SỬ DỤNG MÔ HÌNH ARIMA TRONG DỰ BÁO GIÁ

Cao Hào Thi
Khoa Quản Lý Công Nghiệp
Đại Học Bách Khoa Tp.HCM

TÓM TẮT

Mục tiêu của nghiên cứu này nhằm giới thiệu việc xây dựng mô hình của các quá trình ngẫu nhiên – Mô Hình ARIMA, và ứng dụng mô hình này trong việc dự báo. Mô hình này giải thích sự biến động của chuỗi thời gian bằng cách quan hệ với các giá trị quá khứ và tổng có trọng số các nhiễu ngẫu nhiên hiện hành và các nhiễu ngẫu nhiên có độ trễ. Mô hình cũng được ứng dụng một cách minh họa nhằm dự báo giá cá sông tại Thành Phố Hồ Chí Minh.

ABSTRACT

The objective of this reseach is to introduce the construction of the model of stochastic processes – ARIMA model, and their use in forecasting. This model explains the movement of the time series by relating it to the own past values and to the weighted sum of current and lagged random disturbances. The model is also illustratively applied to forecast the price of riverfish in HoChiMinh City.

I. GIỚI THIỆU

Trong lãnh vực Kinh Tế Lượng, việc dự báo thường dựa trên hai loại mô hình chính là mô hình nhân quả và mô hình chuỗi thời gian. Trong mô hình nhân quả, kỹ thuật phân tích hồi qui được sử dụng để thiết lập mối quan hệ giữa biến phụ thuộc và các biến nguyên nhân. Giá trị của biến phụ thuộc sẽ được dự báo theo giá trị của các biến nguyên nhân. Đối với các chuỗi thời gian, mô hình ARIMA được sử dụng để dự báo các giá trị trong tương lai. Theo mô hình này, giá trị dự báo sẽ phụ thuộc vào các giá trị quá khứ và tổng có trọng số các nhiễu ngẫu nhiên hiện hành và các nhiễu ngẫu nhiên có độ trễ.

Mục tiêu của nghiên cứu này nhằm giới thiệu việc xây dựng mô hình của các quá trình ngẫu nhiên – Mô Hình ARIMA, và ứng dụng của mô hình này trong việc dự báo. Mô hình cũng được ứng dụng một cách minh họa nhằm dự báo giá cá sông tại Thành Phố Hồ Chí Minh.

II. MÔ HÌNH ARIMA

Nhằm mục đích giới thiệu về mô hình Tự Hồi Qui Kết Hợp Trung Bình Trượt (ARIMA–**A**uto**R**egressive **I**ntegrated **M**oving **A**verage), nội dung phần này sẽ trình bày tóm lược một số cơ sở lý thuyết liên quan đến tính dừng (stationary), tính mùa vụ (seasonality), nguyên lý Box-Jenkin; cùng một số nguyên tắc để nhận dạng, xác định các thông số và và các kiểm định về mô hình ARIMA.

Tính dừng

Một quá trình ngẫu nhiên Y_t được xem là dừng nếu như trung bình và phương sai của quá trình không thay đổi theo thời gian và giá trị của đồng phương sai giữa hai thời đoạn chỉ phụ thuộc vào khoảng cách hay độ trễ về thời gian giữa hai thời đoạn này chứ không phụ thuộc vào thời điểm thực tế mà đồng phương sai được tính. Cụ thể:

- Trung bình: $E(Y_t) = \mu = const$
- Phương sai: $Var(Y_t) = \sigma^2 = const$
- Đồng phương sai: $Covar(Y_t, Y_{t-k}) = 0$

Tính dừng của một chuỗi thời gian có thể được nhận biết dựa trên đồ thị của chuỗi thời gian, đồ thị của hàm tự tương quan mẫu hay kiểm định Dickey-Fuller.

- Dựa trên đồ thị $Y_t = f(t)$, một cách trực quan chuỗi Y_t có tính dừng nếu như đồ thị cho thấy trung bình và phương sai của quá trình Y_t không thay đổi theo thời gian.

- Dựa vào hàm tự tương quan mẫu (SAC – Sample Auto Correllation)

$$\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_o} = SAC$$

$$\hat{\gamma}_k = E[(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})] = \frac{\sum (Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{n} = Cov(Y_t, Y_{t-k})$$

$$\hat{\gamma}_o = E[(Y_t - \bar{Y})^2] = \frac{\sum (Y_t - \bar{Y})^2}{n} = Var(Y_t)$$

Nếu SAC = f(t) của chuỗi thời gian giảm nhanh và tắt dần về 0 thì chuỗi có tính dừng.

- Kiểm định Dickey-Fuller (kiểm định nghiệm đơn vị) nhằm xác định xem chuỗi thời gian có phải là Bước Ngẫu Nhiên (Random Walk; nghĩa là $Y_t = \mathbf{1}*Y_{t-1} + \varepsilon_t$) hay không. Nếu chuỗi là Bước Ngẫu Nhiên thì không có tính dừng. Tuy nhiên, Nếu chuỗi không có tính dừng thì chưa chắc là Bước Ngẫu Nhiên.

Để biến đổi chuỗi không dừng thành chuỗi dừng, thông thường nếu lấy sai phân một lần hoặc hai lần thì sẽ được một chuỗi kết quả có tính dừng.

- Chuỗi gốc: Y_t
- Chuỗi sai phân bậc 1: $W_t = Y_t - Y_{t-1}$
- Chuỗi sai phân bậc 2: $V_t = W_t - W_{t-1}$

Tính mùa vụ

Tính mùa vụ là hành vi có tính chu kỳ của chuỗi thời gian trên cơ sở năm lịch. Tính mùa vụ có thể được nhận ra dựa vào đồ thị SAC = f(t). Nếu cứ sau m thời đoạn thì SAC lại có giá trị cao (nghĩa là đồ thị SAC có đỉnh cao) thì đây là dấu hiệu của tính mùa vụ. Chuỗi thời gian có tồn tại tính mùa vụ sẽ không có tính dừng. Phương pháp đơn giản nhất để khử tính mùa vụ là lấy sai phân thứ m. Nếu Y_t có tính mùa vụ với chu kỳ m thời đoạn thì chuỗi $Z_t = Y_t - Y_{t-m}$ sẽ được khảo sát thay vì chuỗi Y_t.

Mô hình ARIMA

Theo Box- Jenkin mọi quá trình ngẫu nhiên *có tính dừng* đều có thể biểu diễn bằng mô hình Tự Hồi Qui Kết Hợp Trung Bình Trượt ARIMA.

- Mô Hình Tự Hồi Qui Bậc p - AR(p)

Trong mô hình tự hồi qui quá trình phụ thuộc vào tổng có trọng số của các giá trị quá khứ và số hạng nhiều ngẫu nhiên

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + \delta + \varepsilon_t$$

- Mô Hình Trung Bình Trượt Bậc q – MA(q)

Trong mô hình trung bình trượt, quá trình được mô tả hoàn toàn bằng tổng có trọng số của các ngẫu nhiên hiện hành có độ trễ

$$Y_t = \mu + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - ... - \theta_q \varepsilon_{t-q}$$

- Mô Hình Hồi Quy Kết Hợp Trung Bình Trượt - ARIMA(p,d,q)

Phương trình tổng quát của mô hình ARIMA là:

$$Y_t = \phi_1 Y_{t-1} + ... + \phi_p Y_{t-p} + \delta + \varepsilon_t - \theta_1 \varepsilon_{t-1} - ... - \theta_q \varepsilon_{t-q}$$

Nhận dạng mô hình

Nhận dạng mô hình ARIMA(p,d,q) là tìm các giá trị thích hợp của p, d, q. Với d là bậc sai phân của chuỗi thời gian được khảo sát, p là bậc tự hồi qui và q là bậc trung bình trượt. Việc xác định p và q sẽ phụ thuộc vào các đồ thị SPAC = f(t) và SAC = f(t). Với SAC đã được giới thiệu ở trên và SPAC là Tự Tương Quan Riêng Phần Mẫu (Sample Partial Auto-Correlation); nghĩa là tương quan giữa Y_t và Y_{t-p} sau khi đã loại bỏ tác động của các Y trung gian.

- Chọn mô hình AR(p) nếu đồ thị SPAC có giá trị cao tại độ trễ 1, 2, ..., p và giảm nhiều sau p và dạng hàm SAC giảm dần.
- Chọn mô hình MA(q) nếu đồ thị SAC có giá trị cao tại độ trễ 1, 2, ..., q và giảm nhiều sau q và dạng hàm SPAC giảm dần. Tóm lại,

Loại mô hình	Dạng đồ thị SAC = f(t)	Dạng đồ thị SPAC = f(t)
AR(p)	Giảm dần	Có đỉnh ở p
MA(q)	Có đỉnh ở q	Giảm dần
ARMA(p, q)	Giảm dần	Giảm dần

Ước lượng các thông số của mô hình ARIMA(p, d, q)

Các thông số ϕ_i và θ_j của mô hình ARIMA sẽ được xác định theo phương pháp bình phương tối thiểu (OLS-Ordinary Least Square) sao cho:

$$\sum (Y_t - \hat{Y}_t)^2 \rightarrow Min$$

Với

$$\varepsilon_t = (Y_t - \hat{Y}_t)$$

Kiểm tra chẩn đoán mô hình

Sau khi xác định p, d, q và các ϕ_i, θ_j; nghĩa là đã xác định được phương trình cho mô hình ARIMA, điều cần phải làm là tiến hành kiểm định xem số hạng ε_t của mô hình có phải là một nhiễu trắng (white noise, nhiễu ngẫu nhiên thuần túy) hay không. Đây là yêu cầu của một mô hình tốt.

Về mặt lý thuyết, ε_t được tạo ra bởi quá trình nhiều trắng nếu:

$$+\varepsilon_t \sim N(0,\sigma_\varepsilon^2) \begin{cases} E(\varepsilon_t)=0 \\ Var(\varepsilon_t)=\sigma_\varepsilon^2=const \end{cases}$$

$$+\gamma_k = Cov(\varepsilon_t,\varepsilon_{t-k})=0$$

Việc kiểm định tính nhiễu trắng sẽ dựa trên đồ thị SAC của chuỗi $\varepsilon_{t.}$

Dự báo

Dựa trên phương trình của mô hình ARIMA, tiến hành xác định giá trị dự báo điểm và khoảng tin cậy của dự báo.

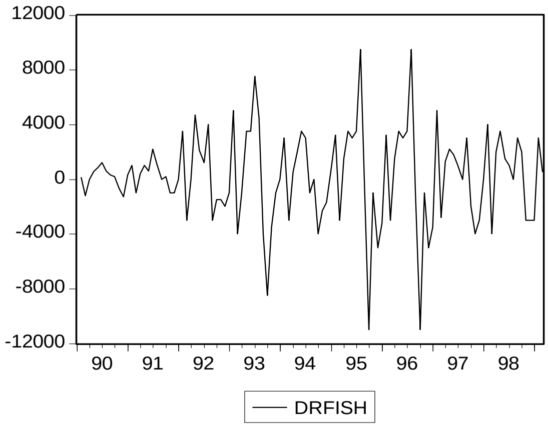
- Dự báo điểm: $\hat{Y}_t$
- Khoảng tin cậy: $\hat{Y}_t - k\sigma(\varepsilon_t) < Y_t < \hat{Y}_t + k\sigma(\varepsilon_t)$

  Với độ tin cậy 95%, k =2.

III. SỬ DỤNG MÔ HÌNH ARIMA TRONG DỰ BÁO GIÁ

Để minh họa, nghiên cứu đã áp dụng mô hình ARIMA trong việc dự báo giá cá sông tại thành phố Hồ Chí Minh. Nghiên cứu đã sử dụng chuỗi gồm 111 dữ liệu tháng từ tháng 1/1990 đến tháng 3/1999 và phần mềm EVIEWS để dự báo giá trị tháng 4/1999.

Các dữ liệu quá khứ của giá cá sông được đặt tên là RFISH và chuỗi sai phân bậc 1 được đặt tên là DRFISH. Đồ thị RFISH = f(t) và DRFISH = f(t) được trình bày như sau:



Đồ thị RFISH cho thấy chuỗi RFISH không có tính dừng. Đồ thị DRFISH cho thấy chuỗi DRFISH cũng không có tính dừng. Qua đồ thị trên và dữ liệu ta nhận thấy chuỗi có tính mùa vụ theo quí. Các kiểm định theo hàm tự tương quan mẫu hay kiểm định Dickey-Fuller trong EVIEWS cũng cho cho thấy chuỗi RFISH và DRFISH không có tính dừng do dữ liệu có tính mùa vụ.

Sử dụng phần mềm EVIEW để khử tính mùa vụ và tiến hành thử nghiệm cho nhiều mô hình ARIMA, cuối cùng ta được mô hình tối ưu có dạng ARIMA(2,1,2) với thời đoạn khử tính mùa vụ là m = 12. Kết quả về các thông số ϕ_i và θ_j được trình bày trong bảng sau:

Dependent Variable: D(RFISH)
Method: Least Squares
Date: 2/3/2002 Time: 18:17
Sample(adjusted): 1991:04 1999:03
Included observations: 96 after adjusting endpoints
Convergence achieved after 50 iterations
Backcast: 1990:02 1991:03

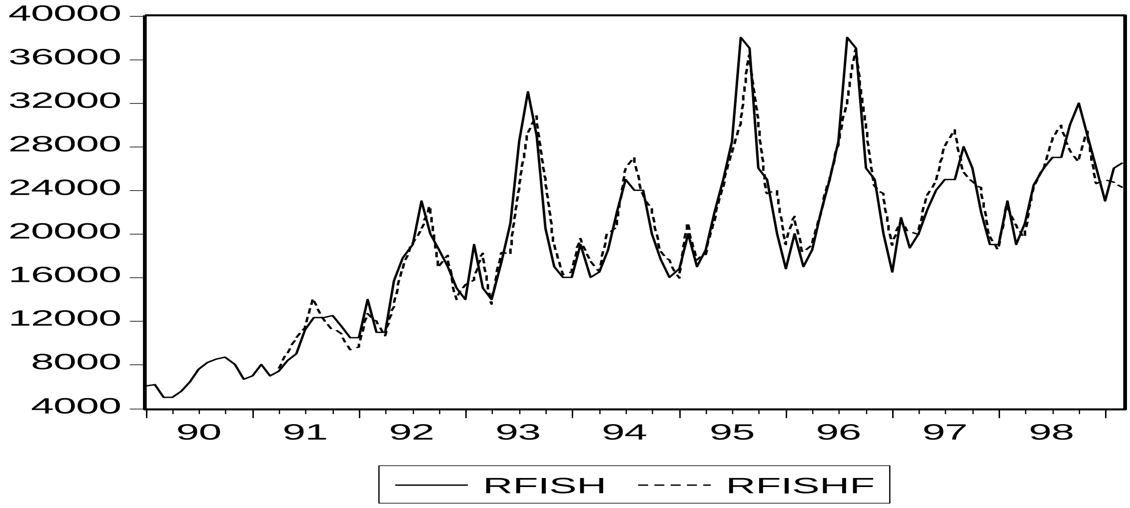
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-283.3601	1010.997	-0.280278	0.7799
AR(2)	0.413278	0.135466	3.050799	0.0030
SAR(12)	0.963121	0.044544	21.62164	0.0000
MA(2)	-0.846851	0.118603	-7.140218	0.0000
SMA(12)	-0.781433	0.078476	-9.957634	0.0000
R-squared	0.614807	Mean dependent var		203.1250
Adjusted R-squared	0.597875	S.D. dependent var		3545.923
S.E. of regression	2248.588	Akaike info criterion		18.32467
Sum squared resid	4.60E+08	Schwarz criterion		18.45823
Log likelihood	-874.5842	F-statistic		36.31124
Durbin-Watson stat	1.718345	Prob(F-statistic)		0.000000

Sau khi xác định được phương trình cho mô hình ARIMA, cần phải tiến hành kiểm định tính nhiễu trắng của ε_t . Kết quả kiểm định dựa trên đồ thị SAC của chuỗi ε_t. cho thấy ε_t có tính nhiễu trắng và được trình bày như sau:

Date: 2/3/2002 Time: 18:20
Sample: 1991:04 1999:03
Included observations: 96
Q-statistic probabilities adjusted for 4 ARMA term(s)

<table>
<tr><th>Autocorrelation</th><th>Partial Correlation</th><th></th><th>AC</th><th>PAC</th><th>Q-Stat</th><th>Prob</th></tr>
<tr><td>. |*. |</td><td>. |*. |</td><td>1</td><td>0.108</td><td>0.108</td><td>1.1485</td><td></td></tr>
<tr><td>.*| . |</td><td>.*| . |</td><td>2</td><td>-0.060</td><td>-0.072</td><td>1.5093</td><td></td></tr>
<tr><td>. | . |</td><td>. | . |</td><td>3</td><td>-0.002</td><td>0.013</td><td>1.5099</td><td></td></tr>
<tr><td>.*| . |</td><td>.*| . |</td><td>4</td><td>-0.107</td><td>-0.115</td><td>2.6913</td><td></td></tr>
<tr><td>. | . |</td><td>. | . |</td><td>5</td><td>-0.046</td><td>-0.021</td><td>2.9138</td><td>0.088</td></tr>
<tr><td>. | . |</td><td>. | . |</td><td>6</td><td>0.003</td><td>-0.005</td><td>2.9147</td><td>0.233</td></tr>
<tr><td>. | . |</td><td>. | . |</td><td>7</td><td>-0.027</td><td>-0.031</td><td>2.9925</td><td>0.393</td></tr>
<tr><td>.*| . |</td><td>.*| . |</td><td>8</td><td>-0.069</td><td>-0.076</td><td>3.5015</td><td>0.478</td></tr>
<tr><td>. | . |</td><td>. | . |</td><td>9</td><td>-0.040</td><td>-0.037</td><td>3.6719</td><td>0.598</td></tr>
<tr><td>. | . |</td><td>. | . |</td><td>10</td><td>-0.008</td><td>-0.011</td><td>3.6789</td><td>0.720</td></tr>
<tr><td>. | . |</td><td>. | . |</td><td>11</td><td>-0.035</td><td>-0.046</td><td>3.8174</td><td>0.801</td></tr>
<tr><td>. |*. |</td><td>. |*. |</td><td>12</td><td>0.176</td><td>0.173</td><td>7.2986</td><td>0.505</td></tr>
<tr><td>. |*. |</td><td>. | . |</td><td>13</td><td>0.069</td><td>0.011</td><td>7.8428</td><td>0.550</td></tr>
<tr><td>. | . |</td><td>. | . |</td><td>14</td><td>0.010</td><td>0.025</td><td>7.8547</td><td>0.643</td></tr>
<tr><td>.*| . |</td><td>.*| . |</td><td>15</td><td>-0.059</td><td>-0.079</td><td>8.2589</td><td>0.690</td></tr>
<tr><td>.*| . |</td><td>.*| . |</td><td>16</td><td>-0.168</td><td>-0.133</td><td>11.559</td><td>0.482</td></tr>
<tr><td>. | . |</td><td>. | . |</td><td>17</td><td>0.021</td><td>0.064</td><td>11.614</td><td>0.560</td></tr>
<tr><td>. | . |</td><td>.*| . |</td><td>18</td><td>-0.050</td><td>-0.085</td><td>11.915</td><td>0.613</td></tr>
<tr><td>. | . |</td><td>. | . |</td><td>19</td><td>0.031</td><td>0.057</td><td>12.036</td><td>0.676</td></tr>
<tr><td>. | . |</td><td>.*| . |</td><td>20</td><td>-0.017</td><td>-0.060</td><td>12.072</td><td>0.739</td></tr>
<tr><td>.*| . |</td><td>.*| . |</td><td>21</td><td>-0.112</td><td>-0.087</td><td>13.648</td><td>0.692</td></tr>
<tr><td>.*| . |</td><td>.*| . |</td><td>22</td><td>-0.094</td><td>-0.099</td><td>14.771</td><td>0.678</td></tr>
<tr><td>.*| . |</td><td>.*| . |</td><td>23</td><td>-0.083</td><td>-0.084</td><td>15.649</td><td>0.681</td></tr>
<tr><td>.*| . |</td><td>**| . |</td><td>24</td><td>-0.177</td><td>-0.248</td><td>19.750</td><td>0.474</td></tr>
<tr><td>. |** |</td><td>. |** |</td><td>25</td><td>0.198</td><td>0.221</td><td>24.964</td><td>0.249</td></tr>
<tr><td>. |*. |</td><td>. | . |</td><td>26</td><td>0.133</td><td>0.014</td><td>27.356</td><td>0.198</td></tr>
<tr><td>. | . |</td><td>.*| . |</td><td>27</td><td>-0.053</td><td>-0.062</td><td>27.733</td><td>0.226</td></tr>
<tr><td>. | . |</td><td>. | . |</td><td>28</td><td>-0.039</td><td>-0.013</td><td>27.944</td><td>0.262</td></tr>
<tr><td>. | . |</td><td>.*| . |</td><td>29</td><td>-0.024</td><td>-0.070</td><td>28.023</td><td>0.307</td></tr>
<tr><td>. | . |</td><td>. | . |</td><td>30</td><td>-0.017</td><td>0.023</td><td>28.065</td><td>0.355</td></tr>
<tr><td>. | . |</td><td>.*| . |</td><td>31</td><td>0.005</td><td>-0.075</td><td>28.069</td><td>0.407</td></tr>
<tr><td>. | . |</td><td>.*| . |</td><td>32</td><td>-0.010</td><td>-0.059</td><td>28.084</td><td>0.460</td></tr>
<tr><td>. | . |</td><td>. | . |</td><td>33</td><td>-0.002</td><td>0.061</td><td>28.084</td><td>0.513</td></tr>
<tr><td>.*| . |</td><td>.*| . |</td><td>34</td><td>-0.072</td><td>-0.075</td><td>28.878</td><td>0.524</td></tr>
<tr><td>. | . |</td><td>. | . |</td><td>35</td><td>-0.040</td><td>-0.028</td><td>29.128</td><td>0.563</td></tr>
<tr><td>. | . |</td><td>. |*. |</td><td>36</td><td>0.045</td><td>0.117</td><td>29.450</td><td>0.596</td></tr>
</table>

Kết quả của mô hình dự báo được trình bày trong tập dữ liệu RFISHF. Đồ thị của RFISH và RFISHF được trình bày chung như sau:



Dựa trên phương trình của mô hình ARIMA, tiến hành xác định giá trị dự báo điểm và khoảng tin cậy của dự báo.
Dự báo điểm là $\hat{Y}_t$ = 26267 Đ và Khoảng tin cậy 95% là [21742 Đ, 30792 Đ]

Sau khi có kết quả dự báo, ta đem so với giá trị thực vào tháng 4/1999 là Y_t = 26000 Đ. Giá trị này nằm trong khoảng tin cậy 95% và xấp xỉ với giá trị dự báo điểm. Sai số dự báo là $(\hat{Y}_t - Y_t)/ Y_t *100$ = (26267 – 26000)/26000 * 100 = 1,03%

KẾT LUẬN

Kết quả dự báo cho thấy đồ thị của mô hình dự báo RFISHF bám rất sát đồ thị của chuỗi dữ liệu gốc RFISH. Điều này chứng tỏ mô hình ARIMA(2,1,2) này đã giải thích được sự sự biến động của chuỗi thời gian về giá cá sông tại Thành Phố Hồ Chí Minh.

Giá trị dự báo xấp xỉ với giá trị trên thực tế (sai số dự báo nhỏ) và khoảng tin cậy 95% cũng chứa giá trị thực` . Điều này chứng tỏ độ tin cậy của mô hình dự báo.

Ngoài ví dụ minh họa trên, nghiên cứu cũng đã áp dụng mô hình ARIMA để dự báo cho hơn 20 loại mặt hàng tại Thành Phố Hồ Chí Minh theo qui trình tương tự và cũng đạt được các kết quả dự báo với độ tin cậy cao. Tóm lại, Mô hình ARIMA là một mô hình đáng tin cậy đối với dự báo ngắn hạn.

TÀI LIỆU THAM KHẢO

Bowerman B.L., and O'Connell R.T., 1993. *Forecasting and Time Series*. 3rd ed., Wadsworth, Inc.

Cao Hào Thi và Các Cộng Sự 1998. *Bản Dịch Kinh Tế Lượng Cơ Sở (Basic Econometrics của Gujarati D.N.).* Chương Trình FulBright về Giảng Dạy Kinh Tế tại Việt Nam.

EVIEWS, 2000. Quantitative Micro Software.

Pindyck R.S., and Rubinfeld D.L., 1991. *Econometric Models and Economic Forecast.* 3rd ed., McGraw-Hill.

Ramanathan R., 2001. *Introductory Econometrics with Applications.* 5th ed., Harcourt College Publishers

Liên hệ : Cao Hào Thi
Địa chỉ : Khoa Quản Lý Công Nghiệp, Đại Học Bách Khoa Tp. HCM
268 Lý Thường Kiệt, Q.10, Tp. HCM
Tel : 84 - 8 - 8650460
Fax : 84 -8 - 8635058
Email : chthi@sim.hcmut.edu.vn