

ỨNG DỤNG ARIMA VÀ SARIMA TRONG DỰ BÁO GIÁ VÀNG

● NGUYỄN THỊ HỒNG THẢO - ĐÀO MINH CHÂU

TÓM TẮT:

Vàng là kim loại quý hiếm, là mặt hàng trao đổi đóng vai trò quan trọng trong nền kinh tế. Vì vậy, dự báo giá vàng là một điều rất cần thiết trong kinh doanh, đầu tư, cũng như trong quản lý nhà nước. Nhóm tác giả đã tìm hiểu mô hình dự báo ARIMA (Autoregressive Integrated Moving Average - mô hình tự hồi quy tích hợp trung bình trượt) và SARIMA (Seasonal Autoregressive Integrated Moving Average - mô hình ARIMA theo mùa), là phương pháp dự báo dữ liệu theo chuỗi thời gian để dự báo giá vàng trong tương lai. Các tác giả đã sử dụng dữ liệu thực tế giá vàng trong 4 năm để phân tích và đánh giá, tìm ra kết quả dự báo tốt nhất của giá vàng.

Từ khóa: mô hình ARIMA, mô hình SARIMA, dự báo giá vàng, giá vàng.

1. Đặt vấn đề

Trong những năm gần đây, việc nghiên cứu dự báo đã được ứng dụng rộng rãi, như: dự đoán kết quả phân tích dự báo đầu tư, đánh giá kết quả phát triển trong một số lĩnh vực,... và được sử dụng trong tiến trình khám phá tri thức. Dự báo mang đến nhiều lợi ích, giúp xử lý rất nhiều thông tin đến từ nhiều nguồn khác nhau và dự báo các thông tin đó; ở những nơi không có chuyên gia, có thể giúp tạo ra được các quyết định từ các dữ liệu có được; các thuật toán có thể giúp xử lý khi dữ liệu không đầy đủ, không chính xác,...

Do dự báo mang đến nhiều lợi ích cho con người, nên nhóm tác giả chọn nghiên cứu vài bài toán dự báo theo chuỗi thời gian, sử dụng mô hình ARIMA và SARIMA để dự báo giá trong tương lai. Dữ liệu để phân tích và dự báo là giá vàng. Vàng là một trong những mặt hàng quan trọng

đánh giá chỉ số lạm phát của một quốc gia. Do đó, dự báo về giá vàng rất cần thiết. Nhóm tác giả đã tiến hành hàng loạt kiểm định và đánh giá mô hình, cũng như phân tích nhân tố biến động theo chuỗi thời gian để tìm ra kết quả tối ưu nhất, nhằm gia tăng hiệu suất dự báo.

2. Cơ sở lý thuyết và phương pháp nghiên cứu

2.1. Các công trình có liên quan

Hiện nay, với sự phát triển dự báo được thuận tiện hơn và có những kết quả dự báo cơ bản đáp ứng được mục đích của từng vấn đề. Nhóm tác giả trích dẫn một số công trình nghiên cứu như sau. Nghiên cứu ảnh hưởng tin tức kinh tế đến giá cả hàng hóa, trong đó có giá vàng trong công trình 7 common factors that influence gold prices của Williams S. (2018) [2]. Một số công trình nghiên cứu dự báo giá vàng bằng cách sử dụng các thuật

toán máy học, như: Predicting future gold rates using machine learning approach của Sami I. (2017) [3], Time series forecasting using a hybrid ARIMA and neural network model [4], Application of SARIMA Model in Cucumber Price Forecast. Applied Mechanics and Materials [5], Gold price forecasting research based on an improved online extreme learning machine algorithm [6], HolteWinters smoothing enhanced by fruit fly optimization algorithm to forecast monthly electricity consumption [7].

Các công trình đã đưa ra được một số mô hình dự báo giá vàng, cũng như giá một số mặt hàng khác bằng các thuật toán máy học. Đối với công trình nghiên cứu giá vàng thường cho kết quả chính xác với giá trị thực nghiệm, nhưng khi thử với những giá trị khác thì cho kết quả có độ sai lệch. Tuy nhiên, kết quả các công trình nghiên cứu này là cơ sở để đánh giá, so sánh thuật toán ARIMA và SARIMA trong dự báo giá vàng.

2.2. Cơ sở lý thuyết

2.2.1. Phương pháp ARIMA

Phương pháp ARIMA là mô hình dự báo theo chuỗi thời gian được phát minh bởi Box-Jenkin, kết quả dự báo phụ thuộc chuỗi giá trị trong quá khứ. Phương pháp ARIMA phân tích định lượng tính tương quan giữa các dữ liệu quan sát để đưa ra mô hình dự báo thông qua các giai đoạn nhận dạng mô hình, ước lượng các tham số từ dữ liệu quan sát và kiểm tra các tham số ước lượng để tìm ra mô hình thích hợp. Phương pháp ARIMA được thực hiện cho dữ liệu dừng (hay dữ liệu đã loại bỏ yếu tố xu thế). ARIMA được kết hợp bởi 3 thành phần chính ARIMA(p,q,d):

AR (p): Autogressive (tự hồi quy)

I (q): Integrated (chuỗi dừng)

MA (d): Moving Average (trung bình động)

Tiến trình xây dựng thuật toán ARIMA gồm có 6 bước:

Bước 1: Tính ACF và PACF của dữ liệu gốc để kiểm tra xem chuỗi gốc có dừng hay không. Nếu dừng, chuyển sang bước 3.

Bước 2: Lấy log rồi lấy sai phân bậc một của

dữ liệu gốc, hoặc lấy sai phân trực tiếp nếu dữ liệu ít biến động, sau đó tính ACF và PACF của dữ liệu chuyển đổi.

Bước 3: Phân tích giản đồ tự tương quan để xác định các mô hình có thể có.

Bước 4: Ước lượng các mô hình dự kiến.

Bước 5: Đối với mỗi mô hình được ước lượng:

Kiểm tra hệ số của độ trễ cao nhất có ý nghĩa thống kê không. Nếu không giảm bớt độ trễ của p và q .

Kiểm tra ACF và PACF đối với phần dư.

So sánh các sai số dự báo.

Phân tích đồ thị phần dư.

Phân tích đồ thị giá trị dự báo và giá trị thực tế.

Khi đánh giá các mô hình ARIMA, nên so sánh giữa các mô hình.

Bước 6: Nếu có thay đổi so với mô hình gốc, quay lại bước 4.

Phương pháp ARIMA được áp dụng đối với những chuỗi thời gian dừng. Hầu hết, dữ liệu chuỗi thời gian là các chuỗi có xu thế (giá trị trung bình của Y_t trong năm này có thể khác năm kia hoặc các chuỗi không dừng). Để áp dụng được mô hình ARIMA thì các chuỗi thời gian phải khử các yếu tố xu thế trong các chuỗi dữ liệu gốc thông qua quy trình lấy sai phân.

2.2.2. Phương pháp SARIMA

Phương pháp SARIMA được phát triển từ phương pháp ARIMA khi chuỗi dữ liệu có tính mùa vụ. Các độ trễ theo mùa của y_t trong mô hình ARIMA để trở thành mô hình SARIMA. Mô hình SARIMA có dạng tổng quát: $(p,d,q)(P,D,Q)_s$, trong đó, P là AR và Q là MA, D là bậc sai phân có tính mùa, s là số kỳ trong một vòng chu kỳ ($s = 12$ khi chuỗi dữ liệu theo tháng). Phân tích ACF và PACF tại các trễ là bội số của độ dài mùa s giúp kết luận các giá trị P, Q phù hợp cho mô hình. Đối với thành phần mùa MA, biểu đồ ACF cho thấy một đỉnh nhọn ở các trễ mùa, còn đối với thành phần mùa AR thì biểu đồ PACF thể hiện đỉnh nhọn này.

Tiến trình xây dựng thuật toán SARIMA thực hiện qua 6 bước sau:

Bước 1: Kiểm định tính dừng, tính mùa vụ của dữ liệu gốc. Các chuỗi được kiểm định tính dừng của dữ liệu thời gian bằng phương pháp kiểm định đơn vị (unit root test) - kiểm định Dickey-Fuller (ADF).

Bước 2: Chuyển dữ liệu sang sai phân. Việc chuyển đổi được thực hiện bằng cách tính sai phân giữa các giá trị quan sát dựa vào giả định các phần khác nhau của các chuỗi thời gian. Nghiên cứu thực nghiệm khi chuyển chuỗi dữ liệu sang sai phân nên dừng lại sai phân bậc 2.

Bước 3: Ước lượng các giá trị tham số của mô hình sau khi chuyển thành chuỗi sai phân bậc d để đạt tính dừng.

Bước 4: Xây dựng mô hình SARIMA.

Bước 5: Lựa chọn mô hình và kiểm định mô hình các tham số của mô hình sẽ được ước lượng bằng MLE.

Bước 6: Dự báo sau khi kiểm định sai số, nếu mô hình là phù hợp, sẽ được sử dụng vào việc dự báo. Nghiên cứu tính chỉ tiêu % sai số của dự báo so với giá trị quan sát được và được tính bằng sự chênh lệch giữa giá trị quan sát so với giá trị dự báo.

2.3. Phương pháp nghiên cứu

Phương pháp định lượng: Dự báo chuỗi thời gian dựa trên phân tích giá vàng và thời gian thay đổi theo từng ngày.

Thu thập và xử lý số liệu số: chuỗi dữ liệu sử dụng 1118 dòng quan sát từ ngày 02/01/2019 đến ngày 25/8/2022. Mỗi dòng quan sát bao gồm giá vàng và thời gian liên tục mỗi ngày lưu trên phần mềm Excel.

Nguồn số liệu: Lấy từ nguồn dữ liệu thực tế truy cập tại: <https://www.tradingview.com>

3. Kết quả thực nghiệm và thảo luận

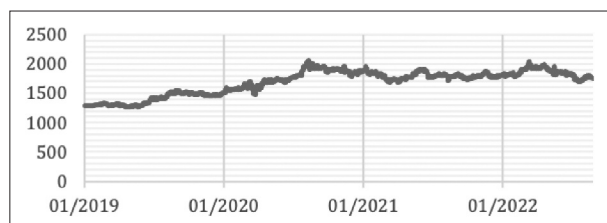
3.1. Phương pháp ARIMA

Với các phương pháp đề xuất, tiến hành thử nghiệm trên chuỗi dữ liệu gồm 1118 dòng quan sát từ ngày 02/01/2019 đến ngày 25/8/2022, được lấy từ các nguồn dữ liệu thực tế thông qua trang web Trading View (www.tradingview.com). Dự đoán chuỗi thời gian dựa trên phân tích giá vàng

và thời gian thay đổi từng ngày. Mỗi dòng quan sát bao gồm giá vàng và thời gian liên tục trong ngày được lưu trữ trong phần mềm Excel.

Các kết quả phân tích dữ liệu dựa trên phần mềm Rstudio. Trước khi tiến hành phân tích dữ liệu, biểu diễn biểu đồ giá vàng trong thời gian quan sát. (Hình 1)

Hình 1: Biểu đồ giá vàng từ ngày 02/01/2019 đến ngày 25/8/2022



Nguồn: www.tradingview.com

Kết quả biểu đồ giá vàng Hình 1, ta thấy giá vàng là chuỗi dữ liệu không dừng vì có yếu tố xu hướng và yếu tố mùa vụ. Để kiểm định lại nhận định trên, cần tiến hành phân tích biểu đồ giảm tự tương quan ACF của tập dữ liệu giá vàng. (Hình 2)

Dựa vào biểu đồ tương quan tính độ trễ, hệ số ACF của tập dữ liệu giảm rất chậm và liên tục nằm ngoài giới hạn tin cậy. Kết luận dữ liệu giá vàng là chuỗi không dừng.

Do dữ liệu giá vàng là chuỗi không dừng, nên cần loại bỏ tính xu hướng và mùa vụ của tập dữ liệu bằng cách tiến hành lấy sai phân hoặc log trước khi vận dụng phương pháp ARIMA.

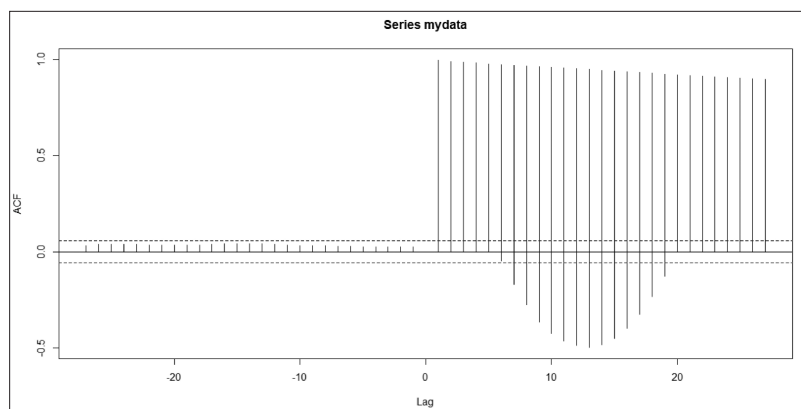
Hình 3 biểu diễn biểu đồ giá vàng sau khi lấy sai phân bậc 1, tất cả các giá trị dao động xung quanh giá trị $[-100, 100]$, có vài điểm nằm ngoài khoảng trên. Sau khi lấy sai phân lần 1 dữ liệu có thể có tính dừng.

Biểu đồ hàm tự tương quan sai phân bậc 1 của chuỗi dữ liệu giá vàng cho thấy chuỗi dữ liệu sau khi lấy sai phân đã loại bỏ thuộc tính mùa vụ và xu hướng, như vậy đã đạt chuỗi dừng. Kết luận $q=1$.

Biểu đồ PACF có độ trễ là 1. Vậy $p=1$.

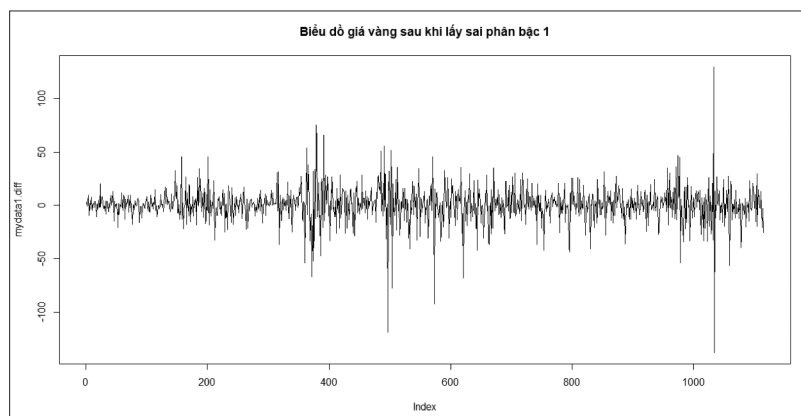
Một số mô hình ARIMA có thể ứng với các giá trị $p=1, q=1, d=1$.

Hình 2: Biểu đồ tương quan của dữ liệu vàng (từ ngày 02/1/2019 đến ngày 25/8/2022)



Nguồn: www.tradingview.com

Hình 3: Biểu đồ giá vàng sau khi lấy sai phân bậc 1



Nguồn: Nhóm tác giả thực hiện

Hình 4: So sánh AIC của các mô hình ARIMA đề xuất với dữ liệu vàng

```
Fitting models using approximations to speed things up...
ARIMA(2,1,2) with drift      : 5337.634
ARIMA(0,1,0) with drift     : 5335.344
ARIMA(1,1,0) with drift     : 5338.015
ARIMA(0,1,1) with drift     : 5337.283
ARIMA(0,1,0)                : 5335.202
ARIMA(1,1,1) with drift     : 5339.947

Now re-fitting the best model(s) without approximations...
ARIMA(0,1,0)                : 5340.798

Best model: ARIMA(0,1,0)

Series: seriesdata2
ARIMA(0,1,0)

sigma^2 estimated as 269.4: log likelihood=-2669.4
AIC=5340.79  AICc=5340.8  BIC=5345.24
```

Nguồn: Nhóm tác giả thực hiện

ARIMA (2,1,2), ARIMA (1,1,0), ARIMA (0,1,1), ARIMA (0,1,0), ARIMA (1,1,1).

Tiếp tục kiểm tra và so sánh các mô hình ARIMA để tìm mô hình ARIMA tối ưu nhất. (Hình 4)

Theo như kết quả so sánh giá trị của chỉ tiêu AIC của các mô hình đề xuất, thì mô hình thứ 5 (ARIMA(0,1,0)) là mô hình có giá trị AIC thấp nhất, điều này có nghĩa là mô hình này phù hợp với tập dữ liệu hơn so với các mô hình còn lại. (Hình 5)

Biểu đồ dự báo Hình 5 ứng với mô hình ARIMA tối ưu nhất, ARIMA(0,1,0). Nhóm tác giả đã so sánh giá trị dự báo và giá trị thực cho mô hình ARIMA. Giá trị dự báo của giá vàng là phần đồ thị thể hiện dấu +.

3.2. Phương pháp SARIMA (Hình 6)

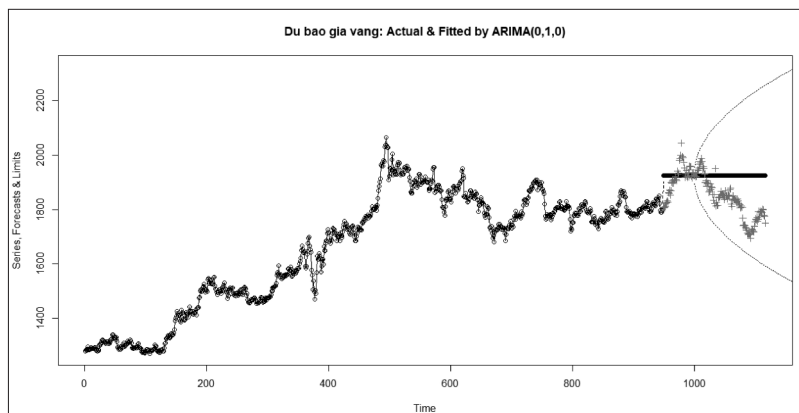
Dựa vào phần biểu đồ tương quan ACF ở trên thì biểu đồ thể hiện biến động tăng và giảm ở các độ trễ nằm trong giới hạn tin cậy, ngoại trừ độ trễ thứ 2 có giá trị của ACF gần với giới hạn tin cậy 95%. Biểu đồ hàm tự tương quan sai phân bậc 2 của chuỗi dữ liệu này cho thấy chuỗi dữ liệu sau khi lấy sai phân bậc 2 đã loại bỏ thuộc tính mùa vụ và xu hướng, như vậy đã đạt chuỗi dừng. Kết luận $q+Q=1$.

Phần biểu đồ tương quan riêng phần PACF, tại độ trễ 2 nằm ngoài phạm vi tin cậy. Chọn $p+P=1$.

Một số mô hình có thể ứng với các giá trị $p+P=1, q+Q=1, d+D=1$

Mô hình SARIMA

Hình 5: Biểu đồ so sánh giá trị thực tế và giá trị dự báo theo mô hình ARIMA (0,1,0)



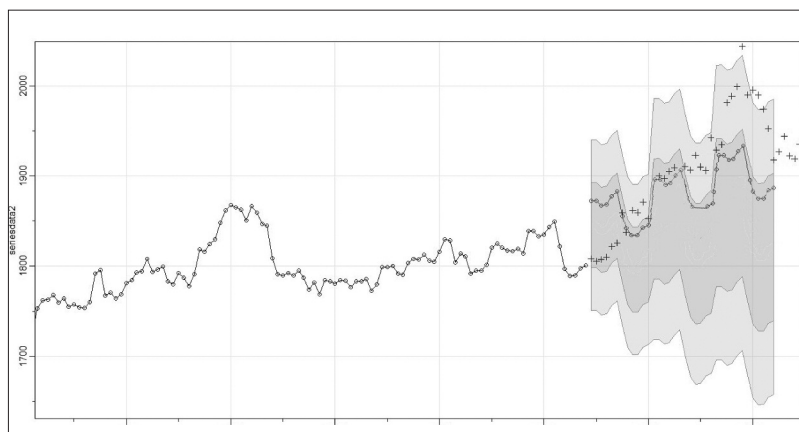
Nguồn: Nhóm tác giả thực hiện

Hình 6: So sánh AIC của các mô hình SARIMA đề xuất với dữ liệu vàng

ARIMA(2,0,2)(0,1,0)[12] with drift	: Inf
ARIMA(0,0,0)(0,1,0)[12] with drift	: 87.51252
ARIMA(1,0,0)(0,1,0)[12] with drift	: 89.77352
ARIMA(0,0,1)(0,1,0)[12] with drift	: 89.09006
ARIMA(0,0,0)(0,1,0)[12]	: 84.7474
ARIMA(1,0,1)(0,1,0)[12] with drift	: Inf
Best model: ARIMA(0,0,0)(0,1,0)[12]	
Series: Train	
ARIMA(0,0,0)(0,1,0)[12]	
sigma^2 estimated as 57.6: log likelihood=-41.17	
AIC=84.35 AICc=84.75 BIC=84.83	

Nguồn: Nhóm tác giả thực hiện

Hình 7: Biểu đồ so sánh giá trị thực tế và giá trị dự báo theo mô hình SARIMA (0,0,0)(0,1,0)₁₂



Nguồn: Nhóm tác giả thực hiện

(2,0,2)(0,1,0)₁₂, Mô hình SARIMA (0,0,0)(0,1,0)₁₂, Mô hình SARIMA (1,0,0)(0,1,0)₁₂, Mô hình SARIMA (0,0,1)(0,1,0)₁₂, Mô hình SARIMA (1,0,1)(0,1,0)₁₂

Theo như kết quả so sánh giá trị của chỉ tiêu AIC của 5 mô hình đề xuất, thì mô hình thứ 2 (tức là mô hình SARIMA (0,0,0)(0,1,0)₁₂) có giá trị AIC thấp nhất, điều này có nghĩa mô hình này phù hợp với tập dữ liệu so với các mô hình còn lại.

Biểu đồ dự báo Hình 7 ứng với mô hình SARIMA tối ưu nhất, SARIMA (0,0,0)(0,1,0)₁₂. Nhóm tác giả đã so sánh giá trị dự báo và giá trị thực cho mô hình ARIMA. Giá trị dự báo của giá vàng là phần thể hiện dấu +.

3.3. Đánh giá và so sánh các kết quả thực nghiệm

Nhóm tác giả đã so sánh các mô hình này bằng cách sử dụng cùng một tập huấn luyện và tập dữ liệu kiểm tra trong cùng một môi trường hoạt động. Dữ liệu được chọn cho thử nghiệm này giá vàng, là dữ liệu giao dịch hàng ngày tính từ ngày 02/01/2019 đến ngày 25/8/2022 được lấy từ cơ sở dữ liệu giao dịch thực tế. Mỗi phần dữ liệu bao gồm giá đóng cửa, mở cửa, giá cao nhất, thấp nhất. Từ kết quả thực nghiệm, nhóm tác giả đã tìm thấy mô hình tối ưu cho ARIMA là ARIMA (0,1,0) và SARIMA là SARIM (0,0,0)(0,1,0)₁₂.

Trong đó:

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

Bảng 1. Đánh giá và so sánh kết quả dự báo của các phương pháp

Mô hình	MAE	MSE	RMSE	MAPE	MPE
<i>Dữ liệu giá vàng từ ngày 02/01/2019 đến ngày 10/06/2021</i>					
ARIMA(0,1,0)	70.26160714	6911.57365357	83.13587465	0.03993752	-0.03185311
SARIMA(0,0,0)(0,1,0) ₁₂	58.56428571	6005.85447714	77.49744820	0.03351040	-0.02915437

Nguồn: Nhóm tác giả thực hiện

MSE: Sai số bình phương trung bình (Mean Squared Error); **RMSE:** Sai số bình phương trung bình gốc (Root Mean Squared Error);

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

MPE: Sai số phần trăm trung bình (Mean Percentage Error); là các thước đo độ chính xác của một mô hình dự báo.

Đối với tập dữ liệu giá vàng tính từ ngày 02/01/2019 đến ngày 25/8/2022 được sử dụng để dự báo theo các mô hình ARIMA và SARIMA, theo như kết quả đánh giá dự báo mô hình SARIMA(0,0,0)(0,1,0)₁₂ có độ chính xác cao hơn, phù hợp để dự báo giá vàng tính từ ngày 02/01/2019 đến ngày 25/8/2022. Từ đó cho thấy dự báo theo mô hình SARIMA là hiệu quả hơn so với mô hình ARIMA. (Bảng 1)

4. Kết luận

Qua việc áp dụng bài toán dự báo dữ liệu chuỗi

thời gian sử dụng mô hình dự báo ARIMA và SARIMA để dự báo giá vàng, nhóm tác giả đã sử dụng dữ liệu nghiên cứu được thu thập từ thực tiễn để huấn luyện và thử nghiệm, nhằm tăng độ chính xác cho quá trình phân tích dự báo một cách đầy đủ và liên tục theo chu kỳ 4 năm. Theo kết quả thực nghiệm tiến hành một loạt thử nghiệm và đánh giá hiệu quả của 2 mô hình ARIMA và SARIMA: mô hình ARIMA (0,1,0) hợp lý nhất để dự báo giá vàng đối với phương pháp ARIMA, mô hình SARIMA (0,0,0)(0,1,0)₁₂ là mô hình cho kết quả tốt nhất so với các mô hình SARIMA. Đối với phương pháp SARIMA, kết quả đánh giá thực nghiệm nghiên cứu dữ liệu giá vàng cho thấy phương pháp SARIMA tối ưu hơn phương pháp ARIMA. Kết quả dự báo này có thể ứng dụng vào thực tế và các phương pháp này có thể được nghiên cứu và mở rộng để dự báo các loại hàng hóa khác trên thị trường ■

TÀI LIỆU THAM KHẢO:

1. Vũ Thanh Nguyên (2004). Ứng dụng các thuật toán máy học trong phân tích dự báo. *Tạp chí Khoa học và Công nghệ, Đại học Quốc gia Thành phố Hồ Chí Minh*, số 1, tháng 2/2004.
2. Williams S (2018). 7 common factors that influence gold prices. Available at: <https://www.fool.com/investing/2016/10/13/7-common-factors-that-influence-gold-prices.aspx>.
3. Sami I., Junejo K.N (2017). Predicting future gold rates using machine learning approach. *International Journal of Advanced Computer Science and Applications*, 8(12).
4. G. Peter Zhang (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50, 159-175.
5. Changshou Luo, Liying Zhou, Qingfeng Weil (2013). Application of SARIMA Model in Cucumber Price Forecast. *Applied Mechanics and Materials*, 373-375, 1686-1690.

6. Futian Weng, Yin hao Chen, Zheng Wang, Muzhou Hou, Jianshu Luo, Zhongchu Tian (2020). Gold price forecasting research based on an improved online extreme learning machine algorithm. *Journal of Ambient Intelligence and Humanized Computing*, 11, 4101- 4111.
7. Weiheng Jiang, Xiaogang Wu, Yi Gong, Wanxin Yu, Xinhui Zhong (2020). HolteWinters smoothing enhanced by fruit fly optimization algorithm to forecast monthly electricity consumption. *Energy*, 193, 116779.

Ngày nhận bài: 14/11/2022

Ngày phản biện đánh giá và sửa chữa: 14/12/2022

Ngày chấp nhận đăng bài: 25/12/2022

Thông tin tác giả:

1. NGUYỄN THỊ HỒNG THẢO¹

2. ĐÀO MINH CHÂU^{1*}

¹Trường Đại học Công nghiệp Thực phẩm Thành phố Hồ Chí Minh

^{*}Tác giả liên hệ

USING ARIMA AND SARIMA MODELS TO FORECAST THE GOLD PRICE

● **NGUYEN THI HONG THAO¹**

● **DAO MINH CHAU^{1*}**

¹ Ho Chi Minh City University of Food Industry

^{*} *Corresponding author*

ABSTRACT:

Gold is a rare and precious metal and it plays an important role in the economy. Hence, it is necessary for businesses, investors, and also government units to forecast the gold price. This study explores the use of ARIMA model and SARIMA model to predict the gold price. These models are based on the time series data forecasting methods. This study uses the actual gold price over four years to analyze and assess these two price forecasting models in order to find the best prediction of gold price.

Keywords: ARIMA model, SARIMA model, gold price prediction.