Ứng dụng mô hình SARIMA dự báo lượng khách quốc tế đến Việt Nam tới năm 2020

NGHIÊM PHÚC HIẾU*

Tóm tắt

Bài viết sử dụng phương pháp Box-Jenkins để xây dựng mô hình ARIMA theo mùa (hay còn gọi là SARIMA - Seasonal Autoregressive Integrated Moving Average) nhằm dự báo lượng khách quốc tế đến Việt Nam dựa trên số liệu công bố hàng tháng của Tổng cục Du lịch Việt Nam. Bài viết cũng đưa ra dự báo thử nghiệm về lượng khách quốc tế đến Việt Nam những tháng cuối năm 2017 với mức độ sai số chấp nhận được từ 1.7% đến 12.4%. Trên cơ sở thử nghiệm này, tác giả tiến hành dự báo lượng khách quốc tế đến Việt Nam trong ba năm tới.

Từ khóa: *dự báo, khách quốc tế, SARIMA*

Summary

With monthly published data from Vietnam National Administration of Tourism, the paper applies Box-Jenkins method to develop SARIMA model so as to predict international visitors to Vietnam. It also forecasts the number of international visitors to Vietnam in the last months of 2017 with acceptable range from 1.7% to 12.4%. Based on this test, the author makes a prediction about international visitors to Vietnam in the next three years.

Keywords: *forecast, international visitors, SARIMA*

GIỚI THIỆU

Với những lợi thế đặc biệt về vị trí địa lý kinh tế và chính trị, Việt Nam có rất nhiều thuận lợi để phát triển du lịch. Nằm ở trung tâm Đông Nam Á, lãnh thổ Việt Nam vừa gắn liền với lục địa vừa thông ra đại dương, có vị trí giao lưu quốc tế thuận lợi cả về đường biển, đường sông, đường sắt, đường bộ và hàng không. Đây là tiền đề rất quan trọng trong việc mở rộng và phát triển du lịch quốc tế.

Để khai thác có hiệu quả các tiềm năng du lịch, tạo dấu ấn tốt trong lòng du khách, khắc phục những rủi ro trong kinh doanh dịch vụ du lịch, lên kế hoạch cho những chặng đường phát triển bền vững tiếp theo, bài viết xây dựng mô hình SARIMA (tức ARIMA theo mùa) phù hợp để dự báo lượng khách quốc tế đến Việt Nam thời gian tới. Trên cơ sở kết quả nghiên cứu, bài viết cũng đưa ra một số hàm ý chính sách để giúp du lịch Việt Nam "cất cánh" trong thời gian tới.

CƠ SỞ LÝ THUYẾT VÀ PHƯƠNG PHÁP NGHIÊN CỨU

Cơ sở lý thuyết

Hai tác giả George Box và Gwilym Jenkins (1976) đã nghiên cứu mô hình tự hồi quy tích hợp trung bình trượt (Autoregressive Integrated Moving Average), viết tắt là ARIMA. ARIMA được kết hợp bởi 3 thành phần chính: AR (thành phần tự hồi quy), I (tính dừng của chuỗi thời gian) và MA (thành phần trung bình trượt).

Mô hình tự tương quan bậc p (viết tắt là AR(p)) là quá trình phụ thuộc tuyến tính của các giá trị trễ và sai số ngẫu nhiên được diễn giải như sau:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + ... + \beta_p Y_{t-p} \quad (1)$$

Mô hình trung bình trượt bậc q, viết tắt là MA(q), là quá trình được mô tả hoàn toàn bằng phương trình tuyến tính có trọng số của các sai số ngẫu nhiên hiện hành và các giá trị trễ của nó. Mô hình được viết như sau:

$$Y_t = \mu + \gamma_0 u_t + \gamma_1 u_{t-1} + ... + \gamma_q u_{t-q} \quad (2)$$

Mô hình tự tương quan tích hợp với trung bình trượt có dạng ARIMA (p,d,q), được xây dựng dựa trên 2 quá trình (1) và (2) được tích hợp. Phương trình tổng quát là:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + ... + \beta_p Y_{t-p} + \gamma_0 u_t + \gamma_1 u_{t-1} + ... + \gamma_q u_{t-q} \quad (3)$$

*ThS., Trường Đại học Bà Rịa - Vũng Tàu | Email: nghiemphuchieu@bvu.edu.vn

Ngày nhận bài: 16/11/2017; Ngày phản biện: 30/11/2017; Ngày duyệt đăng: 12/12/2017

Tuy nhiên, hiện nay, để có những dự báo chính xác các chỉ tiêu kinh tế - xã hội, người ta đã đi sâu tìm hiểu, nghiên cứu và đề xuất một số phương pháp dự báo dữ liệu chuỗi thời gian có yếu tố mùa vụ. Trong đó, mô hình SARIMA được phát triển tiếp từ mô hình ARIMA phù hợp với bất kỳ dữ liệu chuỗi thời gian mùa vụ nào, có thể là 4 quý trong năm; 7 ngày trong tuần; 11, 12 tháng trong một năm... Nếu chuỗi dữ liệu quan sát có tính mùa vụ, thì mô hình ARIMA tổng quát lúc này là SARIMA(p,d,q)(P,D,Q)$_L$ (với P và Q lần lượt là bậc của thành phần mùa AR và MA, D là bậc sai phân có tính mùa, L là số thời đoạn trong một vòng chu kỳ).

Trong những năm qua, có nhiều nghiên cứu được tiến hành để dự báo lượng khách quốc tế sử dụng mô hình SARIMA dựa theo phương pháp chuẩn Box-Jenkins.

Chaitip và cộng sự (2008) áp dụng SARIMA, ARIMA, các mô hình Holt-Winters, mạng thần kinh, VAR, GMM, ARCH-GARCH-M, ARCH-GARCH, TARCH, PARCH và EGARCH, nhằm dự báo lượng khách du lịch tới Thái Lan. Mô hình SARIMA đưa ra kết quả tốt nhất. Tương tự như vậy, Suhartono (2011) cũng thực hiện những phương pháp mới với dữ liệu khách theo đường hàng không tới Bali. Một lần nữa mô hình SARIMA là mô hình tốt nhất dùng để dự báo.

Ngoài ra, mô hình SARIMA cũng được sử dụng trong những lĩnh vực khác, như: Wongkoon và cộng sự (2008) áp dụng mô hình để dự báo số ca sốt xuất huyết ở miền Bắc Thái Lan; K. Rajendran và cộng sự (2011) sử dụng mô hình SARIMA và tuyến tính tổng quát (GLM) để nghiên cứu mối tương quan giữa số ca bệnh dịch tả với thời tiết...

Tại Việt Nam, cũng đã có nhiều công trình sử dụng mô hình SARIMA để dự báo, như: Nguyễn Khắc Hiếu (2014) sử dụng mô hình SARIMA để dự báo lạm phát 6 tháng cuối năm 2014; Vương Quốc Duy và Huỳnh Hải Âu (2014) ứng dụng mô hình SARIMA trong dự báo ngắn hạn lạm phát từ tháng 08/2013 đến tháng 07/2014 cho thấy mô hình SARIMA $(1,0,1)(2,0,3)_{12}$ là phù hợp nhất.

Phương pháp nghiên cứu

Bài viết ứng dụng mô hình SARIMA trong phân tích và dự báo lượng khách quốc tế đến Việt Nam, được thực hiện theo 4 bước sau đây:

Bước 1 - Nhận dạng mô hình: Xác định các giá trị (D, d, p, P, q, Q). Trong đó, trước hết cần xác định bậc sai phân theo mùa vụ D, sai phân thường d và thực hiện biến đổi chuỗi thành chuỗi dừng. Sau đó, kiểm tra biểu đồ của hàm tự tương quan (Autocorrelation Function - ACF), và hàm tự tương quan riêng phần (Partial Autocorrelation Function - PACF) tại các trễ mùa vụ và trễ thường; thực hiện kiểm định nghiệm đơn vị để xác định bậc tự hồi quy p và tự hồi quy mùa vụ P, bậc trung bình trượt q và trung bình trượt mùa vụ Q.

Bước 2 - Ước lượng mô hình: Ước lượng các tham số, sử dụng phương pháp ước lượng cực đại hợp lý để ước lượng giá trị các tham số này.

Bước 3 - Kiểm định: Kiểm định tính hợp lý của mô hình SARIMA được lựa chọn, bao gồm kiểm định các tham số và kiểm định phần dư. Nếu kiểm định mô hình được lựa chọn không thỏa mãn thì quay lại từ giai đoạn nhận dạng để lựa chọn mô hình khác hợp lý hơn.

Bước 4 - Dự báo: Dựa trên mô hình được lựa chọn thực hiện dự báo giá trị tương lai của dữ liệu chuỗi mùa vụ, cũng như đưa ra khoảng tin cậy của dự báo. Giá trị tương lai có thể được dự báo cho thời điểm kế tiếp hoặc mùa vụ kế tiếp.

Dữ liệu được sử dụng trong bài viết là số lượng khách quốc tế đến Việt Nam theo tháng của Tổng cục Thống kê từ tháng 10/2009 đến tháng 10/2017 và được xử lý bằng phần mềm EVIEWS 6.0. Tổng cộng bao gồm 97 quan sát, 92 quan sát từ tháng 10/2009 đến hết tháng 05/2017 sử dụng vào việc thiết lập mô hình, còn lại 5 quan sát từ tháng 06/2017 đến tháng 10/2017 dùng để kiểm tra tính chính xác của dự báo *(Bài viết sử dụng cách viết số thập phân theo chuẩn quốc tế).*

KẾT QUẢ NGHIÊN CỨU

Nhận dạng mô hình

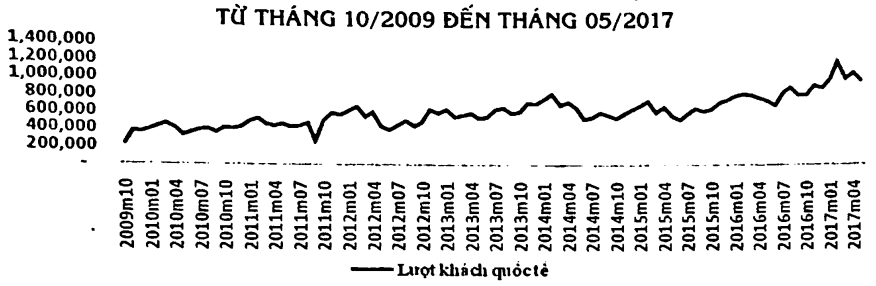
Chuỗi số liệu sử dụng trong mô hình SARIMA theo phương pháp Box-Jenkins được giả định là chuỗi dừng, vì vậy để dự báo lượng khách quốc tế bằng mô hình này cần phải xem xét chuỗi dữ liệu nghiên cứu có dừng hay chưa. Trước tiên, dựa vào việc quan sát đồ thị của chuỗi số liệu, sau đó tiến hành kiểm tra tính dừng này thông qua hai kiểm định phổ biến: Augmented Dickey-Fuller (ADF) và Perron-Phillips (PP) được gọi là kiểm định nghiệm đơn vị (unit root test).

Hình 1 cho thấy, chuỗi dữ liệu nghiên cứu chưa dừng, ta cần lấy sai phân bậc 1 chuỗi dữ liệu và tiến hành hai kiểm định ADF và PP như Bảng 1.

Kết quả của cả hai kiểm định ADF và PP đều cho phép ta bác bỏ giả thuyết H_0 về tính dừng của dữ liệu ở mức ý nghĩa 1% (Bảng 1).

Tiếp đó, để xác định giá trị p, q của mô hình SARIMA, ta phải dựa vào biểu đồ hàm tự tương quan ACF và tự tương quan từng phần PACF. Trong biểu đồ PACF ở Hình 2, các hệ số tương quan riêng phần khác không có ý nghĩa ở các độ trễ 1, 5 và 12 sau đó tắt dần về 0. Còn đối với biểu đồ ACF, ta có các hệ số tương quan khác không có ý nghĩa ở các độ trễ 1, sau đó tắt

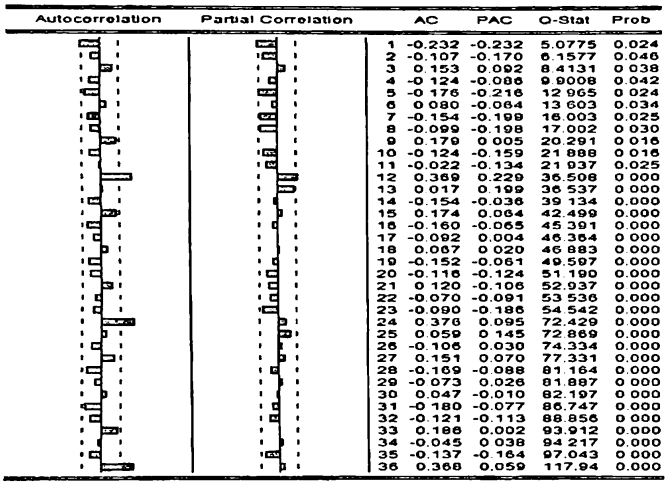
HÌNH 1: LƯỢNG KHÁCH QUỐC TẾ ĐẾN VIỆT NAM
TỪ THÁNG 10/2009 ĐẾN THÁNG 05/2017



BẢNG 1: KẾT QUẢ KIỂM ĐỊNH ADF VÀ PP

Kiểm định		Giá trị t		Xác suất
ADF	-	12.0821		0.0001
PP	-	12.1572		0.0001
Các giá trị tới hạn ở mức ý nghĩa thống kê 1%, 5%, 10% tương ứng là: -3.505, -2.894, -2.584				

HÌNH 2: BIỂU ĐỒACF VÀ PACF



	AC	PAC	Q-Stat	Prob
1	-0.232	-0.232	5.0775	0.024
2	-0.107	-0.170	6.1577	0.046
3	0.153	0.092	8.4131	0.038
4	-0.124	-0.086	9.9008	0.042
5	-0.176	-0.216	12.965	0.024
6	0.080	-0.064	13.603	0.034
7	-0.154	-0.199	16.003	0.025
8	-0.099	-0.198	17.002	0.030
9	0.179	0.005	20.291	0.016
10	-0.124	-0.159	21.888	0.015
11	-0.022	-0.134	21.937	0.025
12	0.369	0.229	36.508	0.000
13	0.017	0.199	36.537	0.000
14	-0.154	-0.036	39.134	0.000
15	0.174	0.064	42.499	0.000
16	-0.160	-0.065	45.391	0.000
17	-0.092	0.004	46.354	0.000
18	0.067	0.020	46.883	0.000
19	-0.152	-0.061	49.597	0.000
20	-0.116	-0.124	51.190	0.000
21	0.120	-0.106	52.937	0.000
22	-0.070	-0.091	53.536	0.000
23	-0.090	-0.186	54.542	0.000
24	0.376	0.095	72.429	0.000
25	0.059	0.145	72.869	0.000
26	-0.106	0.030	74.334	0.000
27	0.151	0.070	77.331	0.000
28	-0.169	-0.088	81.164	0.000
29	-0.073	0.026	81.887	0.000
30	0.047	-0.010	82.197	0.000
31	-0.180	-0.077	86.747	0.000
32	-0.121	-0.113	88.856	0.000
33	0.186	0.002	93.912	0.000
34	-0.045	0.038	94.217	0.000
35	-0.137	-0.164	97.043	0.000
36	0.368	0.059	117.94	0.000

BẢNG 2: CÁC MÔ HÌNH SARIMA(P,D,Q)(P,D,Q)$_L$ THỬ NGHIỆM

Mô hình	*R^2 điều chỉnh*	*AIC*	*SC*
SARIMA $(1,1,1)\ (1,1,3)_{12}$	0.717	24.283	24.494
SARIMA $(5,1,1)\ (1,1,3)_{12}$	0.718	24.319	24.537
SARIMA $(12,1,1)\ (1,1,3)_{12}$	0.814	23.798	24.029

BẢNG 3: KẾT QUẢ ƯỚC LƯỢNG CỦA MÔ HÌNH SARIMA $(12,1,1)\ (1,1,3)_{12}$

Biến	Hệ số	Sai số chuẩn	Thống kê t	Xác suất
C	-81,221.500	105,270	-0.771	0.443
AR (12)	-0.372	0.124	-3.006	0.004
SAR (12)	1.052	0.059	17.762	0.000
MA (1)	-0.480	0.124	-3.878	0.000
SMA (12)	-0.336	0.055	-6.110	0.000
SMA (24)	-0.308	0.000	-232529.8	0.000
SMA (36)	0.905	0.058	15.662	0.000

Nguồn: Xử lý số liệu của tác giả

dần về 0. Như vậy, biểu đồ chỉ ra rằng ta nên chọn p (1,5,12) và q(1) cho thành phần không có tính mùa.

Hình 1 cũng cho thấy có những đỉnh nhọn ở các độ trễ 12, 24 và 36 trên ACF sau đó tắt hết về 0, gợi ý rằng thành phần MA mùa cần được xem xét trong mô hình. Trên PACF tồn tại những đỉnh nhọn ở độ trễ 12 sau đó tắt hết về 0, do đó thành phần AR mùa cũng phải cần được bao gồm. Điều đó có nghĩa ta nên chọn P = 1, Q = 3 và L = 12 cho thành phần có tính mùa.

Tóm lại, các dạng mô hình SARIMA được nhận diện bao gồm: SARIMA$(1,1,1)(1,1,3)_{12}$, SARIMA$(5,1,1)(1,1,3)_{12}$, SARIMA$(12,1,1)(1,1,3)_{12}$.

Ước lượng mô hình

Các mô hình đã nhận diện được so sánh tính phù hợp dựa trên các thông số kiểm định sau hồi quy bao gồm: R^2 điều chỉnh, công cụ thông tin Akaike (AIC), công cụ Schwarz (SC) để lựa chọn mô hình phù hợp nhất. Thông số R^2 điều chỉnh phải càng lớn, trong khi đó AIC và SC phải càng nhỏ thì càng tốt, mô hình sẽ càng phù hợp.

Từ Bảng 2 ta thấy, mô hình SARIMA$(12,1,1)(1,1,3)_{12}$ là mô hình thỏa mãn nhiều nhất các tiêu chuẩn sử dụng, do đó đây là mô hình được vận dụng vào việc dự báo ngoài mẫu.

Kết quả hồi quy Bảng 3 cho thấy, có 6 hệ số có ý nghĩa ở mức 1%. Cũng trong Bảng 3, SAR (thể hiện điều kiện chạy mô hình mang tính thời vụ) được thêm vào mô hình khi ACF ở khoảng thời gian mùa vụ (12 tháng) là dương và SMA (thể hiện điều kiện chạy mô hình mang tính thời vụ) được thêm vào nếu như ACF ở khoảng thời gian mùa vụ (12 tháng) là âm.

Mô hình sau đó được kiểm tra mức độ phù hợp với chuỗi dữ liệu nghiên cứu bằng cách phân tích phần dư.

Kiểm định phần dư

Biểu đồ ACF của phần dư ở Hình 3 cho thấy, không có thanh nào vượt quá 2 đường biên cho thấy sai số là một nhiễu trắng. Ngoài ra, kết quả kiểm định Breusch-Godfrey ở mức ý nghĩa 1% cũng cho thấy không tồn tại hiện tượng tự tương quan bậc 2.

Kết quả kiểm tra mô hình SARIMA$(12,1,1)(1,1,3)_{12}$ bằng kiểm định Breusch - Godfrey là thích hợp và có thể sử dụng để dự báo (Bảng 4).

Dự báo

Bảng 5 thể hiện các giá trị dự báo trong 5 tháng từ tháng 06/2017 tới tháng 10/2017 và so sánh với các giá trị thực tế. Kết quả cho thấy chênh lệch giữa giá trị dự báo và thực tế lượng khách quốc tế đến Việt Nam trong giai đoạn này khá thấp, chỉ trừ trường hợp tháng 08/2017 có chênh lệch

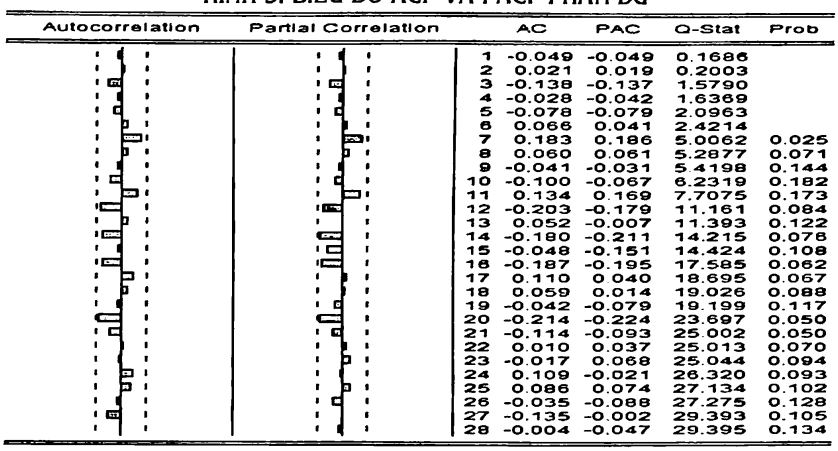
nhiều so với thực tế do ngành du lịch và các công ty lữ hành đã có nhiều biện pháp tổ chức hiệu quả các hoạt động văn hóa, du lịch nhằm thu hút khách quốc tế.

Từ đó, ta dự báo lượng khách quốc tế đến Việt Nam trong 3 năm sắp tới (Bảng 6). Dự báo cho rằng lượng khách quốc tế có tốc độ tăng nhanh trong những năm tiếp theo và tới năm 2020 sẽ vượt mốc 20 triệu lượt khách. Điều này là một tín hiệu rất tốt cho ngành du lịch Việt Nam và cũng là thách thức khiến chúng ta cần phải chuẩn bị nhiều nguồn lực để đón tiếp khách quốc tế một cách chu đáo nhất.

KẾT LUẬN VÀ KIẾN NGHỊ

Kết quả nghiên cứu cho thấy, lượng khách quốc tế đến Việt Nam có tốc độ tăng nhanh trong những năm tiếp theo và tới năm 2020 sẽ vượt mốc 20 triệu lượt khách. Theo đó, để sẵn sàng cho công tác tiếp đón du khách quốc tế với số lượng rất lớn trong thời gian tới, ngành du lịch cần tập trung chú trọng phát triển cơ sở vật chất hạ tầng du lịch, nâng cao chất lượng sản phẩm dịch vụ, chuẩn bị tốt nguồn nhân lực du lịch đáp ứng yêu cầu về chất lượng, ngoại ngữ tốt, cơ cấu ngành nghề và tính chuyên nghiệp,tăng cường khai thác các công nghệ thông tin hiện đại, khai thác hiệu quả internet, báo chí, truyền thông để phục vụ cho công tác quảng bá du lịch Việt Nam tại các thị trường trọng điểm, tăng cường hội nhập hợp tác quốc tế về du lịch.❑

HÌNH 3: BIỂU ĐỒ ACF VÀ PACF PHẦN DƯ



	AC	PAC	Q-Stat	Prob
1	-0.049	-0.049	0.1686	
2	0.021	0.019	0.2003	
3	-0.138	-0.137	1.5790	
4	-0.028	-0.042	1.6369	
5	-0.078	-0.079	2.0963	
6	0.066	0.041	2.4214	
7	0.183	0.186	5.0062	0.025
8	0.060	0.061	5.2877	0.071
9	-0.041	-0.031	5.4198	0.144
10	-0.100	-0.067	6.2319	0.182
11	0.134	0.169	7.7075	0.173
12	-0.203	-0.179	11.161	0.084
13	0.052	-0.007	11.393	0.122
14	-0.180	-0.211	14.215	0.076
15	-0.048	-0.151	14.424	0.108
16	-0.187	-0.195	17.585	0.062
17	0.110	0.040	18.695	0.067
18	0.059	0.014	19.026	0.088
19	-0.042	-0.079	19.199	0.117
20	-0.214	-0.224	23.697	0.050
21	-0.114	-0.093	25.002	0.050
22	0.010	0.037	25.013	0.070
23	-0.017	0.068	25.044	0.094
24	0.109	-0.021	26.320	0.093
25	0.086	0.074	27.134	0.102
26	-0.035	-0.088	27.275	0.128
27	-0.135	-0.002	29.393	0.105
28	-0.004	-0.047	29.395	0.134

BẢNG 4: KẾT QUẢ KIỂM ĐỊNH BREUSCH-GODFREY

Kiểm định	Giá trị t	Xác suất
Breusch-Godfrey	2.168	0.373

BẢNG 5: KẾT QUẢ DỰ BÁO

Thời gian	Lượng khách thực tế	Lượngkhách dự báo	Chênh lệch
06/2017	949,362	889,549	-6.30%
07/2017	1,036,880	1,005,056	-3.07%
08/2017	1,229,163	1,076,734	-12.40%
09/2017	975,952	1,037,479	6.30%
10/2017	1,024,899	1,007,445	-1.7%

BẢNG 6: DỰ BÁO KHÁCH QUỐC TẾ TỚI 2020

Thời gian	Dự báo lượng khách
Năm 2018	15,314,765
Năm 2019	19,151,568
Năm 2020	24,080,227

Nguồn: Xử lý số liệu của tác giả

TÀI LIỆU THAM KHẢO

1. Tổng cục Thống kê (2009-2017). *Báo cáo tình hình kinh tế - xã hội*, từ tháng 10/2009 đến tháng 10/2017

2. Vương Quốc Duy, Huỳnh Hải Âu (2014). Dự báo lạm phát Việt Nam giai đoạn 8/2013-7/2014, *Tạp chí khoa học Đại học Cần Thơ*, số 30, 34-41

3. Nguyễn Khắc Hiếu (2014). Mô hình ARIMA và dự báo lạm phát 6 tháng cuối năm 2014, *Tạp chí Kinh tế và Dự báo*, số 16/2014, 16-18

4. Box, G.E.P., and G.M. Jenkins (1976). *Time Series Analysis: Forecasting and Control*, Revised Edition, Holden Day, San Francisco

5. Chaitip, P., Chaiboonsri and R. Mukhjang (2008). Time Series Models for Forecasting International Visitor Arrivals to Thailand, *International Conference on Applied Economics*, 2008, 159-163

6. K. Rajendran, A. Sumi, M. K. Bhattachariya, B. Manna, D. Sur, N. Kobayashi and T. Ramamurthy (2011). Influence of relative humidity in Vibrio cholerae infection: a time series model, *Indian J Med Res*,133, 138-145

7. S. Wongkoon, M. Pollar, M. Jaroensutasinee and K. Jaroensutasinee (2008). Predicting DHF Incidence in Northern Thailand using Time Series Analysis Technique, *International Journal of Biological and LifeSciences*, 4(3)

8. Suhartono (2011). Time Series Forecasting by using Seasonal Autoregressive Integrated Moving Average: Subset, Multiplicative or Additive Model, *Journal of Mathematics and Statistics*, 7(1), 20-27