A Shiny application with ARIMA model to evaluate pandemic COVID-19

Nguyen Quoc Duong[1*], Le Phuong Thao[1], Dinh Thi Quynh Nhu[1]
Cao Thi Ai Loan[2], Phung Thi Hong Diem[2], Le Thanh Binh[2]

[1]Faculty of Education, Quy Nhon University
[2]Faculty of Mathematics and Statistics, Quy Nhon University

[*]Corresponding author*: nguyenquocduongqnu1999@gmail.com*

Abstract:

Nowadays, data visualization is an important tool for exploring and representing findings in medical research, and specially in epidemiological surveillance. For forecast of time series, ARIMA model is one of the best modeling techniques. The objective of this study is to use the Shiny package, an available package of R language, and ARIMA model for creating a COVID-19 pandemic trend prediction website in all over the world. The COVID-19 Dashboard app provides daily updated data analysis of COVID-19 pandemic.

Keywords: ***ARIMA, auto.arima, COVID-19, forecast, Shiny package.***

Classification number: *1.1*

Ứng dụng Shiny kết hợp với mô hình ARIMA để đánh giá đại dịch COVID-19

Nguyễn Quốc Dương[1*], Lê Phương Thảo[1], Đinh Thị Quỳnh Như[1], Cao Thị Ái Loan[2], Phùng Thị Hồng Diễm[2], Lê Thanh Bính[2]

[1]Khoa Sư phạm, Trường Đại học Quy Nhơn
[2]Khoa Toán và Thống kê, Trường Đại học Quy Nhơn

Tóm tắt:

Ngày nay, trực quan hóa dữ liệu là một công cụ quan trọng để khám phá và mô tả những phát hiện trong nghiên cứu y học và đặc biệt là trong giám sát dịch tễ học. Đối với dự báo chuỗi thời gian, ARIMA là một trong những kỹ thuật mô hình hóa tốt nhất. Mục đích của nghiên cứu này là sử dụng gói lệnh Shiny, gói lệnh có sẵn của R, và mô hình ARIMA để xây dựng một website để dự báo khuynh hướng đại dịch COVID-19 cho mỗi quốc gia trên toàn thế giới. Ứng dụng COVID-19 Dashboard cập nhật dữ liệu phân tích hàng ngày của đại dịch COVID-19.

Từ khóa: ***ARIMA, auto.arima, COVID-19, forecast, Shiny package.***

Chỉ số phân loại: ***1.1***

Đặt vấn đề

Trong khi đại dịch COVID-19 đang diễn ra, rất nhiều nhóm nghiên cứu trên thế giới đã tìm cách góp phần chuyên môn của mình vào việc kiểm soát dịch. Không chỉ trong chuyên ngành dịch tễ học, mà rất nhiều chuyên gia từ các chuyên ngành khác nhau cũng tham gia trực tiếp hay gián tiếp. Điển hình như giới miễn dịch học, toán và thống kê, … Trong đó, các mô hình dự báo sự lây lan của dịch bệnh đóng một vai trò rất quan trọng trong việc hoạch định nhằm đưa ra các phương hướng xử lý tối ưu nhất cho mỗi quốc gia.

Có rất nhiều mô hình dự báo khác nhau đã được các nhà nghiên cứu sử dụng như ETS (Exponential Smoothing), FLM (Functional Linear Model), SEIR (Susceptible - Exposed - Infectious - Recovered), … [1]. Trong số các mô hình dự báo đó, một vài nghiên cứu đã sử dụng ARIMA như là một công cụ hữu ích trong việc dự báo khuynh hướng dịch bệnh COVID-19. Nguyễn Quốc Dương và các cộng sự đã dự báo số ca nhiễm mới COVID-19 trên toàn thế giới bằng mô hình ARIMA [2]. Hay Alzahrani SI và các cộng sự đã sử dụng mô hình ARIMA để dự báo sự lây lan của đại dịch COVID-19 tại Saudi Arabia [3]. Hơn nữa, Lutfi Bayyurt và Burcu Bayyurt đã áp dụng mô hình ARIMA để dự báo số ca nhiễm mới và số ca tử vong mới [4]. Do đó, mô hình ARIMA có thể được xem như là một công cụ dự báo tốt giúp các cơ quan y tế giám sát, hoạch

định các chính sách nhằm kiểm soát sự lây lan của dịch bệnh COVID-19.

Để thực hiện tính toán nhanh chóng và chính xác cho các mô hình, phần mềm phân tích số liệu là một công cụ không thể thiếu đối với các nhà thống kê. Nó như là một công cụ hỗ trợ cho việc thực hiện các ước lượng, tính toán nhanh chóng trong quá trình phân tích dữ liệu. Hiện nay, có rất nhiều phần mềm chuyên dụng phục vụ cho việc xử lý và phân tích số liệu thống kê như: SAS, SPSS, STATA, R … Trong đó, R là một ngôn ngữ rất mạnh và được nhiều nhà thống kê học sử dụng cho công việc của mình. Tuy R là mã nguồn mở nhưng chức năng của nó không thua kém các phần mềm thương mại đắt tiền khác. Tất cả những bài toán, mô hình mà các phần mềm thương mại có thể làm được thì R cũng có thể làm được. R có lợi thế là khả năng phân tích biểu đồ rất tuyệt vời. Không một phần mềm nào có thể sánh với R về phần biểu đồ. Hơn nữa, R gắn liền với giới học thuật nên hầu hết những mô hình thống kê mới nhất đều được hỗ trợ bởi R.

Xuất phát từ những ưu điểm của R, các nhà nghiên cứu đã sử dụng phần mềm này như là công cụ rất hữu ích để phân tích dữ liệu COVID-19 [1]. Đặc biệt, để đưa các kết quả nghiên cứu đến với mọi người, R còn có chức năng xây dựng một ứng dụng website với gói lệnh "Shiny" [1]. Một số nhà nghiên cứu đã sử dụng mô hình ARIMA kết hợp với gói lệnh "Shiny" để xây dựng website dự báo khuynh hướng dịch bệnh COVID-19. Fabio Caironi và các cộng sự đã xây dựng ứng dụng dự báo dịch bệnh COVID-19 cho các tỉnh và thành phố ở Italy bằng các mô hình ARIMA, logistic và SEIR [1]. Hay Jamal Kay Rogers và Yvonne Grace Arandela đã xây dựng website dự báo dịch bệnh COVID-19 cho Philippine bằng mô hình ARIMA [1]. Do đó, ARIMA kết hợp với gói lệnh "Shiny" để xây dựng website dự báo là một ứng dụng thiết thực và hữu ích. Xuất phát từ những điều trên, chúng tôi xây dựng một website dự báo khuynh hướng về số ca nhiễm mới, số ca tử vong mới, số ca phục hồi mới cho từng quốc gia trên toàn thế giới bằng mô hình ARIMA kết hợp với gói lệnh "Shiny".

Đối tượng và phương pháp

Nguồn dữ liệu

Để đảm bảo tính chính xác và độ tin cậy của website, chúng tôi sử dụng bộ dữ liệu từ Johns Hopkins University (https://github.com/CSSEGISandData/COVID-19), dữ liệu được cập nhật hàng ngày bởi nhóm ESRI Living Atlas. Hầu hết các nghiên cứu phân tích dữ liệu COVID-19 đều sử dụng nguồn dữ liệu từ hệ thống này [1].

Phần mềm R và gói lệnh "Shiny"

Ứng dụng của chúng tôi được xây dựng trên nền tảng RStudio phiên bản 1.2.5033, sử dụng gói lệnh "Shiny" phiên bản 1.4.0. Shiny là một R package cho phép tạo các ứng dụng web tương tác (Interactive Web Applications) vô cùng đơn giản trực tiếp trong môi trường của R [5, 6]. Tất cả các phân tích được thực hiện trong môi trường R với phiên bản 4.0.1.

Mô hình ARIMA

Vào năm 1970, Box và Jenkins đã đưa ra dạng tổng quát của mô hình ARIMA. Mô hình ARIMA là sự kết hợp

giữa ba quá trình AR (Auto-Regressive), MA (Moving Average) và quá trình sai phân bậc d [7, 8].

Mô hình AR(p) tổng quát có dạng:

$$\phi(\mathrm{B})\mathrm{x_t} = \mathrm{w_t},$$

trong đó $\phi(\mathrm{B}) = 1 - \phi_1 B - \phi_2 \mathrm{B}^2 - \cdots - \phi_\mathrm{p}\mathrm{B}^\mathrm{p}$ là toán tử tự hồi quy bậc p với $\phi_1, \phi_2, \ldots, \phi_\mathrm{p}$ $(\phi_p \neq 0)$ là các hằng số, $\mathrm{x_t}$ là chuỗi có tính dừng, $\mathrm{w_t}$ là chuỗi nhiễu trắng Gaussian và B là toán tử backshift.

Mô hình MA(q) tổng quát có dạng:

$$\mathrm{x_t} = \theta(\mathrm{B})\mathrm{w_t},$$

trong đó $\theta(\mathrm{B}) = 1 + \theta_1 B + \theta_2 \mathrm{B}^2 + \cdots + \theta_\mathrm{q} B^\mathrm{q}$ là toán tử trung bình trượt bậc q với $\theta_1, \theta_2, \ldots, \theta_\mathrm{q}$ $(\theta_\mathrm{q} \neq 0)$ là các tham số, x_t là chuỗi có tính dừng, w_t là chuỗi nhiễu trắng Gaussian và B là toán tử backshift.

Quá trình sai phân bậc d có dạng: $\Delta^\mathrm{d}\mathrm{x_t} = (1 - \mathrm{B})^\mathrm{d}\mathrm{x_t}$.

Kết hợp ba quá trình trên, ta được *phương trình ARIMA tổng quát có dạng*:

$$\phi(\mathrm{B})(1 - \mathrm{B})^\mathrm{d}\mathrm{x_t} = \theta(\mathrm{B})\mathrm{w_t}.\mathrm{p}$$

Để chọn các tham số thích hợp cho mô hình ARIMA, chúng ta sử dụng các tiêu chuẩn thông tin *AIC* (Akaike's Information Criterion), *AICc* (AIC hiệu chỉnh) và *BIC* (Bayesian Information Criterion). Mô hình được chọn là tốt nhất nếu một trong ba tiêu chí thông tin là nhỏ nhất [8]. Công thức của các tiêu chí được đưa ra như sau:

$$\mathrm{AIC} = -2\ln(L) + 2\mathrm{k},$$

$$\mathrm{AIC_C} = \mathrm{AIC} + \frac{2(\mathrm{p}+1)(\mathrm{p}+2)}{\mathrm{n}-\mathrm{p}},$$

$$\mathrm{BIC} = \mathrm{AIC} + \mathrm{p}(\ln(\mathrm{T}) - 2),$$

trong đó L là giá trị likelihood, k là số các tham số được ước lượng trong mô hình và T là cỡ mẫu. Khoảng tin cậy 95% dự báo cho mô hình ARIMA được tính theo công thức

$$\hat{\mathrm{y}}_{\mathrm{T+h|T}} \pm 1.96\sqrt{\mathrm{v_{T+h|T}}},$$

trong đó $v_{T+h|T}$ đề cập đến phương sai của $\mathrm{y_{T+h|y_1,\ldots,y_T}}$.

Hàm auto.arima và hàm forecast

Chúng tôi sử dụng gói lệnh "forecast" được phát triển bởi *Hyndman-Khandakar (2008)* để thực hiện phân tích dữ liệu [9, 10]. Trong gói lệnh này, hàm *auto.arima* có chức năng đưa ra mô hình phù hợp nhất để tiến hành dự báo. Thuật toán và các bước chọn tự động mô hình ARIMA của hàm *auto.arima* được trình bày chi tiết tại [10]. Sau khi chọn được mô hình phù hợp nhất, chúng tôi sử dụng hàm *forecast* để dự báo khuynh hướng cho 7 ngày tiếp theo.

Kết quả và bàn luận

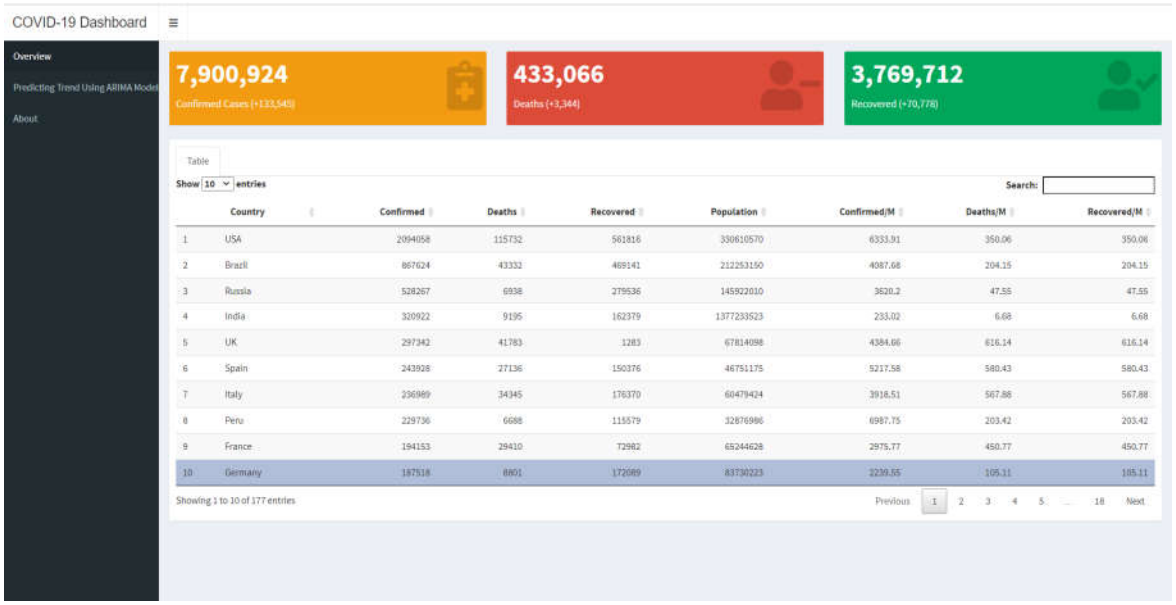
Xuất phát từ tính năng hữu ích của gói lệnh Shiny và mô hình ARIMA, chúng tôi xây dựng một website có tên là COVID-19 Dashboard. Mục tiêu của website này là để theo dõi tình dịch bệnh tại mỗi quốc gia trên toàn thế giới và dự báo khuynh hướng dịch bệnh cho 7 ngày tiếp theo. Ứng dụng này gồm 2 thẻ chính là "Overview" và "Predicting Trend Using Arima Model". Link website: https://nguyenquocduong.shinyapps.io/NCKH.

Thẻ "Overview"

Trong thẻ này, chúng tôi trình bày

tổng quan về tình hình dịch COVID-19 bằng cách hiển thị các con số thống kê dịch bệnh. Ba hộp trên đầu trang web với ba màu yellow, red, green lần lượt là tổng số ca nhiễm tích lũy (confirmed), tổng số ca tử vong tích lũy (deaths), tổng số ca phục hồi tích lũy (recovered) trên toàn thế giới. Đồng thời, chúng tôi hiển thị thêm dữ liệu về số ca nhiễm mới, số ca tử vong mới và số ca phục hồi mới tại mỗi hộp giúp người dùng có cái nhìn trực quan hơn về mức độ lây nhiễm mới hàng ngày của đại dịch COVID-19.

Để truy cập thông tin dịch bệnh của mỗi quốc gia, chúng tôi xây dựng một bảng hiển thị đầy đủ thông tin về dữ liệu của các quốc gia có dịch. Tùy vào mục đích quan sát, người dùng có thể khai thác thông tin về quốc gia mà mình muốn theo dõi một cách nhanh chóng và tiện lợi. Tag này được mô tả như hình 1.

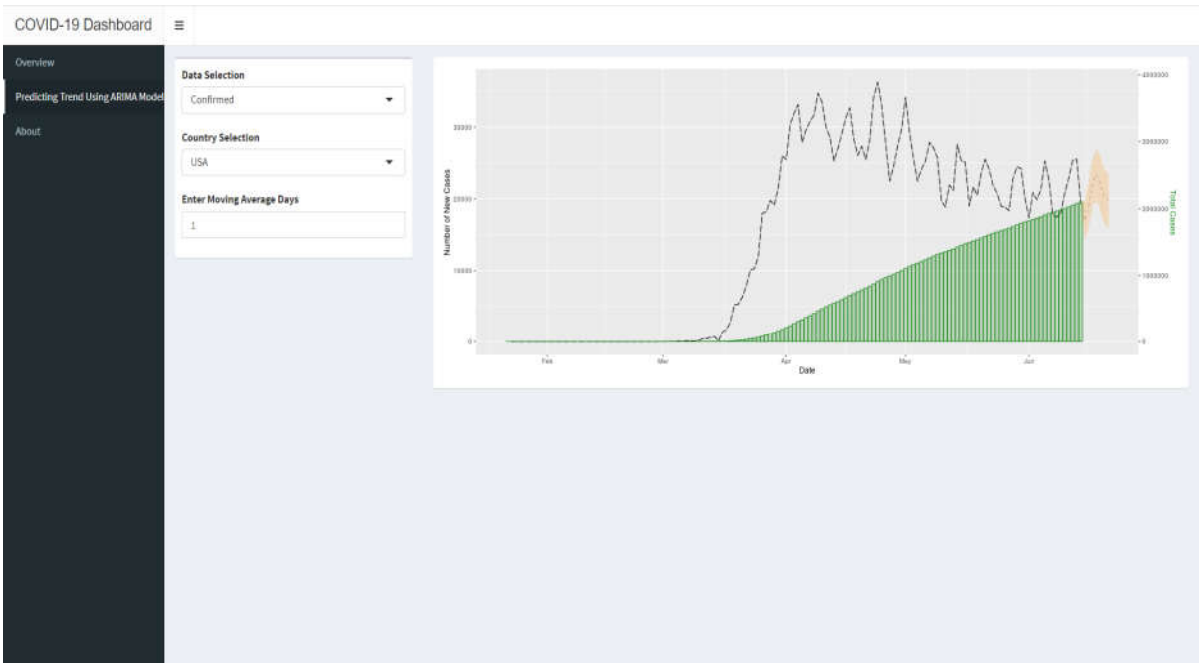


Hình 1. Thẻ overview của website COVID-19 Dashboard (15/06/2020).

Thẻ "Predicting Trend Using ARIMA Model"

Trong thẻ này, chúng tôi sử dụng mô hình ARIMA kết hợp với gói lệnh "Shiny" để dự báo khuynh hướng dịch bệnh COVID-19 cho từng quốc gia cụ thể. Khuynh hướng được dự báo bao gồm 3 tùy chọn là số ca nhiễm mới, số ca tử vong mới và số ca phục hồi mới. Đồng thời, với mỗi bộ dữ liệu được phân tích, chúng tôi vẽ đồ thị trực quan dữ liệu tích lũy các bộ dữ liệu tương ứng (số ca nhiễm, số ca phục hồi và số ca tử vong).

Để truy cập quan sát cho quốc gia nào, chúng ta chỉ cần chọn quốc gia đó và lựa chọn dữ liệu tương ứng (confirmed, deaths và recovered). Ngoài ra, chúng tôi cung cấp thêm đường trung bình trượt (moving average) để làm mịn chuỗi thời gian quan sát và mặc định là 1 nên đường trung bình trượt trùng với số trường hợp mới. Tag này được mô tả như hình 2.



Hình 2. Thẻ dự báo khuynh hướng dịch bệnh bằng mô hình ARIMA (15/06/2020).

Quan sát hình 2, chúng ta thấy đường đồ thị màu xanh biểu thị cho tổng số trường hợp tích lũy tương ứng với số ca nhiễm. Số liệu quan sát tương ứng được hiển thị ở trục tung bên phải đồ thị, tức tương ứng với dòng chữ "total cases" màu xanh.

Chuỗi thời gian màu xám biểu thị số trường hợp mới tương ứng với dữ liệu được lựa chọn để quan sát. Số liệu quan sát tương ứng được hiển thị tại trục tung bên trái đồ thị, tức tương ứng với dòng chữ "number of new cases".

Chúng tôi sử dụng mô hình ARIMA để dự báo khuynh hướng cho 7 ngày tiếp theo (vùng màu hồng tương ứng với khoảng dự báo 95%). Từ khuynh hướng và khoảng dự báo này, người dùng có thể đưa ra các phán đoán cho 7 ngày tiếp theo và có những hoạch định tốt nhất cho công việc của mình.

Ứng dụng COVID-19 Dashboard có thể là một công cụ rất hữu ích để theo dõi khuynh hướng và giám sát dịch tễ học của dịch COVID-19 cho từng quốc gia trên toàn thế giới. Từ khuynh hướng đó, website là cơ sở rõ ràng trực quan, kịp thời và đúng lúc giúp các trung tâm kiểm soát và nhà hoạch định đưa ra các biện pháp kiểm soát dịch tối ưu nhất cho quốc gia của mình.

Chúng tôi tiếp tục lên kế hoạch cải tiến cho ứng dụng bằng cách thêm thẻ bản đồ phân bố dịch nhằm trực quan bảng số liệu được xây dựng ban đầu. Từ bản đồ này, chúng ta có thể so sánh và đánh giá được mức độ nguy hiểm của dịch bệnh đối với từng quốc gia. Hơn nữa, mỗi quốc gia đều có đặc điểm về thể lực, môi trường, khí hậu và phong tục sinh hoạt khác nhau, do đó, mô hình này có thể thích hợp với quốc gia này nhưng không phù hợp với quốc gia khác. Chính

vì vậy, việc đối sánh giữa các mô hình dự báo với nhau là rất quan trọng. Do đó, chúng tôi sẽ xây dựng thêm một thẻ đối sánh giữa các mô hình dự báo và tự động đưa ra mô hình phù hợp nhất đối với từng quốc gia cụ thể. Chúng tôi dự định sẽ so sánh giữa mô hình ARIMA với các mô hình lai ghép (hybrid) của nó như *ARIMA-ANN* (Artificial Neural Network), *ARIMA-LSTM* (Long Short Term Memory), …. Từ đó, chúng tôi sẽ thiết lập để website tự động đưa ra mô hình phù hợp nhất dựa vào các thước đo sai số từ các mô hình. Việc đối sánh này không chỉ giúp xác định mô hình nào phù hợp với dân số nghiên cứu nào mà còn tăng độ tin cậy chắc chắn của ứng dụng. Tóm lại, ứng dụng này dễ sử dụng và có thể sử dụng để tham khảo trong việc đưa ra các kịch bản kiểm soát dịch cho từng quốc gia cụ thể.

Kết luận

Trong nghiên cứu này, chúng tôi đã kết hợp gói lệnh có sẵn "Shiny" trong R và mô hình lý thuyết ARIMA để xây dựng website COVID-19 Dashboad nhằm theo dõi và dự báo khuynh hướng cho dịch COVID-19 cho 7 ngày tiếp theo. Qua đó, người dùng có được cái nhìn trực quan để đánh giá ban đầu về tình hình dịch bệnh. Đối với ở Việt Nam, kết quả của chúng tôi hoàn toàn mới mẻ trong việc dùng Shiny để xây dựng website dự báo. Đó là một bước đầu quan trọng để chúng tôi đưa ra các công cụ dự báo trực tuyến chính xác và mạnh mẽ hơn cho các loại bệnh khác trong tương lai.

TÀI LIỆU THAM KHẢO

[1] Antoine Soetewey (2020), "Top 100 R resources on Novel COVID-19 Coronavirus", Towards Data Science, https://bitly.com.vn/CzdP3.

[2] Nguyen Quoc Duong, et al. (2020), "Predicting the Pandemic COVID-19 using ARIMA Model", *VNU Journal of Science: Mathematics-Physics*, **36(4)**.

[3] Alzahrani SI, et al. (2020), "Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions", *Journal of Infection and Public Health*.

[4] Lutfi Bayyurt, Burcu Bayyurt (2020), "Forecasting of COVID-19 Cases and Deaths Using ARIMA Models", *The preprint server for health sciences*.

[5] Keon-Woong Moon (2017), "Learn ggplot2 Using Shiny App", *Springer*.

[6] Chris Beeley (2013), "Web Application Development with R Using Shiny", *Packt Publishing*.

[7] R. H. Shumway, D. S. Stoffer (2006), "Time Series Analysis and Its Applications", *Springer Publisher*.

[8] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, G. M. Liung (2016), "Time Series Analysis: Forecasting and Control, 5th edition", *Publisher Wiley*.

[9] R. J. Hyndman, G. Athanasopoulos (2018), "Forecasting: Principles and Practice, 2nd edition", *OTexts Publisher*.

[10] R. J. Hyndman, Y. Khandakar (2008), "Automatic Time Series Forecasting: The forecast Package for R", *Journal of Statistical Software*, **27(3)**.