Mục lục

THÔNG TIN KẾT QUẢ NGHIÊN CỨU CỦA ĐỀ TÀI

1. Thông tin chung:

- Tên đề tài: Phân tích chuỗi thời gian bằng mô hình ARIMA với phần mềm R
- Mã số: **S2019.564.01**
- Sinh viên thực hiện: - Cao Thị Ái Loan[1]
  - Phùng Thị Hồng Diễm[1]
  - Nguyễn Quốc Dương[2]
  - Lê Phương Thảo[2]
  - Đinh Thị Quỳnh Như[2]

[1]Lớp Toán học K39 Khoa: Toán và Thống kê Năm thứ: 4 Số năm đào tạo: 4

[2]Lớp Sư Phạm Toán K40 Khoa: Sư Phạm Năm thứ: 3 Số năm đào tạo: 4

- Người hướng dẫn: TS. Lê Thanh Bính

2. Mục tiêu đề tài:

- Nắm vững lý thuyết cơ bản về chuỗi thời gian và mô hình ARIMA; xây dựng quy trình dự báo bằng mô hình ARIMA gồm 5 bước sau: kiểm tra tính dừng của chuỗi thời gian; chuyển một chuỗi không dừng thành dừng; xác định bậc p, q, P, Q; kiểm tra độ chính xác của mô hình; bước cuối cùng là, dự báo.

- Nắm vững quy trình xây dựng mô hình ARIMA (với 5 bước nêu trên) bằng phần mềm R cho các dữ liệu thời gian thực trong những lĩnh vực khác nhau như khí tượng thủy văn; dịch tễ học; giáo dục; tài chính; chứng khoán.

3. Tính mới và sáng tạo: Áp dụng mô hình ARIMA để nghiên cứu độc lập trên các tập dữ liệu thực tế, ứng dụng trực tiếp cho địa phương và dự báo dịch bệnh COVID-19. Hơn nữa, chúng tôi so sánh khả năng dự báo của mô hình ARIMA với NNAR để có cái nhìn khách quan hơn.

4. Kết quả nghiên cứu: Phân tích tổng quan tình hình COVID-19 trên toàn thế giới, phân tích và dự báo số ca nhiễm mới COVID-19 tại Mỹ và số ca tử vong mới COVID-19 tại Italy, dự báo lượng mưa tại trạm quan trắc Quy Nhơn và tạo một website Dashboard COVID-19 để theo dõi xu hướng dịch bệnh bằng mô hình ARIMA.

5. Đóng góp về mặt kinh tế-xã hội, giáo dục và đào tạo, an ninh, quốc phòng và khả năng áp dụng của đề tài: Đề tài hoàn thành là tài liệu tham khảo hữu ích cho những ai quan tâm đến phân tích dữ liệu chuỗi thời gian, đặc biệt là mô hình ARIMA với phần mềm R. Hơn nữa, đề tài còn góp phần vào công cuộc chống đại dịch COVID-19 bằng cách xây dựng website Dashboard COVID-19.

6. Công bố khoa học của sinh viên từ kết quả nghiên cứu của đề tài:

- Bài báo "**Monthly Rainfall Forecast of Quy Nhon using SARIMA**

Model" được chấp nhận đăng trên *Tạp chí khoa học Trường đại học Quy Nhơn.*

- Bài báo "**Predicting the Pandemic COVID-19 using ARIMA Model**" được chấp nhận đăng trên tạp chí *VNU Journal of Science: Mathematics-Physics*, Đại học Quốc gia Hà Nội (tạp chí được xuất bản bằng tiếng Anh).

- Đã gửi bản thảo bài báo "**Modeling Total Vehicle Sales data in USA to forecasting: A comparison between the Holt-Winters, ARIMA and NNAR models**" đến tạp chí *International Journal of Applied Mathematics and Statistics* (India).

Ngày 28 tháng 05 năm 2020

Sinh viên chịu trách nhiệm chính thực hiện đề tài

Cao Thị Ái Loan

Nhận xét của người hướng dẫn về những đóng góp khoa học của sinh viên thực hiện đề tài:

Nhóm sinh viên thực hiện đề tài đã dành rất nhiều thời gian, công sức để tập trung tìm tòi và nỗ lực đọc hiểu kĩ các tài liệu bằng tiếng Anh với lượng kiến thức rất lớn liên quan đến xác suất thống kê, lý thuyết chuỗi thời gian, mô hình ARIMA và các công cụ của phần mềm R. Trong quá trình thực hiện nghiên cứu, có những vấn đề nảy sinh không có trong tài liệu chuyên khảo hoặc đề cập nhưng không rõ ràng, nhóm tác giả đã chủ động trao đổi trên diễn đàn thống kê học máy với các chuyên gia và đã học hỏi được nhiều kiến thức. Tôi đánh giá rất cao phương pháp và tinh thần chủ động học tập đó. Hơn nữa, nhóm sinh viên đã giành rất nhiều thời gian để nghiên cứu và học cách trình bày kết quả nghiên cứu dưới dạng bài báo bằng tiếng Anh. Chính vì vậy, tôi dành lời khen rất lớn về thái độ, tinh thần làm việc, say mê nghiên cứu của nhóm sinh viên. Một lần nữa, tôi cho rằng nhóm sinh viên thực hiện đề tài đã hoàn thành rất xuất sắc công việc mà người hướng dẫn đã đặt ra ban đầu.

Ngày 28 tháng 05 năm 2020

Xác nhận của Khoa **Người hướng dẫn**

PGS. TS. Lê Công Trình **TS. Lê Thanh Bính**

THÔNG TIN VỀ SINH VIÊN
CHỊU TRÁCH NHIỆM CHÍNH THỰC HIỆN ĐỀ TÀI

I. SƠ LƯỢC VỀ SINH VIÊN:

Ảnh 4×6

Họ và tên: Cao Thị Ái Loan
Sinh ngày: 22/07/1997
Nơi sinh: Thị Xã An Nhơn - Bình Định
Lớp: Toán học K39 **Khóa:** 39
Khoa: Toán và Thống kê
Địa chỉ liên hệ: 56 Võ Mười, Thành phố Quy Nhơn
Điện thoại: 0965375088 **Email:** caoailoan@gmail.com

II. QUÁ TRÌNH HỌC TẬP:

**Năm thứ 1:*
Ngành học: Cử nhân Toán học **Khoa:** Toán và Thống kê
Kết quả xếp loại học tập: Khá
Sơ lược thành tích:

**Năm thứ 2:*
Ngành học: Cử nhân Toán học **Khoa:** Toán và Thống kê
Kết quả xếp loại học tập: Khá
Sơ lược thành tích:

**Năm thứ 3:*
Ngành học: Cử nhân Toán học **Khoa:** Toán và Thống kê
Kết quả xếp loại học tập: Khá
Sơ lược thành tích:

Ngày 28 tháng 05 năm 2020

Xác nhận của Khoa **Sinh viên chịu trách nhiệm chính**

PGS. TS. Lê Công Trình **Cao Thị Ái Loan**

Mở đầu

1. Tổng quan tình hình nghiên cứu thuộc lĩnh vực đề tài:

Vào năm 1970, mô hình ARIMA được nghiên cứu và phát hiện bởi hai nhà thống kê học *G. E. P. Box và G. M. Jenkins*. Vì vậy, loại mô hình này còn được biết đến với tên gọi là phương pháp Box-Jenkins. Có thể hiểu, ARIMA là mô hình được sử dụng để dự đoán và khai phá dữ liệu trong nhiều lĩnh vực khác nhau như trong lĩnh vực giáo dục để dự báo số sinh viên nhập học của một trường đại học; hay trong lĩnh vực khác như dự báo nhu cầu sử dụng điện, hay dự báo khí tượng thủy văn, ... Đây là một phương pháp nghiên cứu độc lập thông qua việc dự báo theo các chuỗi thời gian. Mô hình ARIMA được nghiên cứu ở đây là một công cụ mạnh, nó thích ứng hầu hết cho chuỗi thời gian có tính dừng và tuyến tính. Sau đó, các nhà nghiên cứu sẽ sử dụng các thuật toán dự báo độ trễ để đưa ra mô hình phù hợp.

Trên thực tế, phân tích chuỗi thời gian từ tập dữ liệu được thống kê về sự lây lan của các bệnh có yếu tố dịch tễ rất hữu ích trong việc phát triển các giả thuyết để giải thích và dự báo về sự lây lan của nó. Có rất nhiều nghiên cứu đã sử dụng mô hình ARIMA để dự báo xu hướng của các bệnh có yếu tố dịch tễ. 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]. Hay 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]. Gần hơn với đại dịch COVID-19 là 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]. Vì vậy, chúng tôi thấy rằng mô hình ARIMA là một công cụ hữu ích trong việc theo dõi và dự báo xu hướng thay đổi trong các bệnh truyền nhiễm.

Hơn nữa, mô hình ARIMA theo mùa được sử dụng rộng rãi để dự báo khí tượng thủy văn. 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 [4]. 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

[5]. Momani sử dụng thành công mô hình ARIMA để dự báo xu hướng lượng mưa của Jordan [6]. 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 [7]. Chính vì vậy, mô hình ARIMA theo mùa có thể được xem là một công cụ hữu ích để dự báo hiện tượng khí tượng thủy văn, đặc biệt là lượng mưa.

Để đơn giản cho việc tính toán và đồ thị hóa, chúng ta cần có sự hỗ trợ của các công cụ phần mềm thống kê hiện đại. Trong nước, hầu hết các bài toán dự báo chuỗi thời gian đều sử dụng các phần mềm thương mại đắt tiền như SPSS, Eviews,... Trong khi đó, R là một công cụ hoàn toàn miễn phí và hỗ trợ đầy đủ các tính năng mà các phần mềm thương mại hiện có. Tuy nhiên, nó chưa được sử dụng rộng rãi tại các trường đại học tại Việt Nam. Chính vì vậy, chúng tôi quyết định cho R như là một công cụ đắc lực cho bài nghiên cứu.

2. Lý do chọn đề tài :

Sử dụng mô hình ARIMA để tiến hành dự báo là một phương pháp hiệu quả và phổ biến. Bên cạnh đó, tính năng và ưu điểm vượt trội của phần mềm R đã và đang thu hút sự quan tâm của các nhà nghiên cứu khi vận dụng vào bài toán thực tiễn ngày càng phức tạp. Vì vậy, việc kết hợp mô hình ARIMA với phần mềm R có rất nhiều ý nghĩa, đó là động lực thúc đẩy nhóm sinh viên lựa chọn và thực hiện đề tài này.

3. Mục tiêu đề tài :

- Về lý thuyết: tìm hiểu lý thuyết cơ bản về chuỗi thời gian và mô hình ARIMA; nắm vững quy trình dự báo bằng mô hình ARIMA gồm 5 bước sau: kiểm tra tính dừng của chuỗi thời gian; chuyển một chuỗi không dừng thành dừng; xác định bậc p, q, P và Q; kiểm tra độ chính xác của mô hình; bước cuối cùng là, dự báo.

- Về thực hành: Nắm vững quy trình xây dựng mô hình ARIMA (với 5 bước nêu trên) bằng phần mềm R cho các dữ liệu thời gian thực trong những lĩnh vực khác nhau như là: khí tượng thủy văn; dịch tễ học; giáo dục; tài chính; chứng khoán.

4. Phương pháp nghiên cứu:

Sinh viên đọc hiểu các tài liệu tham khảo về lý thuyết dự báo, phân tích chuỗi thời gian và tài liệu hướng dẫn sử dụng R; trực tiếp thực hành xây dựng mô hình trên R cho các bộ dữ liệu khác nhau.

5. Đối tượng và phạm vi nghiên cứu:

- Đối tượng nghiên cứu: Phương pháp/công cụ dùng trong phân tích chuỗi thời gian, có hỗ trợ của phần mềm thống kê.

- Phạm vi nghiên cứu: Mô hình ARIMA trên phần mềm R.

Chương 1

Tổng quan về lý thuyết chuỗi thời gian và mô hình ARIMA

Trong Chương 1, chúng tôi giới thiệu tổng quan một số kiến thức về lý thuyết chuỗi thời gian, hồi quy cổ điển của chuỗi thời gian bao gồm các định nghĩa về *hàm trung bình*, *hàm tự hiệp phương sai mẫu*, *hàm tự tương quan mẫu* và tính chất *phân phối mẫu lớn của ACF* (Auto-Correlation Function). Chi tiết các định nghĩa được chúng tôi trình bày trong bản báo cáo tổng kết. Từ các khái niệm và tính chất này, chúng tôi tiến hành xây dựng mô hình ARIMA.

1.1 Mô hình ARIMA

Mô hình tự hồi quy tích hợp trung bình trượt (ARIMA - Autoregressive Integrated Moving Average) là một lớp mô hình tuyến tính có khả năng biểu diễn cả chuỗi thời gian dừng lẫn không dừng. Mô hình ARIMA dựa vào các mẫu tự tương quan trong bản thân của chuỗi thời gian để sinh ra dự báo. Hệ thống các phương pháp dùng để xác định, kiểm tra và cải tiến mô hình ARIMA có sự đóng góp rất lớn của hai nhà thống kê *G. E. P. Box và G. M. Jenkins (1970)*. Do đó, việc mô hình hóa và dự báo dựa trên mô hình ARIMA còn được gọi là *phương pháp luận Box-Jenkins*. Mô hình ARIMA là sự kết hợp giữa 3 quá trình: *quá trình tự hồi quy, quá trình trung bình trượt và sai phân bậc d*. Quy trình xây dựng mô hình ARIMA, các định nghĩa và phương trình tổng quát được chúng tôi trình bày chi tiết trong bản tổng kết.

1.2 Các bước để xây dựng mô hình ARIMA

Bước 1: Kiểm tra tính dừng của chuỗi thời gian: Quan sát biểu đồ chuỗi thời gian, quan sát đồ thị ACF, *kiểm định Dickey-Fuller mở rộng.*
Bước 2: Chuyển một chuỗi không có tính dừng về chuỗi có tính dừng giúp ta chọn được tham số d của mô hình.
Bước 3: Chọn mô hình phù hợp nhất

- Quan sát đồ thị ACF và PACF, chúng ta chọn được tham số p và q hoặc dùng *thuật toán Hyndman-Khandakar (2008)* để xác định các tham số p, d, q một cách tự động.
- Chúng tôi dùng các tiêu chí thông tin AIC, AICc, BIC để chọn mô hình phù hợp nhất.

Bước 4: Kiểm định độ chính xác của mô hình.

- Dựa vào các *thước đo sai số RMSE, MAE, MAPE*, chúng tôi đánh giá được sai số của dự báo từ mô hình.
- Ta kiểm tra xem các phần dư ước lượng từ mô hình có tính dừng hay không. Nếu phần dư có tính dừng với giá trị trung bình bằng 0 và phương sai là 1 thì mô hình là phù hợp; còn nếu không ta lặp lại các bước ở trên cho đến khi tìm được mô hình tốt nhất.

Bước 5: Dự báo.

1.3 Mô hình ARIMA theo mùa

Đối với dữ liệu có tính mùa vụ, *mô hình ARIMA theo mùa (SARIMA - Seasonal Autoregressive Integrated Moving Average)* là một lựa chọn hợp lý. Mô hình ARIMA theo mùa có cấu trúc tương tự như mô hình ARIMA, được hình thành từ các quá trình AR, MA và quá trình tích hợp để đưa chuỗi dữ liệu về chuỗi có tính dừng. Định nghĩa và phương trình tổng quát của mô hình ARIMA theo mùa được chúng tôi trình bày chi tiết trong bản báo cáo tổng kết.

Trong toàn bộ Chương 1, chúng tôi đã trình bày lý thuyết cơ sở của chuỗi thời gian, hồi quy cổ điển của chuỗi thời gian và mô hình ARIMA. Trong chương tiếp theo, chúng tôi sẽ áp dụng kiến thức lý thuyết của mô hình đã giới thiệu để phân tích dữ liệu chuỗi thời gian với sự hỗ trợ của phần mềm R.

Chương 2

Phân tích chuỗi thời gian bằng mô hình ARIMA với phần mềm R

Trong chương này, chúng tôi trình bày tổng quan về lịch sử của R, các gói lệnh chính được sử dụng trong bài báo cáo và chức năng của chúng, thực hành phân tích dữ liệu chuỗi thời gian bằng mô hình ARIMA với sự hỗ trợ của phần mềm R. Trên mỗi tập dữ liệu được phân tích, chúng tôi không chỉ thực hiện xây dựng mô hình ARIMA và phân tích kết quả đầu ra thu được, mà còn sử dụng phần mềm vào việc vẽ các đồ thị cần thiết nhằm trực quan hóa các kết quả phân tích.

2.1 Phân tích và dự báo trên tập dữ liệu COVID-19

Trong phần này, chúng tôi trình bày các bước xử lý số liệu thô, phân tích tình hình COVID-19 trên toàn thế giới bằng các hình ảnh được chúng tôi vẽ được. Áp dụng mô hình ARIMA để dự báo số ca nhiễm mới tại Mỹ và số ca tử vong mới tại Ý. Xuyên suốt bài báo cáo này, code dùng để vẽ hình và xây dựng mô hình ARIMA được tìm thấy tại `https://github.com/nguyenquocduongqnu/NCKH-2020`. Qua mỗi đồ thị, chúng tôi rút ra một số nhận xét được trình bày chi tiết trong bản báo cáo tổng kết.

Chúng tôi lấy dữ liệu toàn cầu dưới dạng thô được thống kê hàng ngày của đại học Johns Hopkins tại `https://github.com/CSSEGISandData/COVID-19`. Phần lớn các nhà thống kê học, dịch tễ học và các nhà nghiên cứu đều dùng bộ dữ liệu này để phân tích.

2.1.1 Dự báo số ca nhiễm mới COVID-19 tại nước Mỹ

Trong phần này, chúng tôi sẽ phân tích tình hình diễn biến dịch bệnh tại Mỹ và dự báo số ca nhiễm mới của quốc gia này trong 7 ngày tiếp theo. Dữ liệu được trích xuất từ ngày 22/1/2020 đến ngày 4/4/2020. Các bước đào tạo và xây dựng mô hình được chúng tôi trình bày chi tiết trong bản báo cáo tổng kết. Bảng 2.1 đánh giá sai số từ

Bảng 2.1: Đánh giá sai số từ mô hình ARIMA(0,2,3)

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%

mô hình ARIMA(0, 2, 3), chúng tôi rút ra một số nhận xét như sau:

(i) Phần trăm sai số giữa các giá trị dự báo và giá trị thực tế tương đối nhỏ.

(ii) Giá trị dự báo thấp hơn so với giá trị thực tế. Điều này cho chúng ta thấy rằng số ca nhiễm mới tại Mỹ đang tăng rất nhanh và có nguy cơ sẽ dẫn đến khủng hoảng trên diện rộng trong vài ngày tới. Vì vậy các giá trị dự báo được đánh giá là phù hợp và có ý nghĩa.

Do đó, mô hình này có thể là mô hình có ích để đưa ra một cảnh báo cấp thiết để các cấp chính quyền và người dân có hành động cao hơn nữa nhằm kiểm soát tốt dịch bệnh trong thời gian tới. Số ca nhiễm mới và đồ thị dự báo tại Mỹ trong 7 ngày tiếp theo được chúng tôi trình bày chi tiết trong bản báo cáo tổng kết. Từ kết quả dự báo, chúng tôi sẽ đưa ra một cảnh báo cho 7 ngày tiếp theo (5/4/2020 - 11/4/2020) rằng số ca nhiễm tại nước Mỹ vẫn tiếp tục tăng cao, thậm chí vượt khỏi giới hạn trên của khoảng dự báo 95% nếu Mỹ không đưa ra các biện mạnh hơn để kiểm soát.

2.1.2 Dự báo số ca tử vong mới theo ngày do COVID-19 gây ra ở Italy

Phân tích nguyên nhân dẫn đến số ca tử vong tăng vọt tại Ý được chúng tôi trình bày chi tiết trong bản báo cáo tổng kết. Tiếp theo, chúng tôi sẽ dự báo số ca tử vong mới tại quốc gia này trong 7 ngày tiếp theo bằng mô hình ARIMA.

Dữ liệu được phân tích từ ngày 22/01/2020 đến ngày 07/04/2020. Các bước đào tạo và xây dựng mô hình được chúng tôi trình bày chi tiết trong bản báo cáo tổng kết.

Quan sát bảng 2.2, chúng ta thấy được phần trăm sai số từ mô hình ARIMA(3, 1, 3)

Bảng 2.2: Đánh giá sai số dự báo số ca tử vong mới tại Italy

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%

rất nhỏ. Vì vậy mô hình này có thể dùng để dự báo tổng quan cho 7 ngày tiếp theo tại Italy. Số ca tử vong mới và đồ thị dự báo tại Italy trong 7 ngày tiếp theo được chúng tôi trình bày chi tiết trong bản báo cáo tổng kết. Từ kết quả dự báo, chúng ta thấy số ca tử vong mới ở Italy có xu hướng giảm trong tuần tới. Tuy nhiên con số tử vong vẫn còn rất cao. Do đó, chính phủ Italy không nên chủ quan mà cần nỗ lực hơn nữa để đưa ra kịch bản kiểm soát dịch và phác đồ điều trị tối ưu nhất.

2.2 Dự báo lượng mưa hàng tháng tại trạm quan trắc Quy Nhơn

Trong phần này, chúng tôi sẽ phân tích tổng quan lượng mưa tại trạm quan trắc Quy Nhơn và dự báo lượng mưa trung bình hàng tháng cho 2 năm tiếp theo (2019-2020) bằng *mô hình mô hình tự hồi quy tích hợp trung bình trượt theo mùa* (SARIMA).

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. Các bước xây dựng mô hình ARIMA theo mùa được chúng tôi trình bày chi tiết trong bài báo cáo tổng kết. Sai số từ mô hình ARIMA$(1,0,0)(2,1,1)_{12}$ được mô tả trong bảng 2.3.

Các giá trị dự báo cho mùa mưa (tháng 9 đến tháng 12) cho kết quả sai số rất thấp. Điều này phù hợp với đặc điểm lượng mưa tại Quy Nhơn. Do đó, mô hình ARIMA$(1, 0, 0)(2, 1, 1)_{12}$ có thể dùng dự báo lượng mưa tại Quy Nhơn. Vì vậy, chúng tôi sử dụng mô hình ARIMA$(1, 0, 0)(2, 1, 1)_{12}$ để dự báo lượng mưa hàng tháng cho 2 năm tiếp theo. Từ kết quả dự báo này, chúng tôi hy vọng sẽ giúp ích trong việc hoạch định và điều tiết hệ thống tài nguyên nước.

Bảng 2.3: Đối sánh giữa kết quả dự báo và tập test cho lượng mưa

	Point Actual(A)	Point Forecast (F)	Lo 80	Hi 80	Lo 95	Hi 95	$\frac{A-F}{A}.100\%$
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 Webstie Dashboard COVID-19

Nhằm góp phần vào việc kiểm soát dịch, chúng tôi đã tạo ra một website hỗ trợ quan sát xu hướng dịch bệnh COVID-19 cho từng quốc gia bằng mô hình ARIMA với package "shiny" trong R. Link: `https://nguyenquocduong.shinyapps.io/NCKH`.

2.4 So sánh mô hình ARIMA với mô hình NNAR

Tập dữ liệu được chúng tôi sử dụng để đánh giá mô hình là dữ liệu lượng mưa tại trạm quan trắc Quy Nhơn. Đối với tập dữ liệu này, mô hình ARIMA thể hiện khả năng dự báo tốt hơn mô hình NNAR (Neural Network Auto-Regression). Chi tiết quy trình thực hiện so sánh được chúng tôi trình bày trong bản báo cáo tổng kết. Vì lí do thời gian còn hạn chế, chúng tôi chỉ so sánh giữa ARIMA và NNAR. Trong thời gian tới, chúng tôi sẽ mở rộng so sánh thêm với một số mô hình hiện đại như LSTM (Long Short Term Memory), ARIMA-LSTM, ARIMA-ANN, ...

Kết luận chương 2

Trong Chương 2, chúng tôi đã sử dụng các gói lệnh trong R để phân tích tổng quan dữ liệu COVID-19, dự báo số ca nhiễm mới tại Mỹ, dự báo số ca tử vong mới ở Italy và dự báo lượng mưa hàng tháng tại trạm quan trắc Quy Nhơn. Qua từng bộ dữ liệu được phân tích, chúng tôi thấy được mô hình ARIMA đã cho khả năng dự báo rất tốt. Hơn nữa, chúng tôi đã khai thác rất hiệu quả tính năng của R để xây dựng được website Dashboard COVID-19 nhằm góp phần vào công cuộc chống đại dịch toàn cầu.

Thảo luận

Những mặt thuận lợi và hạn chế của mô hình ARIMA

Dự báo chính xác là một nhiệm vụ quan trọng nhưng thường là khó khăn đối với các nhà hoạch định chính sách trong nhiều lĩnh vực. Mặc dù có rất nhiều mô hình được ứng dụng trong việc dự báo nhưng mỗi mô hình đều có thuận lợi và hạn chế riêng. Do đó, chúng ta cần nắm bắt được những điểm thuận lợi và hạn chế của mô hình được đưa ra.

Những điểm thuận lợi khi sử dụng ARIMA

- 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.
- ARIMA phát huy thế mạnh trong việc sử dụng để dự báo tài chính, chứng khoán, kinh tế lượng, khí tượng thủy văn, ...
- ARIMA thích hợp cho các bài toán dự báo ngắn hạn.
- 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 chế khi sử dụng ARIMA

- Trong ARIMA, một cấu trúc tương quan tuyến tính được giả định giữa các giá trị trong chuỗi thời gian. Do đó, đối với dữ liệu có tính chất phi tuyến thì mô hình ARIMA không phát hiện được.
- 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 xu hướng dịch bệnh cho mỗi quốc gia bằng mô hình ARIMA với 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 như: dịch tễ học, y tế cộng đồng, khí tượng thủy văn, ...

Trong thời gian tới, chúng tôi mong muốn thực hiện các nghiên cứu về khoa học dữ liệu với R (data science, mô hình phân tích sống sót) trên số liệu thực gắn với lĩnh vực Y sinh, sức khỏe cộng đồng và đời sống kinh tế-xã hội của địa phương, trong nhiều lĩnh vực khác nhau, để có thể mang lại những ứng dụng thiết thực cho tỉnh Bình Định.

Tài liệu tham khảo

[1] 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.

[2] 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.

[3] 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.

[4] M. Rahman, Islam AHMS, Nadvi SYM, R. M. Rahman, ***Comparative study of ANFIS and ARIMA model for weather forecasting in Dhaka***. Informatics, Electronics & Vision (ICIEV), 2013 International Conference on, Dhaka; 2013. p. 1-6.

[5] Dizon CQ, ***ARIMA modeling of a stochastic process appropriate for the angat reservoir***. Philipp. Eng. J. 2007;28:1-20

[6] Momani PENM, ***Time series analysis model for rainfall data in Jordan: Case study for using time series analysis***, Am. J. Environ. Sci. 2009;5:599-604.

[7] Nguyễn Hữu Quyền, ***Nghiên cứu ứng dụng mô hình ARIMA để dự báo lượng mưa vụ đông xuân ở một số tỉnh vùng Đồng bằng Bắc Bộ***, Luận văn thạc sĩ khoa học, 2013.

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

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

[10] R. J. Hyndman and G. Athanasopoulos, ***Forecasting: Principles and Practice,*** 2^{nd} ***edition***, OTexts Publisher, USA(2018).

[11] R. Krispin, ***Hands-On Time Series Analysis with R***, Packt Publisher, UK(2019).

[12] Website:https://nguyenquocduong.shinyapps.io/NCKH/.