịch như giúp truy soát đủ số ca bệnh hằng ngày, chuẩn bị cho tình hình khẩn cấp, thiếu giường bệnh, đưa ra chiến lược chống dịch tốt nhất. Nghiên cứu cũng cho thấy xu hướng tăng hiện tại số ca ở cả Bình Định và Hà Nội.

Tài liệu liên quan

[1] J. K. Ord, A. B. Koehler, and R. D. Snyder, “Estimation and prediction for a class of dynamic nonlinear statistical models,” *Journal of the American Statistical Association*, vol. 92, no. 440, pp. 1621–1629, 1997. DOI: 10.1080/01621459.1997.104 eprint: https://doi.org/10.1080/01621459.1997.10473684. [Online]. Available: https://doi.org/10.1080/01621459.1997.10473684.