

**ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**

Nguyễn Ngọc Thiệp

**MỘT SỐ PHƯƠNG PHÁP KHAI PHÁ DỮ LIỆU QUAN
HỆ TRONG TÀI CHÍNH VÀ CHỨNG KHOÁN
(MÔ HÌNH ARIMA)**

KHÓA LUẬN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY

Ngành : Công nghệ thông tin

HÀ NỘI - 2010

**ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**

Nguyễn Ngọc Thiệp

**MỘT SỐ PHƯƠNG PHÁP KHAI PHÁ DỮ LIỆU QUAN
HỆ TRONG TÀI CHÍNH VÀ CHỨNG KHOÁN
(MÔ HÌNH ARIMA)**

KHÓA LUẬN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY

Ngành : Công nghệ thông tin

Cán bộ hướng dẫn : PGS-TS Hà Quang Thụy

Cán bộ đồng hướng dẫn : Th.s Nguyễn Thị Oanh.

HÀ NỘI - 2010

LỜI CẢM ƠN

Lời đầu tiên, em xin bày tỏ lòng biết ơn tới các thầy, cô giáo trong trường Đại học Công Nghệ - Đại học Quốc Gia Hà nội. Các thầy cô đã dạy bảo, chỉ dẫn em và luôn tạo điều kiện tốt nhất cho chúng em học tập trong suốt quá trình học đại học đặc biệt là trong thời gian làm khóa luận tốt nghiệp.

Em xin bày tỏ lòng biết ơn sâu sắc tới thầy giáo PGS.TS Hà Quang Thụy cùng cô giáo ThS Trần Thị Oanh, và các anh chị trong phòng LAB 102 đã hướng dẫn em tận tình trong năm học vừa qua.

Tôi cũng xin cảm ơn những người bạn của mình, các bạn đã luôn bên tôi, giúp đỡ và cho tôi những ý kiến đóng góp quý báu trong học tập cũng như trong cuộc sống.

Cuối cùng con xin gửi tới bố mẹ và toàn thể gia đình lòng biết ơn và tình cảm yêu thương nhất.

Hà Nội, ngày 10/05/2010

Nguyễn Ngọc Thiệp

MỞ ĐẦU

Bài toán dự báo tài chính ngày càng được nhiều người quan tâm trong bối cảnh phát triển kinh tế xã hội. Đầu tư vào thị trường chứng khoán đòi hỏi nhiều kinh nghiệm và hiểu biết của các nhà đầu tư. Các kỹ thuật khai phá dữ liệu được áp dụng nhằm dự báo sự lên xuống của thị trường là một gợi ý giúp các nhà đầu tư có thể ra quyết định giao dịch.

Mô hình ARIMA được xây dựng với chức năng nhận dạng mô hình, ước lượng các tham số và đưa ra kết quả dự báo dựa trên các tham số ước lượng đã được lựa chọn một cách tối ưu.

Khóa luận nghiên cứu, thi hành mô hình ARIMA (từ các nghiên cứu của Box-Jenkins) và ứng dụng vào bài toán khai phá dữ liệu chuỗi thời gian trong dự báo tài chính, chứng khoán. Khóa luận đã thực nghiệm trên dữ liệu vnIndex và đã thu được kết quả bước đầu.

Với nội dung trình bày những lý thuyết cơ bản về mô hình ARIMA cho dữ liệu thời gian thực (time series) và cách áp dụng vào bài toán thực tế - dự báo sự lên xuống của thị trường chứng khoán. Khóa luận được tổ chức theo cấu trúc như sau :

Chương 1. GIỚI THIỆU CHUNG giới thiệu sơ lược về khai phá dữ liệu nói chung và bài toán dự báo đang được quan tâm trong khai phá dữ liệu . Bài toán dự báo được áp dụng dưới khía cạnh sử dụng mô hình ARIMA cho chuỗi thời gian thực.

Chương 2. MÔ HÌNH ARIMA VÀ PHẦN MỀM EVIEW trình bày một số nội dung cơ sở lý thuyết về mô hình ARIMA, cũng như những công cụ sẽ được áp dụng vào trong mô hình mà khóa luận đề cập : Hàm tự tương quan ACF, hàm tự tương quan riêng phần PACF... Các bước phát triển mô hình : xác định mô hình, ước lượng các tham số, kiểm định độ chính xác và dự báo. Mô hình ARIMA là một quá trình thử và sai : khi một kiểm định nào đó không thỏa mãn, phải xác định lại mô hình. Tiếp đến giới thiệu qua về phần mềm Eviews 5.1 cho quá trình thi hành.

Chương 3. ÁP DỤNG MÔ HÌNH ARIMA VÀO BÀI TOÁN TÀI CHÍNH, CHỨNG KHOÁN trình bày thực nghiệm mô hình ARIMA cho dữ liệu tài chính, chứng khoán. Các bước trong quá trình thi hành chương trình với phần mềm Eviews 5.1, đưa ra kết quả và đánh giá với thực tế.

Phần Kết luận tổng kết kết quả của khóa luận và phương hướng nghiên cứu tiếp theo.

MỤC LỤC

MỞ ĐẦU	4
Chương 1. GIỚI THIỆU CHUNG	7
1.1. Bài toán dự báo	7
1.2. Dữ liệu chuỗi thời gian	9
1.2.1. Khái niệm chuỗi thời gian thực	10
1.2.2. Thành phần xu hướng dài hạn	10
1.2.3. Thành phần mùa	11
1.2.4. Thành phần chu kỳ	11
1.2.5. Thành phần bất thường.....	12
CHƯƠNG 2. MÔ HÌNH ARIMA VÀ PHẦN MỀM EViews	13
2.1. Mô hình ARIMA	13
2.1.1. Hàm tự tương quan ACF	13
2.1.2. Hàm tự tương quan từng phần PACF	14
2.1.3. Mô hình AR(p)	17
2.1.4. Mô hình MA(q)	17
2.1.5. Sai phân I(d)	18
2.1.6. Mô hình ARIMA	18
2.1.7. Các bước phát triển mô hình ARIMA	22
2.2. Phần mềm ứng dụng Eviews	22
2.2.1. Giới thiệu Eviews	22
2.2.2. Áp dụng Eviews thi hành các bước mô hình ARIMA	27
Tóm tắt chương 2	29
Chương 3. ÁP DỤNG MÔ HÌNH ARIMA VÀO BÀI TOÁN TÀI CHÍNH, CHỨNG KHOÁN...	30
3.1. Mô hình ARIMA cho dự báo tài chính, chứng khoán	30
3.1.1. Dữ liệu tài chính	30
3.1.2. Mô hình ARIMA cho bài toán dự báo tài chính.....	30

3.1.3. Thiết kế mô hình ARIMA cho dữ liệu	31
3.2. Áp dụng	33
3.2.1. Môi trường thực nghiệm.....	33
3.2.2.Dữ liệu.....	33
3.2.3.Kiểm tra tính dừng của chuỗi chứng khoán AAM	34
3.2.4.Nhận dạng mô hình	35
3.2.5.Ước lượng và kiểm định với mô hình ARIMA	37
3.2.6Thực hiện dự báo.....	38
KẾT LUẬN	41

Chương 1. GIỚI THIỆU CHUNG

1.1. Bài toán dự báo

Sự phát triển của công nghệ thông tin và việc ứng dụng công nghệ thông tin trong nhiều lĩnh vực của đời sống, kinh tế xã hội trong nhiều năm qua cũng đồng nghĩa với lượng dữ liệu đã được các cơ quan thu thập và lưu trữ ngày một tích lũy nhiều lên. Họ lưu trữ các dữ liệu này vì cho rằng trong nó ẩn chứa những giá trị nhất định nào đó. Tuy nhiên, theo thống kê thì chỉ có một lượng nhỏ của những dữ liệu này (khoảng từ 5% đến 10%) là luôn được phân tích, số còn lại họ không biết sẽ phải làm gì hoặc có thể làm gì với chúng nhưng họ vẫn tiếp tục thu thập rất tốn kém với ý nghĩ lo sợ rằng sẽ có cái gì đó quan trọng đã bị bỏ qua sau này có lúc cần đến nó. Mặt khác, trong môi trường cạnh tranh, người ta ngày càng cần có nhiều thông tin với tốc độ nhanh để trợ giúp việc ra quyết định và ngày càng có nhiều câu hỏi mang tính chất định tính cần phải trả lời dựa trên một khối lượng dữ liệu khổng lồ đã có. Với những lý do như vậy, các phương pháp quản trị và khai thác cơ sở dữ liệu truyền thống ngày càng không đáp ứng được thực tế đã làm phát triển một khuynh hướng kỹ thuật mới đó là kỹ thuật phát hiện tri thức và khai phá dữ liệu (KDD – **K**nowledge **D**iscovery and **D**ata Mining).

Kỹ thuật phát hiện tri thức và khai phá dữ liệu đã và đang được nghiên cứu, ứng dụng trong nhiều lĩnh vực khác nhau ở các nước trên thế giới, tại Việt Nam kỹ thuật này tương đối còn mới mẻ tuy nhiên cũng đang được nghiên cứu và dần đưa vào ứng dụng.

Từ thuở xa xưa, những nhà tiên tri đã giữ một vị trí quan trọng trong cộng đồng. Khi văn minh nhân loại phát triển đã làm gia tăng các mối quan hệ phức tạp của các giai đoạn trong cuộc sống, con người có nhu cầu quan tâm đến tương lai của họ.

Như trình bày trong [2, 3], kỹ thuật dự báo đã hình thành từ thế kỉ thứ 19, tuy nhiên dự báo có ảnh hưởng mạnh mẽ khi công nghệ thông tin phát triển vì bản chất mô phỏng của các phương pháp dự báo rất cần thiết sự hỗ trợ của máy tính. Đến năm những 1950, các lý thuyết về dự báo cùng với các phương pháp luận được xây dựng và phát triển có hệ thống.

Dự báo là một nhu cầu không thể thiếu cho những hoạt động của con người trong bối cảnh bùng nổ thông tin. Dự báo sẽ cung cấp những cơ sở cần thiết cho các hoạch định, và có thể nói rằng nếu không có khoa học dự báo thì những dự định tương lai của con người vạch ra sẽ không có sự thuyết phục đáng kể.

Trong công tác phân tích dự báo, vấn đề quan trọng hàng đầu cần đặt ra là việc nắm bắt tối đa thông tin về lĩnh vực dự báo. Thông tin ở đây có thể hiểu một cách cụ

thể gồm : (1) các số liệu quá khứ của lĩnh vực dự báo, (2) diễn biến tình hình hiện trạng cũng như động thái phát triển của lĩnh vực dự báo và (3) đánh giá một cách đầy đủ nhất các nhân tố ảnh hưởng cả về định lượng lẫn định tính.

Căn cứ vào nội dung phương pháp và mục đích của dự báo, người ta chia dự báo thành hai loại: Phương pháp định tính và phương pháp định lượng.

Phương pháp định tính thường phụ thuộc rất nhiều vào kinh nghiệm của một hay nhiều chuyên gia trong lĩnh vực liên quan. Phương pháp này thường được áp dụng, kết quả dự báo sẽ được các chuyên gia trong lĩnh vực liên quan nhận xét, đánh giá và đưa ra kết luận cuối.

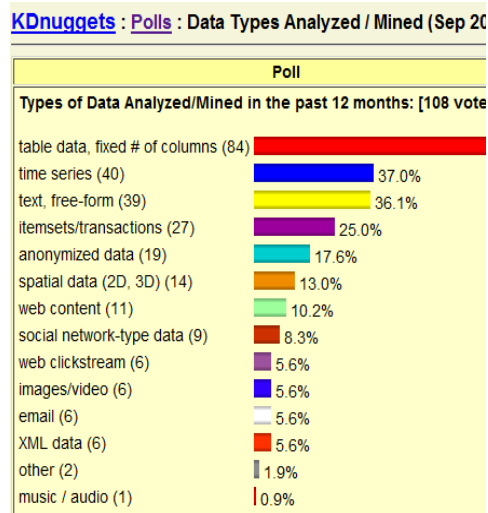
Phương pháp định lượng sử dụng những dữ liệu quá khứ theo thời gian, dựa trên dữ liệu lịch sử để phát hiện chiều hướng vận động của đối tượng phù hợp với một mô hình toán học nào đó và đồng thời sử dụng mô hình đó làm mô hình ước lượng. Tiếp cận định lượng dựa trên giả định rằng giá trị tương lai của biến số dự báo sẽ phụ thuộc vào xu thế vận động của đối tượng đó trong quá khứ. Phương pháp dự báo theo chuỗi thời gian là một phương pháp định lượng.

Phương pháp chuỗi thời gian sẽ dựa trên việc phân tích chuỗi quan sát của một biến duy nhất theo biến số độc lập là thời gian. Giả định chủ yếu là biến số dự báo sẽ giữ nguyên chiều hướng phát triển đã xảy ra trong quá khứ và hiện tại.

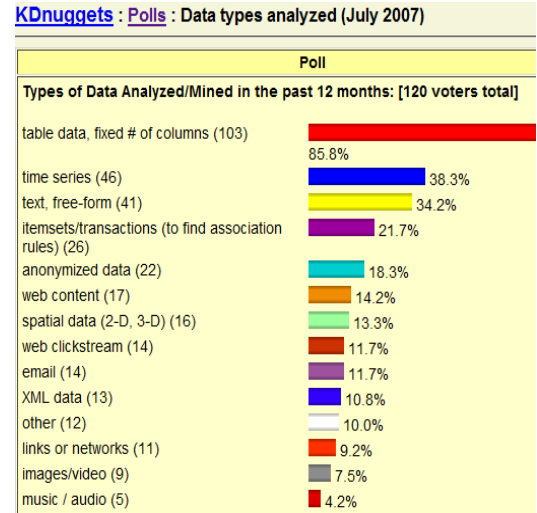
Khóa luận tập trung nghiên cứu mô hình ARIMA để thực hiện phân tích dữ liệu chứng khoán hướng tới việc dự báo chứng khoán. Mô hình ARIMA (**A**uto**R**egressive **I**ntegrate **M**oving **A**verage) do Box-Jenkins đề nghị năm 1976 [6, 11, 13], dựa trên mô hình tự hồi quy AR và mô hình trung bình động MA. ARIMA là mô hình dự báo định lượng theo thời gian, giá trị tương lai của biến số dự báo sẽ phụ thuộc vào xu thế vận động của đối tượng đó trong quá khứ. Mô hình ARIMA phân tích 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. Mô hình kết quả của quá trình trên gồm các tham số thể hiện mức độ tương quan trên dữ liệu, và được chọn để dự báo giá trị tương lai. Giới hạn độ tin cậy của dự báo được tính dựa trên phương sai của sai số dự báo.

1.2. Dữ liệu chuỗi thời gian

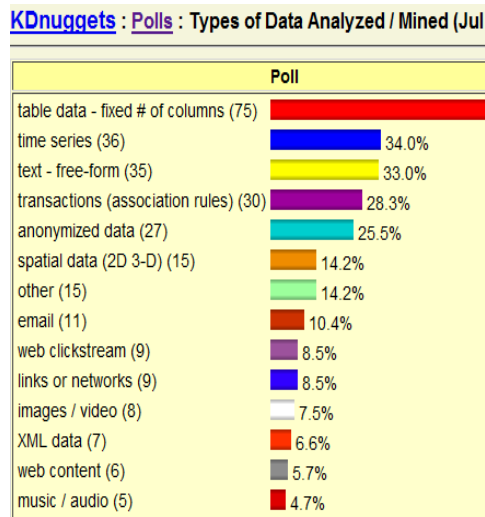
Trong các bài toán dự báo nói chung và các bài toán dự báo tài chính và chứng khoán nói riêng, dữ liệu thường được biểu diễn dưới dạng chuỗi thời gian. Trong các dạng dữ liệu được phân tích thì dữ liệu chuỗi thời gian luôn thuộc top đầu về tính phổ biến. Các bảng thống kê thăm dò về các kiểu dữ liệu được phân tích trong 4 năm 2005-2008¹ (Hình 1) là một minh chứng về điều này.



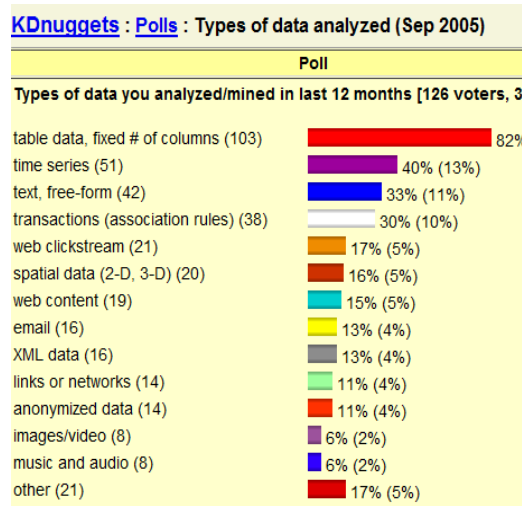
<http://www.kdnuggets.com/polls/2008/data-types-analyzed-data-mined.htm>



http://www.kdnuggets.com/polls/2007/data_types_analyzed.htm



http://www.kdnuggets.com/polls/2006/types_data_analyzed_mined.htm



http://www.kdnuggets.com/polls/2005/data_types.htm

Hình 1. Chuỗi thời gian là kiểu dữ liệu được phân tích phổ biến

¹ <http://www.kdnuggets.com/>

1.2.1. Khái niệm chuỗi thời gian thực

Theo [13, 16], dữ liệu thời gian thực hay chuỗi thời gian là một chuỗi các giá trị của một đại lượng nào đó được ghi nhận là thời gian.

Ví dụ : Số lượng hàng hóa được bán ra trong 12 tháng năm 2009 của một công ty.

Các giá trị của chuỗi thời gian của đại lượng X được kí hiệu là $X_1, X_2, X_3, \dots, X_t, \dots, X_n$ với X là giá trị của X tại thời điểm t .

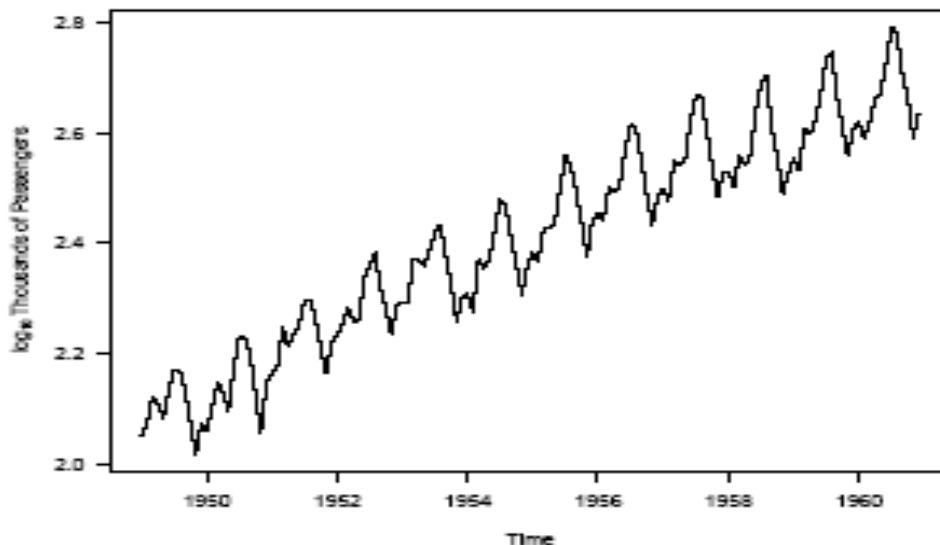
Các thành phần của dữ liệu chuỗi thời gian thực

Các nhà thống kê thường chia chuỗi theo thời gian thành 4 thành phần:

- Thành phần xu hướng dài hạn (long –term trend component)
- Thành phần mùa (seasonal component)
- Thành phần chu kỳ (cyclical component)
- Thành phần bất thường (irregular component)

1.2.2. Thành phần xu hướng dài hạn

Thành phần này dùng để chỉ xu hướng tăng hay giảm của đại lượng X trong thời gian dài. Về mặt đồ thị thành phần này có thể biểu diễn bởi một đường thẳng hay một đường cong trơn.



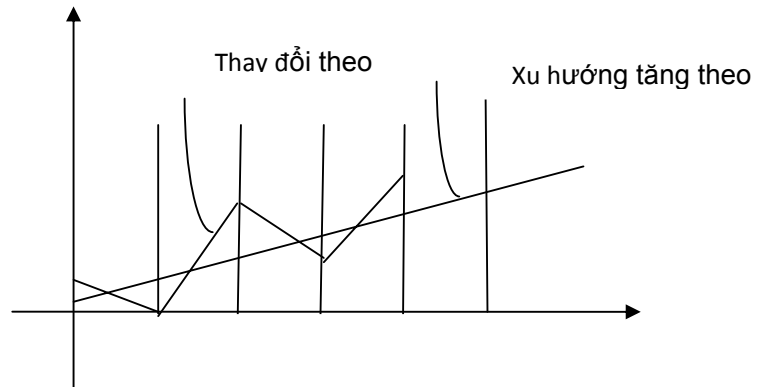
Hình 1a. Xu hướng tăng theo thời gian [16]

1.2.3. Thành phần mùa

Thành phần này dùng để chỉ xu hướng tăng hay giảm của đại lượng X tính theo mùa trong năm (có thể tính theo tháng trong năm)

Ví dụ : Lượng tiêu thụ chất đốt sẽ tăng vào mùa đông và giảm vào mùa hè, ngược lại, lượng tiêu thụ xăng sẽ tăng vào mùa hè và giảm vào mùa đông.

Lượng tiêu thụ đồ dùng học tập sẽ tăng vào mùa khai trường

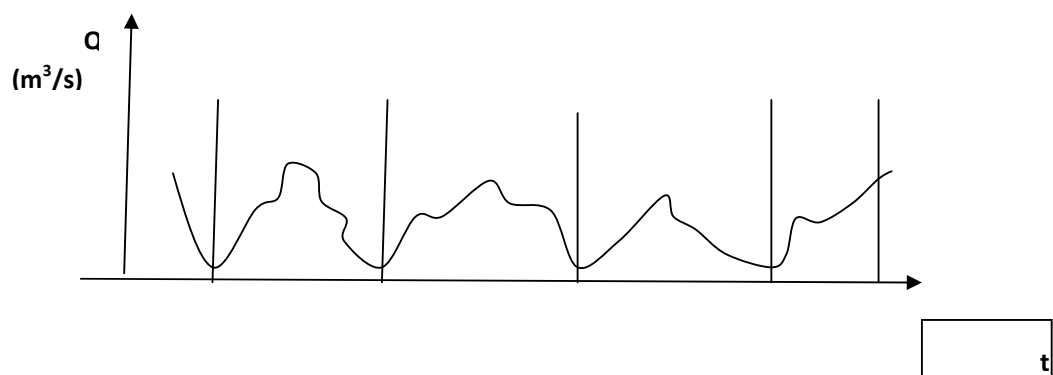


Hình 2. Thành phần mùa [1]

1.2.4. Thành phần chu kỳ

Thành phần này chỉ sự thay đổi của đại lượng X theo chu kỳ. Thành phần này khác thành phần mùa ở chỗ chu kỳ của đại lượng X kéo dài hơn 1 năm. Để đánh giá thành phần này các giá trị của chuỗi thời gian được quan sát hàng năm.

Ví dụ, Lượng dòng chảy đến hồ Trị An từ năm 1959 – 1985



Hình 3. Thành phần chu kỳ [1]

1.2.5. Thành phần bất thường

Thành phần này dùng để chỉ sự thay đổi bất thường của các giá trị trong chuỗi thời gian. Sự thay đổi này không thể dự đoán bằng các số liệu kinh nghiệm trong quá khứ, về mặt bản chất thành phần này không có tính chu kỳ.

CHƯƠNG 2. MÔ HÌNH ARIMA VÀ PHẦN MỀM EVIEWS

2.1. Mô hình ARIMA

2.1.1. Hàm tự tương quan ACF

Hàm tự tương quan đo lường phụ thuộc tuyến tính giữa các cặp quan sát $y(t)$ và $y(t+k)$, ứng với thời đoạn $k = 1, 2, \dots$ (k còn gọi là độ trễ). Với mỗi độ trễ k , hàm tự tương quan tại độ trễ k được xác định qua độ lệch giữa các biến ngẫu nhiên Y_t , Y_{t+k} so với các giá trị trung bình, và được chuẩn hóa qua phương sai.

Dưới đây, giả thiết rằng các biến ngẫu nhiên trong chuỗi dừng thay đổi quanh giá trị trung bình μ với phương sai hằng số δ^2 . Hàm tự tương quan tại các độ trễ khác nhau sẽ có giá trị khác nhau.

Trong thực tế, ta có thể ước lượng hàm tự tương quan tại độ trễ thứ k qua phép biến đổi trung bình của tất cả các cặp quan sát, phân biệt bằng các độ trễ k , với giá trị trung bình mẫu là μ , được chuẩn hóa bởi phương sai σ^2 . Chẳng hạn, cho mỗi chuỗi N điểm, giá trị r_k của hàm tự tương quan tại độ trễ thứ k được tính như sau :

$$r_k = \frac{\frac{1}{N} \sum_{t=1}^{N-k} (y_t - \mu)(y_{t+k} - \mu)}{\delta^2}$$

(1.1)

$$\text{với } \mu = \frac{1}{N} \sum_{t=1}^N (y_t) \quad \delta^2 = \frac{1}{N} \sum_{t=1}^N (y_t - \mu)^2 \quad (1.2)$$

y_t : chuỗi thời gian dừng tại thời điểm t

y_{t+k} : chuỗi thời gian dừng tại thời điểm $t+k$

μ^{\wedge} : giá trị trung bình của chuỗi dừng

r_k : giá trị tương quan giữa y_t và y_{t+k} tại độ trễ k

$r_k = 0$ thì không có hiện tượng tự tương quan

Về mặt lý thuyết, chuỗi dừng khi tất cả các $r_k = 0$ hay chỉ vài r_k khác không. Do chúng ta xem xét hàm tự tương quan mẫu, do đó sai số mẫu sẽ xuất hiện vì vậy, hiện tượng tự tương quan khi $r_k = 0$ theo ý nghĩa thống kê.

Khi hàm tự tương quan ACF giảm đột ngột, có nghĩa r_k rất lớn ở độ trễ 1, 2 và có ý nghĩa thống kê ($|t| > 2$). Những r_k này được xem là những “đỉnh” và ta nói rằng

hàm tự tương quan ACF giảm đột ngột sau độ trễ k nếu không có những “đỉnh” ở độ trễ k lớn hơn k . Hầu hết hàm tự tương quan ACF sẽ giảm đột ngột sau độ trễ 1, 2.

Nếu hàm tự tương quan ACF của chuỗi thời gian không dừng không giảm đột ngột mà trái lại giảm nhanh nhưng đều : không có đỉnh, ta gọi chiều hướng này là “tắt dần”. Xem minh họa trong hình 4, hàm tự tương quan ACF có thể “tắt dần” trong vài dạng sau :

Dạng phân phối mẫu (**hình 4a và hình 4b**)

Dạng sóng sin (**hình 4c**)

Kết hợp cả hai dạng 1 và 2.

Sự khác nhau giữa hiện tượng “tắt dần” nhanh và “tắt dần” chậm đều được phân biệt khá tùy tiện.

2.1.2. Hàm tự tương quan từng phần PACF

Song song với việc xác định hàm tự tương quan giữa các cặp $y(t)$ và $y(t+k)$, ta xác định hàm tự tương quan từng phần cũng có hiệu lực trong việc can thiệp đến các quan sát $y(t+1)$, ..., $y(t+k-1)$. Hàm tự tương quan từng phần tại độ trễ k C_{kk} được ước lượng bằng hệ số liên hệ $y(t)$ trong mỗi kết hợp tuyến tính bên dưới. Sự kết hợp được tính dựa trên tầm ảnh hưởng của $y(t)$ và các giá trị trung gian $y(t+k)$.

$$y(t+k) = C_{k1}y(t+k-1) + C_{k2}y(t+k-2) + \dots + C_{k,k-1}y(t+1) + C_{kk}y(t) + e(t)$$

(1.3)

Giải phương trình hồi quy dựa trên bình phương tối thiểu vì hệ số hồi quy C_{kj} phải được tính ở mỗi độ trễ k , với j chạy từ 1 đến k .

Giải pháp ít tốn kém hơn do Durbin [14] phát triển dùng để xấp xỉ đệ quy hệ số hồi quy cho mô hình ARIMA chuỗi dừng, sử dụng giá trị hàm tự tương quan tại độ trễ k r_k và hệ số hồi quy của độ trễ trước. Dưới đây là phương pháp Durbin sử dụng cho 3 độ trễ đầu tiên.

Độ trễ 1 : Khởi tạo, giá trị của hàm tự tương quan từng phần tại độ trễ 1 có cùng giá trị với hàm tự tương quan tại độ trễ 1 vì không có trung gian giữa các quan sát kết tiếp : $C_{11} = r_1$

Độ trễ 2 : Hai giá trị C_{22} và C_{21} được tính dựa vào hàm tự tương quan r_2 và r_1 , cùng với hàm tự tương quan từng phần trước đó

$$C_{22} = \frac{r_{22} - C_{11}r_1}{1 - C_{11}r_1}$$

$$C_{21} = C_{11} - C_{22}C_{11}$$

Độ trễ 3 : Tương tự, ba giá trị C_{33} , C_{32} , C_{31} được tính dựa vào các hàm tự tương quan trước r_3, r_2, r_1 cùng với các hệ số được tính ở độ trễ thứ 2 : C_{22} và C_{21} .

$$C_{33} = \frac{r_3 - C_{21}r_2 - C_{22}r_1}{1 - C_{22}r_2 - C_{21}r_1}$$

$$C_{32} = C_{21} - C_{33}C_{22}$$

$$C_{31} = C_{22} - C_{33}C_{21}$$

Tổng quan, hàm tự tương quan từng phần được tính theo Durbin :

$$C_{kk} = \frac{r_k - \sum ((C_{k-1,j})r_{k-j})}{1 - \sum (C_{k-1,j})r_j} \quad (1.4)$$

Trong đó :

r_k : Hàm tự tương quan tại độ trễ k

v : Phương sai

C_{kj} : Hàm tự tương quan từng phần cho độ trễ k , loại bỏ những ảnh hưởng của các độ trễ can thiệp.

$$C_{kj} = C_{k-1,j} - (C_{kk}).C_{(k-1,k-j)} \quad k = 2, \dots, j = 1, 2, \dots, k-1$$

$$C_{22} = (r_2 - r_1^2) / (1 - r_1^2)$$

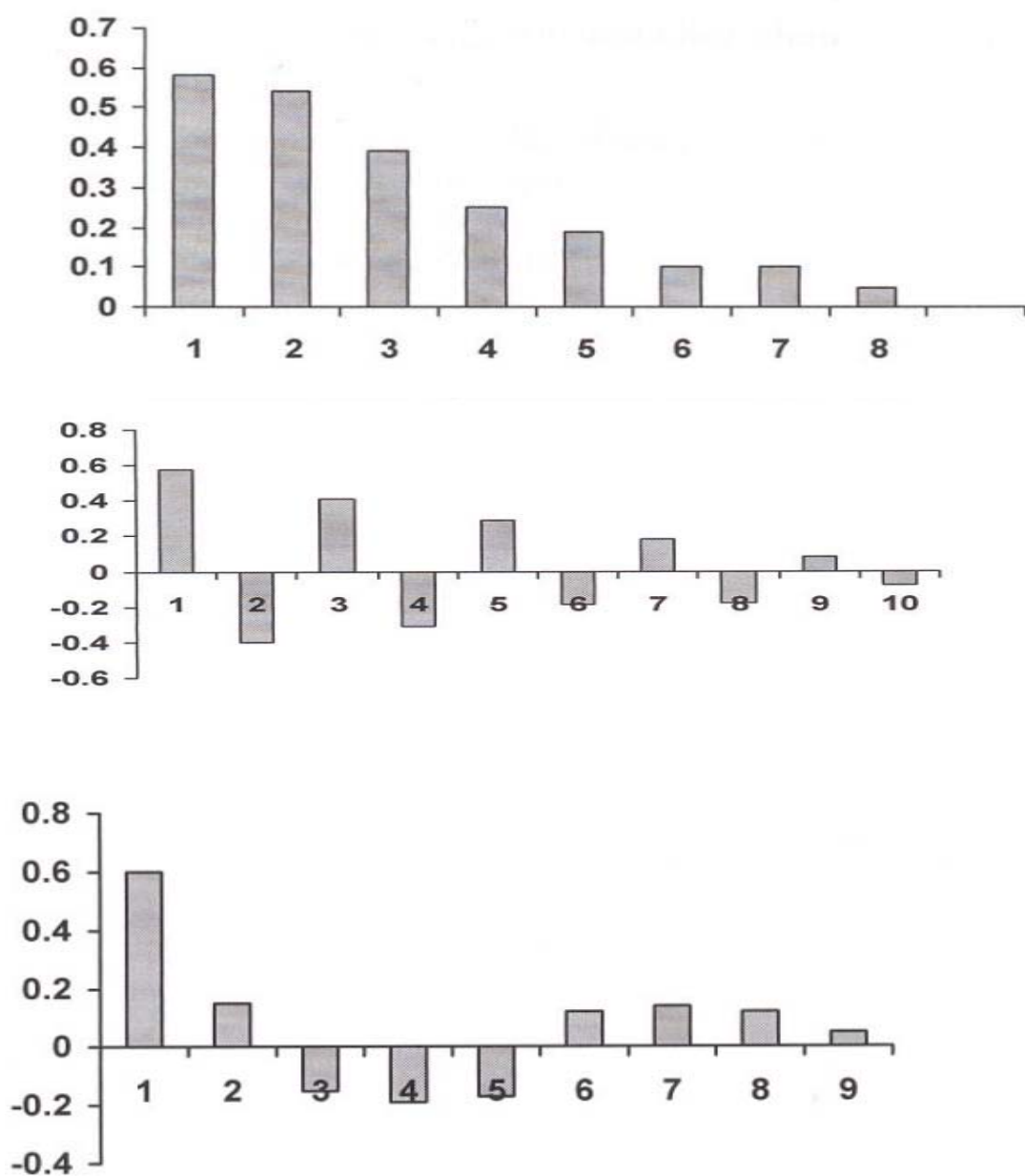
$$C_{11} = r_1$$

Khi độ trễ tăng, số các hệ số tăng theo. Phương pháp của Durbin cho phép việc tính đệ quy dựa vào việc sử dụng kết quả trước đó.

Tóm lại, hàm tự tương quan ACF và hàm tự tương quan từng phần PACF của chuỗi thời gian có các đặc tính khác nhau. Hàm tự tương quan ACF đo mức độ phụ thuộc tuyến tính giữa các cặp quan sát. Hàm tự tương quan từng phần PACF đo mức độ phụ thuộc tuyến tính từng phần. ARIMA khai thác những điểm khác biệt này để xác định cấu trúc mô hình cho chuỗi thời gian.

Xu hướng vận động của hàm tự tương quan từng phần PACF có thể giảm đột ngột (thường sau độ trễ 1 hoặc 2) hay có thể giảm đều. Cũng như hàm tự tương quan

ACF, xu hướng giảm đều của hàm tự tương quan từng phần PACF cũng có các dạng phân phối mũ, dạng sóng hình sin hoặc kết hợp cả 2 dạng này (hình 1-4)



Hình 4 : Ví dụ về chiều hướng giảm đều khác nhau [2]

- a) Dao động hàm mũ tắt dần (Damped Exponential)
- b) Dao động tắt dần theo quy luật số mũ (Damped exponential oscillation)
- c) Dao động sóng tắt dần theo quy luật hình sin (Damped sine wave)

2.1.3. Mô hình AR(p)

Theo [6, 11, 16], ý tưởng chính của mô hình AR(p) là hồi quy trên chính số liệu quá khứ ở những chu kì trước.

$$Y(t) = a_0 + a_1y(t-1) + a_2y(t-2) + \dots + a_p y(t-p) + e(t) \quad (1.5)$$

Trong đó :

$y(t)$: quan sát dừng hiện tại

$y(t-1), y(t-2), \dots$: quan sát dừng quá khứ (thường sử dụng không quá 2 biến này)

a_0, a_1, a_2, \dots : các tham số phân tích hồi quy.

e_t : sai số dự báo ngẫu nhiên của giai đoạn hiện tại. Giá trị trung bình được mong đợi bằng 0.

$Y(t)$ là một hàm tuyến tính của những quan sát dừng quá khứ $y(t-1), y(t-2), \dots$. Nói cách khác khi sử dụng phân tích hồi quy $y(t)$ theo các giá trị chuỗi thời gian dừng có độ trễ, chúng ta sẽ được mô hình AR (yếu tố xu thế đã được tách khỏi yếu tố thời gian, chúng ta sẽ mô hình hóa những yếu tố còn lại – đó là sai số).

Số quan sát dừng quá khứ sử dụng trong mô hình hàm tự tương quan là bậc p của mô hình AR. Nếu ta sử dụng hai quan sát dừng quá khứ, ta có mô hình tương quan bậc hai AR(2).

Điều kiện dừng là tổng các tham số phân tích hồi quy nhỏ hơn 1 :

$$a_1 + a_2 + \dots + a_p < 1$$

Mô hình AR(1) : $y(t) = a_0 + a_1y(t-1) + e(t)$

Mô hình AR(2) : $y(t) = a_0 + a_1y(t-1) + a_2y(t-2) + e(t)$

2.1.4. Mô hình MA(q)

Quan sát dừng hiện tại $y(t)$ là một hàm tuyến tính phụ thuộc các biến sai số dự báo quá khứ và hiện tại. Mô hình bình quân di động là một trung bình trọng số của những sai số mới nhất.

$$y(t) = b_0 + e(t) + b_1e(t-1) + b_2e(t-2) + \dots + b_q e(t-q) \quad (1.6)$$

Trong đó :

$y(t)$: quan sát dừng hiện tại

$e(t)$: sai số dự báo ngẫu nhiên, giá trị của nó không được biết và giá trị trung bình của nó là 0.

$e(t-1), e(t-2), \dots$: sai số dự báo quá khứ (thông thường mô hình sẽ sử dụng không quá 2 biến này)

b_0, b_1, b_2, \dots : giá trị trung bình của $y(t)$ và các hệ số bình quân di động.

q : sai số quá khứ được dùng trong mô hình bình quân di động, nếu ta sử dụng hai sai số quá khứ thì sẽ có mô hình bình quân di động bậc 2 là MA(2).

Điều kiện cần là tổng các hệ số bình quân di động phải nhỏ hơn 1 :

$$b_1 + b_2 + \dots + b_q < 1$$

Mô hình MA(1) : $y(t) = b_0 + e(t) + b_1e(t-1)$

Mô hình MA(2) : $y(t) = b_0 + e(t) + b_1e(t-1) + b_2e(t-2)$

2.1.5. Sai phân I(d)

Chuỗi dừng : Chuỗi thời gian được coi là dừng nếu như trung bình và phương sai của nó không đổi theo thời gian và giá trị của đồng phương sai giữa hai thời đoạn chỉ phụ thuộc vào khoảng cách và 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.

Sai phân chỉ sự khác nhau giữa giá trị hiện tại và giá trị trước đó. Phân tích sai phân nhằm làm cho ổn định giá trị trung bình của chuỗi dữ liệu, giúp cho việc chuyển đổi chuỗi thành một chuỗi dừng.

Sai phân lần 1 (**I(1)**) : $z(t) = y(t) - y(t-1)$

Sai phân lần 2 (**I(2)**) : $h(t) = z(t) - z(t-1)$

2.1.6. Mô hình ARIMA

Mô hình ARMA(p,q) : là mô hình hỗn hợp của AR và MA. Hàm tuyến tính sẽ bao gồm những quan sát dừng quá khứ và những sai số dự báo quá khứ và hiện tại :

$$y(t) = a_0 + a_1y(t-1) + a_2y(t-2) + \dots + a_py(t-p) + e(t) \\ + b_1e(t-1) + b_2e(t-2) + \dots + b_qe(t-q)$$

(1.7)

Trong đó :

$y(t)$: quan sát dừng hiện tại

$y(t-p)$, và $e(t-q)$: quan sát dừng và sai số dự báo quá khứ.

$a_0, a_1, a_2, \dots, b_1, b_2, \dots$: các hệ số phân tích hồi quy

Ví dụ : ARMA(1,2) là mô hình hỗn hợp của AR(1) và MA(2)

Đối với mô hình hỗn hợp thì dạng $(p,q) = (1,1)$ là phổ biến. Tuy nhiên, giá trị p và q được xem là những độ trễ cho ACF và PACF quan trọng sau cùng. Cả hai điều kiện bình quân di động và điều kiện dừng phải được thỏa mãn trong mô hình hỗn hợp ARMA.

Mô hình ARIMA(p,d,q) : Do mô hình Box-Jenkins chỉ mô tả chuỗi dừng hoặc những chuỗi đã sai phân hóa, nên mô hình ARIMA(p,d,q) thể hiện những chuỗi dữ liệu không dừng, đã được sai phân (ở đây, d chỉ mức độ sai phân).

Khi chuỗi thời gian dừng được lựa chọn (hàm tự tương quan ACF giảm đột ngột hoặc giảm đều nhanh), chúng ta có thể chỉ ra một mô hình dự định bằng cách nghiên cứu xu hướng của hàm tự tương quan ACF và hàm tự tương quan từng phần PACF. Theo lý thuyết, nếu hàm tự tương quan ACF giảm đột biến và hàm tự tương quan từng phần PACF giảm mạnh thì chúng ta có mô hình tự tương quan. Nếu hàm tự tương quan ACF và hàm tự tương quan từng phần PACF đều giảm đột ngột thì chúng ta có mô hình hỗn hợp.

Về mặt lý thuyết, không có trường hợp hàm tự tương quan ACF và hàm tự tương quan từng phần cùng giảm đột ngột. Trong thực tế, hàm tự tương quan ACF và hàm tự tương quan từng phần PACF giảm đột biến khá nhanh. Trong trường hợp này, chúng ta nên phân biệt hàm nào giảm đột biến nhanh hơn, hàm còn lại được xem là giảm đều. Do đôi lúc sẽ có trường hợp giảm đột biến đồng thời khi quan sát biểu đồ hàm tự tương quan ACF và hàm tự tương quan từng phần PACF, biện pháp khắc phục là tìm vài dạng hàm dự định khác nhau cho chuỗi thời gian dừng. Sau đó, kiểm tra độ chính xác mô hình tốt nhất.

Mô hình ARIMA (1, 1, 1) : $y(t) - y(t-1) = a_0 + a_1(y(t-1) - y(t-2)) + e(t) + b_1e(t-1)$

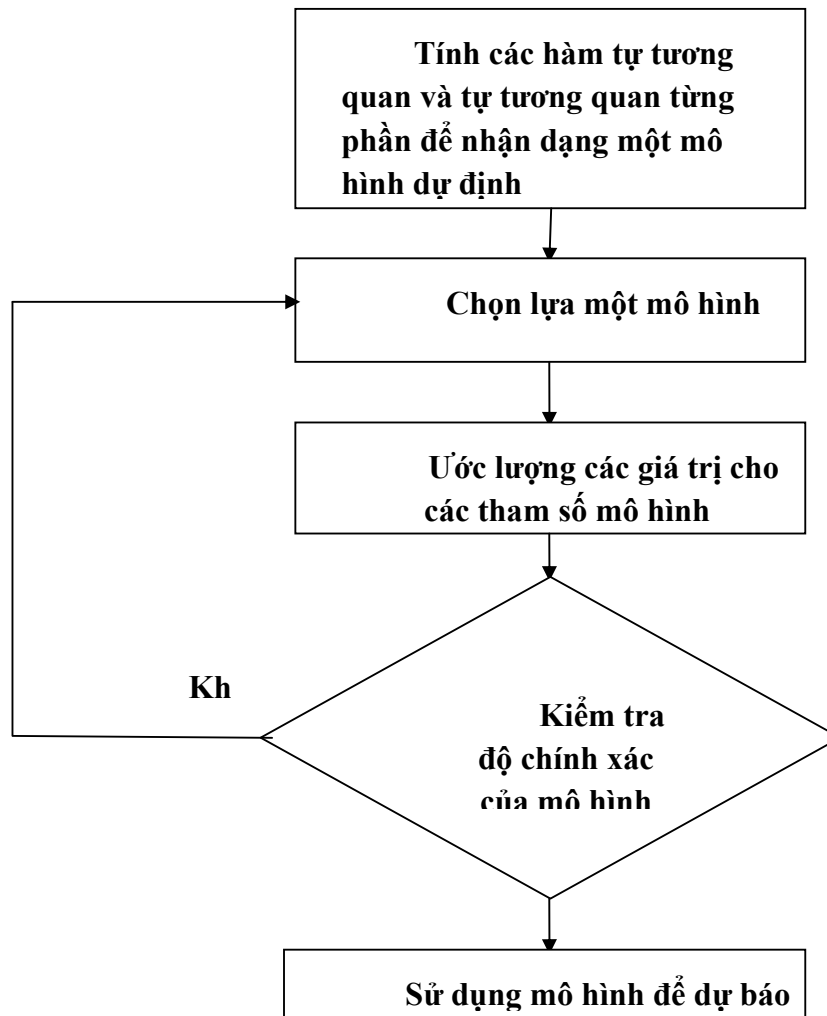
Hoặc $z(t) = a_0 + a_1z(t-1) + e(t) + b_1e(t-1)$,

Với $z(t) = y(t) - y(t-1)$ ở sai phân đầu tiên : $d = 1$.

Tương tự ARIMA(1,2,1) : $h(t) = a_0 + a_1z(t-1) + e(t) + b_1e(t-1)$,

Với $h(t) = z(t) - z(t-1)$ ở sai phân thứ hai : $d = 2$.

Theo [6], trong thực hành d lớn hơn 2 rất ít được sử dụng.



Hình 5. Sơ đồ mô phỏng mô hình Box-Jenkins [3].

2.1.7. Các bước phát triển mô hình ARIMA

Theo [3, 6], phương pháp Box – Jenkins bao gồm các bước chung:

- Xác định mô hình
 - Ước lượng tham số
 - Kiểm định độ chính xác
 - Dự báo.
- **Xác định mô hình** : Mô hình ARIMA chỉ được áp dụng đối với chuỗi dừng. Mô hình có thể trình bày theo dạng AR, MA hay ARMA. Phương pháp xác định mô hình thường được thực hiện qua nghiên cứu chiều hướng biến đổi của hàm tự tương quan ACF hay hàm tự tương quan từng phần PACF.
 - **Chuỗi ARIMA không dừng** : cần phải được chuyển đổi thành chuỗi dừng trước khi tính ước lượng tham số bình phương tối thiểu. Việc chuyển đổi này 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 đều được xem xét tương tự, ngoại trừ các khác biệt ở giá trị trung bình. Nếu việc chuyển đổi này không thành công, sẽ áp dụng tiếp các kiểu chuyển đổi khác (chuyển đổi logarithm chẳng hạn).
 - **Ước lượng tham số** : tính những ước lượng khởi đầu cho các tham số $a_0, a_1, \dots, a_p, b_1, \dots, b_q$ của mô hình dự định. Sau đó xây dựng những ước lượng sau cùng bằng một quá trình lặp.
 - **Kiểm định độ chính xác** : Sau khi các tham số của mô hình tổng quát đã xây dựng, ta kiểm tra mức độ chính xác và phù hợp của mô hình với dữ liệu. Chúng ta kiểm định phần dư $(Y_t - \hat{Y}_t)$ và có ý nghĩa cũng như mối quan hệ các tham số. Nếu bất cứ kiểm định nào không thỏa mãn, mô hình sẽ nhận dạng lại các bước trên được thực hiện lại.
 - **Dự báo** : Khi mô hình thích hợp với dữ liệu đã tìm được, ta sẽ thực hiện dự báo tại thời điểm tiếp theo t . Do đó, mô hình ARMA(p,q) :

$$y(t+1) = a_0 + a_1 y(t) + \dots + a_p y(t - p + 1) + e(t+1) + b_1 e(t) + \dots + b_q e(t - q + 1)$$

(X)

2.2. Phần mềm ứng dụng Eviews

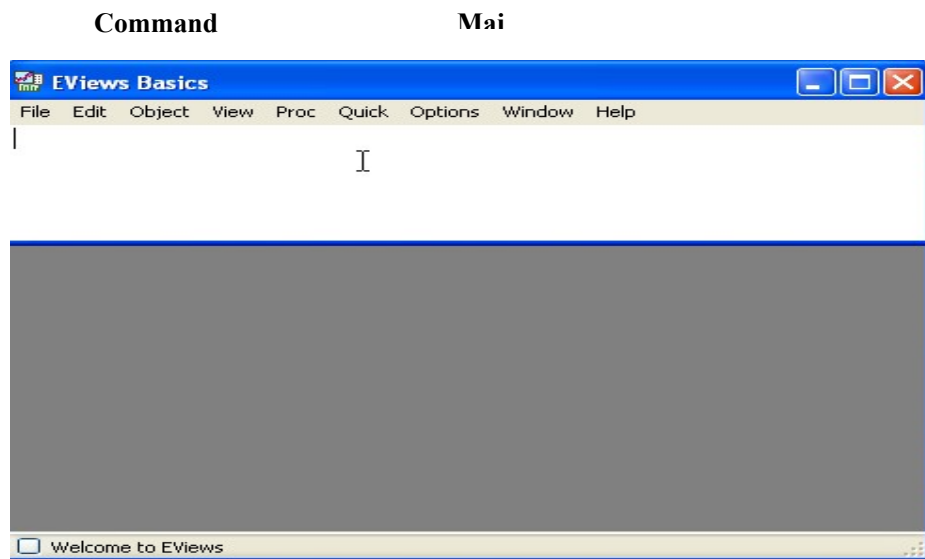
2.2.1. Giới thiệu Eviews

Eviews là một gói phần mềm thống kê cho Windows, được sử dụng chính vào phân tích kinh tế hướng đối tượng chuỗi thời gian. Nó do Quantitative Micro Software (QMS) phát triển. Bản 1.0 được ra đời vào tháng 3 năm 1994 [1].

Phùng Thanh Bình [5] đã giới thiệu tương đối cụ thể về Eviews và các tình huống sử dụng Eviews. Eviews cung cấp các công cụ phân tích dữ liệu phức tạp, hồi quy và dự báo chạy trên Windows. Với Eviews, chúng ta có thể nhanh chóng xây dựng mối quan hệ kinh tế lượng từ dữ liệu có sẵn và sử dụng mối quan hệ này để dự báo các giá trị tương lai. Eviews có thể hữu ích trong tất cả các loại nghiên cứu như đánh giá và phân tích dữ liệu khoa học, phân tích tài chính, mô phỏng và dự báo vĩ mô, dự báo doanh số, và phân tích chi phí. Đặc biệt, Eviews là một phần mềm rất mạnh cho phân tích dữ liệu thời gian.

Eview đưa ra nhiều cách nhập dữ liệu rất thông dụng và dễ sử dụng như nhập bằng tay, từ các file có dưới dạng excel hay text, dễ dàng mở rộng file dữ liệu có sẵn. Eviews trình bày các biểu đồ, kết quả ấn tượng và có thể in trực tiếp hoặc chuyển quan các loại định dạng văn bản khác nhau. Eviews giúp người sử dụng dễ dàng ước lượng và kiểm định các mô hình kinh tế lượng. Ngoài ra, Eviews còn giúp người nghiên cứu có thể xây dựng các file chương trình cho dự án nghiên cứu của mình.

Khi khởi động chương trình có dạng :



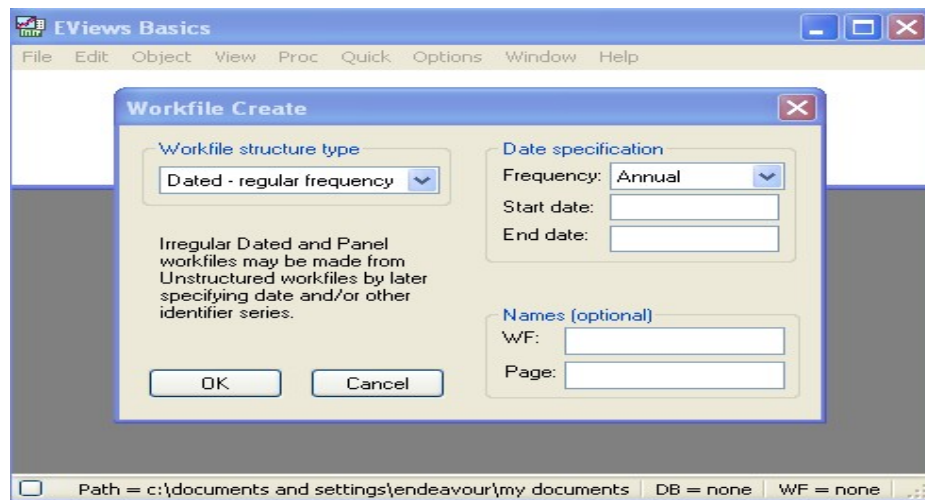
Hình 6. Eviews 5 Users Guide.

Tạo một tập tin Eviews

Có nhiều cách tạo một tập tin mới.

- Eviews sẽ tạo ra một tập tin mới để ta nhập dữ liệu vào một cách thủ công từ bàn phím hoặc copy và paste

File/ New Workfile... từ thực đơn chính để mở hộp thoại **Workfile Create**. Ở góc bên trái mô tả cấu trúc cơ bản của dữ liệu. Ta có thể chọn giữa **Dated-Regular Frequency, Unstructured, Balanced Panel**. Với dữ liệu thời gian ta chọn **Dated-Regular Frequency**, nếu dữ liệu đơn giản ta chọn **Balanced Panel**, các trường hợp khác chọn **Unstructured**.

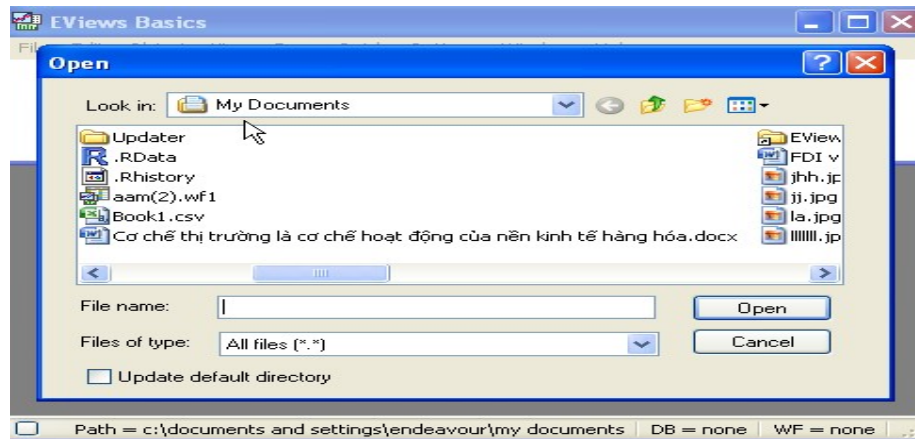


Hình 7. Lựa chọn cấu trúc cơ bản của quá trình tạo Workfile

Nếu là dữ liệu năm, thì ở ô **Frequency** ta chọn **Annual**; ở các ô **Start date** và **End date** ta nhập năm bắt đầu và năm kết thúc của chuỗi dữ liệu. Nếu dữ liệu là quý, thì ở ô **Frequency** ta chọn **Quarrterly**...

- Mở và đọc dữ liệu từ một nguồn bên ngoài (không thuộc định dạng của Eviews) như Text, Excel, Stata

File/open/Foreign Data as Workfile,... để đến hộp thoại Open, chọn **Files of type**



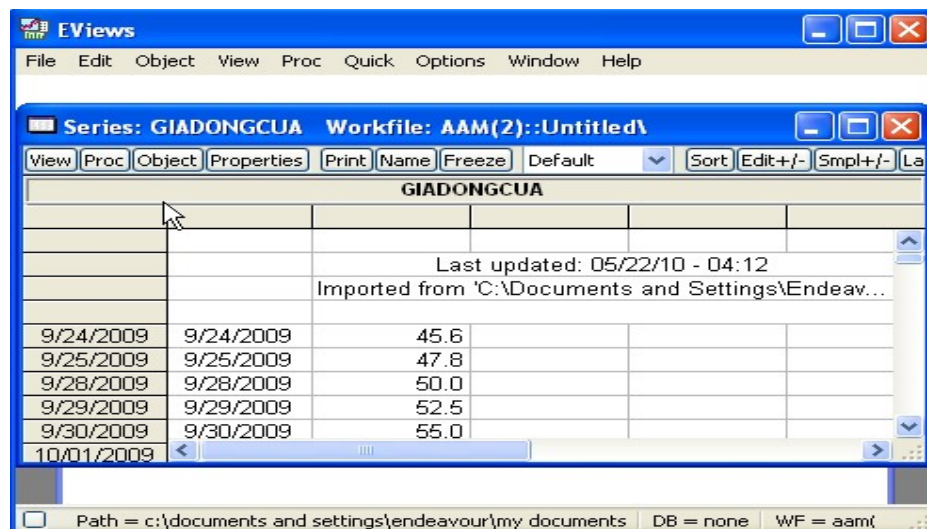
Hình 8. Mở một file có sẵn với Eviews 5

Sau khi tạo một tập tin Eviews, ta lưu lại dưới định dạng Eviews bằng cách chọn **File/Save As...** hay **File/Save...**

Trình bày dữ liệu

- **Trình bày dữ liệu của một chuỗi**

Để xem nội dung của một biến nào đó, ví dụ **giadongcua** trong tập tin. Ta kích đúp vào.

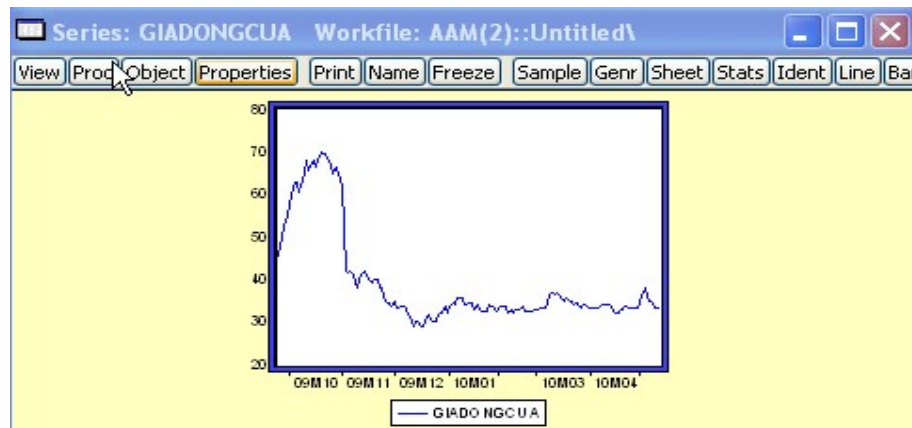


Hình 9. Miêu tả chuỗi dữ liệu

- **Vẽ đồ thị**

Có hai cách biểu hiện đồ thị dạng Line của biến. Thứ nhất, từ chuỗi (lấy chuỗi **giadongcua** làm ví dụ) ta chọn **View/Graph/Line**. Thứ 2, từ cửa sổ

Workfile trên thanh Main menu ta chọn **Quick/Graph/Line Graph,...** rồi nhập tên biến **giadongcua**



Hình 10: Đồ thị của chuỗi GIADONGCUA

Đơn giản để copy đồ thị ra word ta chỉ cần Ctrl + C và paste sang word.

- **Tạo một biến mới**

Eviews hỗ trợ chuyển đổi để tạo biến mới bằng cách click **Genr** rồi gõ hàm chuyển đổi. Thông thường : $\text{loggiadongcua} = \log(\text{giadongcua})$.

- **Biến trễ, tới, sai phân và mùa vụ**

Biến trễ , tới một giai đoạn $(x_{t-1}) : x(-1)$, $(x_{t+1}) : x(+1)$

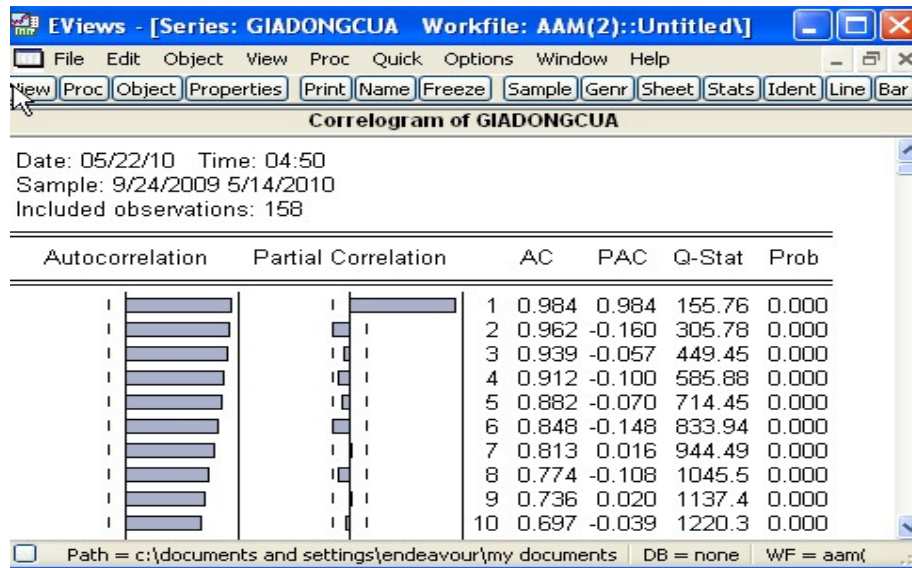
Biến trễ k giai đoạn $(x_{t-k}) : x(-k)$, $(x_{t+k}) : x(+k)$

Sai phân bậc một $(d(x) = x_t - x_{t-1})$

Sai phân bậc k $(d(x,k) = x_t - x_{t-k})$

- **Biểu đồ tương quan.**

View/Correlogram...



Hình 11. Biểu đồ hàm tự tương quan, tự tương quan từng phần.

- Hàm và các phép toán trong Eviews

- Các phép toán số học : +, -, *, /

- Các phép toán chuỗi

Eviews cho phép tính toán hoặc tạo một chuỗi mới từ một hoặc nhiều chuỗi đã có sẵn bằng các toán tử thông thường như trên. Ví dụ :

$$2*y + 3, x/y + z...$$

-Các hàm chuỗi : Hầu hết các hàm Eviews đều bắt đầu bằng ký hiệu @, ví dụ @mean(y) : Giá trị trung bình của chuỗi y

@abs(x) : Hàm giá trị tuyệt đối

@sqrt(x) : Hàm căn bậc hai...

2.2.2. Áp dụng Eviews thi hành các bước mô hình ARIMA

2.2.2.1. Xác định mô hình

- Đưa dữ liệu vào :** Do dữ liệu trong quá trình dự báo sử dụng mô hình ARIMA là đủ lớn, dữ liệu đầu vào được đề xuất : Mở và đọc dữ liệu từ một nguồn bên ngoài (không thuộc định dạng của Eviews) như Text, Excel, Stata

File/open/Foreign Data as Workfile,... để đến hộp thoại Open, chọn **Files of type** (xem thêm ở 2.2.1)

- Kiểm tra tính dừng của chuỗi dữ liệu :** kích đúp vào biến “GiaDongCua”,

View/Graph/line : đưa ra ý tưởng về một chuỗi thời gian là dừng hay không.

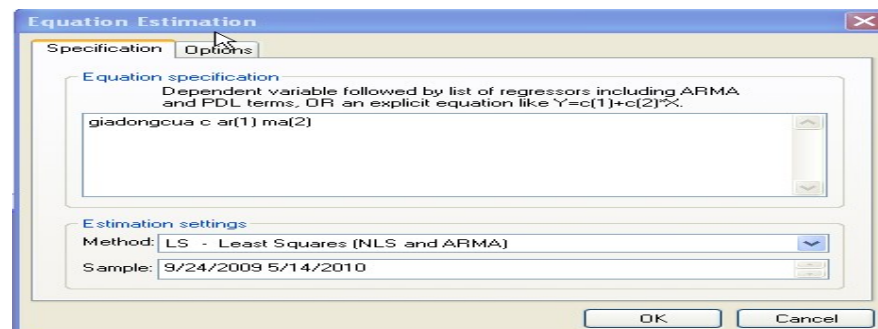
View/Correlogram : Xác định các thành phần p,d,q của mô hình.

2.2.2.2. Ước lượng mô hình, kiểm tra mô hình

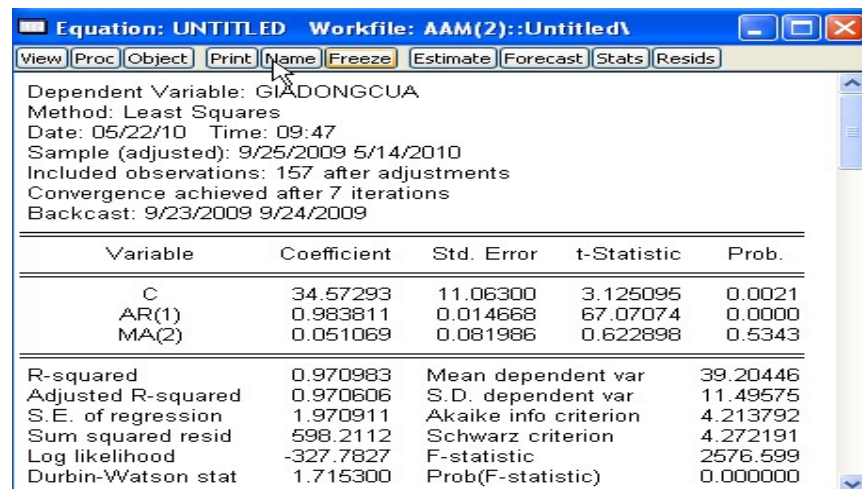
Từ biểu đồ tương quan, xác định được các thành phần p,d,q cho mô hình. Tiếp theo ta xây dựng mô hình theo các bước :

- Chọn **Quick/estimate Equation** gõ vào mục **Equation Specification** mô hình đã được xác định ở 2.2.2.1.

Type : 'giadongcua c ar(1) ma(2)', 'giadongcua c ar(1)', 'giadongcua c ma(2)' (Tùy thuộc vào mô hình đã được xác định)



Hình12. Ước lượng mô hình.



Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	34.57293	11.06300	3.125095	0.0021
AR(1)	0.983811	0.014668	67.07074	0.0000
MA(2)	0.051069	0.081986	0.622898	0.5343

R-squared	0.970983	Mean dependent var	39.20446
Adjusted R-squared	0.970606	S.D. dependent var	11.49575
S.E. of regression	1.970911	Akaike info criterion	4.213792
Sum squared resid	598.2112	Schwarz criterion	4.272191
Log likelihood	-327.7827	F-statistic	2576.599
Durbin-Watson stat	1.715300	Prob(F-statistic)	0.000000

Hình 13. Kết quả quá trình ước lượng

- Chọn **View/Residual tests/correlogram-Q-Statistic** : Dùng để xác định tính nhiễu trắng của mô hình.

Mô hình được gọi là nhiễu trắng(white noise) có trung bình và phương sai không đổi theo thời gian hay hàm tự tương quan và tự tương quan riêng phần dao động quanh một vị trí trung bình của chuỗi [17].

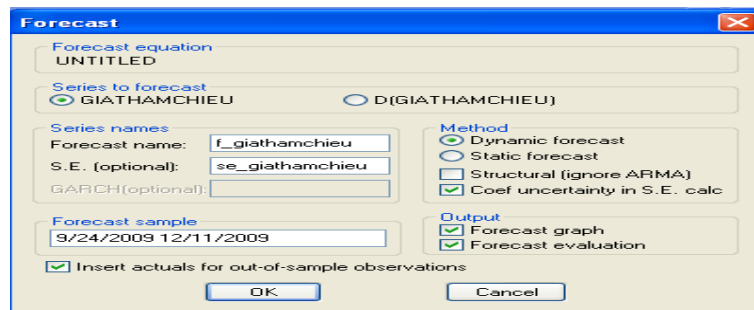
Khi một mô hình được xác định là nhiễu trắng, ta có thể dừng ở mô hình đó mà không cần đến mô hình tiếp theo.

- **Các tiêu chuẩn để đánh giá một mô hình là tốt nhất [18] :**
 - BIC nhỏ (Schwarz criterion được xác định bởi : $n \cdot \text{Log}(\text{SEE}) + K \cdot \text{Log}(n)$)
 - SEE nhỏ
 - R^2 lớn
 - Q-statistics và đồ thị tương quan chỉ ra phần dư là nhiễu trắng.

Sau đó có thể thử với các mô hình khác và so sánh kết quả theo các tiêu chuẩn đánh giá.

2.2.2.3. Dự báo

Tại cửa sổ Equation của phương trình, bấm nút forecast



Hình 14. Chọn các yêu cầu thích hợp cho dự báo

Tóm tắt chương 2

Chương này nhằm giới thiệu về mô hình ARIMA: (1) hàm tự tương quan ACF, (2) hàm tự tương quan từng phần PACF, (3) mô hình thành phần AR(p), (4) mô hình MA(q), sai phân I(d), các bước trong quá trình xây dựng mô hình ARIMA. Giới thiệu sơ bộ về phần mềm ứng dụng Eviews 5.1 phục vụ cho bài toán dự báo bằng mô hình ARIMA.

Chương 3. ÁP DỤNG MÔ HÌNH ARIMA VÀO BÀI TOÁN TÀI CHÍNH, CHỨNG KHOÁN

3.1. Mô hình ARIMA cho dự báo tài chính, chứng khoán

3.1.1. Dữ liệu tài chính

Dữ liệu chúng ta sử dụng là dữ liệu chuỗi thời gian. Đặc điểm chính để phân biệt giữa dữ liệu có phải là thời gian thực hay không đó chính là sự tồn tại của cột thời gian được đính kèm trong đối tượng quan sát. Nói cách khác, dữ liệu thời gian thực là một chuỗi các giá trị quan sát của biến Y :

$Y = \{y_1, y_2, y_3, \dots, y_{t-1}, y_t, y_{t+1}, \dots, y_n\}$ với y_t là giá trị của biến Y tại thời điểm t .

Mục đích chính của việc phân tích chuỗi thời gian thực là thu được một mô hình dựa trên các giá trị trong quá khứ của biến quan sát $y_1, y_2, y_3, \dots, y_{t-1}, y_t$ cho phép ta dự đoán được giá trị của biến Y trong tương lai, tức là có thể dự đoán được các giá trị $y_{t+1}, y_{t+2}, \dots, y_n$.

Trong bài toán của chúng ta, dữ liệu chứng khoán được biết tới như một chuỗi thời gian đa dạng bởi có nhiều thuộc tính cùng được ghi tại một thời điểm nào đó. Với dữ liệu đang xét, các thuộc tính đó là : Open, High, Low, Close, Volume

Open : Giá cổ phiếu tại thời điểm mở cửa trong ngày.

High : Giá cổ phiếu cao nhất trong ngày

Low : Giá cổ phiếu thấp nhất trong ngày

Close : Giá cổ phiếu được niêm yết tại thời điểm đóng cửa sàn giao dịch

Volume : Khối lượng giao dịch cổ phiếu (bán, mua) trong ngày.

3.1.2. Mô hình ARIMA cho bài toán dự báo tài chính

Dựa vào trình tự cơ bản của phương pháp luận (phần 1.7) cùng cấu trúc và hoạt động của mô hình ARIMA trong chương 2. Để áp dụng mô hình ARIMA vào bài toán dự báo tài chính, ta xây dựng mô hình dự báo.

Mô hình gồm 3 quá trình chính :

- Xác định mô hình : Với đầu vào là tập dữ liệu chuỗi thời gian trong tài chính giúp cho việc xác định ban đầu các thành phần trong mô hình p , d , q , S .
- Ước lượng, kiểm tra : Mô hình ARIMA là phương pháp lặp, sau khi xác định các thành phần, mô hình sẽ ước lượng các tham số, sau đó thì kiểm tra độ chính xác của mô hình : Nếu hợp lý, tiếp bước sau, nếu không hợp lý, quay trở lại bước xác định
- Dự báo : Sau khi đã xác định các tham số, mô hình sẽ đưa ra dự báo cho ngày tiếp theo.

3.1.3. Thiết kế mô hình ARIMA cho dữ liệu

Việc thiết kế thành công mô hình ARIMA phụ thuộc vào sự hiểu biết rõ ràng về vấn đề, về mô hình, có thể dựa vào kinh nghiệm của các chuyên gia dự báo...

Trong quá trình tìm hiểu, khóa luận sẽ đưa ra các bước để xây dựng một mô hình như sau :

1. Chọn tham biến
2. Chuẩn bị dữ liệu
 - Xác định tính dừng của chuỗi dữ liệu
 - Xác định yếu tố mùa vụ
 - Xác định yếu tố xu thế
3. Xác định các thành phần p , q trong mô hình ARMA
4. Ước lượng các tham số và chẩn đoán mô hình phù hợp nhất
5. Dự báo ngắn hạn

3.1.3.1 Chọn tham biến

Hướng tiếp cận phổ biến trong dữ liệu tài chính là tập trung xây dựng mô hình dự báo giá cổ phiếu đóng cửa sau khi kết thúc mỗi phiên giao dịch (Close).

3.1.3.2 Chuẩn bị dữ liệu

- Xác định tính dừng của chuỗi dữ liệu : Dựa vào đồ thị của chuỗi và đồ thị của hàm tự tương quan.
- Nếu đồ thị của chuỗi $Y = f(t)$ một cách trực quan nếu chuỗi được coi là dừng khi đồ thị của chuỗi cho trung bình hoặc phương sai không đổi theo thời gian (chuỗi dao động quanh giá trị trung bình của chuỗi)

- Dựa vào đồ thị của hàm tự tương quan ACF nếu đồ thị cho ta một chuỗi giảm mạnh và tắt dần về 0 sau q độ trễ.
- Xác định yếu tố mùa vụ cho chuỗi dữ liệu : Dựa vào đồ thị của chuỗi dữ liệu $Y = f(t)$. (Xem phần chương 1.1)
- Xác định yếu tố xu thế cho chuỗi dữ liệu : Xem lại phần **2.1.2** (Trong giới hạn của khóa luận)

3.1.3.3 Xác định thành phần p, q trong mô hình ARMA

Sau khi loại bỏ các thành phần : Xu thế, mùa vụ, tính dừng thì dữ liệu trở thành dạng thuần có thể áp dụng mô hình ARMA cho quá trình dự báo. Việc xác định 2 thành phần p và q.

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

3.1.3.4 Ước lượng các thông số của mô hình và kiểm định mô hình phù hợp nhất

Có nhiều phương pháp khác nhau để ước lượng. Ở đây, khóa luận tập trung vào : Khi đã chọn được mô hình, các hệ số của mô hình sẽ được ước lượng theo phương pháp tối thiểu tổng bình phương các sai số. Kiểm định các hệ số a, b của mô hình bằng thống kê t. Ước lượng sai số bình phương trung bình của phần dư S^2 :

$$S^2 = \frac{\sum_{t=1}^n e_t^2}{n-r} = \frac{\sum_{t=1}^n (Y_t - \hat{Y}_t)^2}{n-r} \quad \dots$$

Trong đó : $e_t = Y_t - \hat{Y}_t$ = phần dư tại thời điểm t

n = số phần dư

r = tổng số hệ số ước lượng

Tuy nhiên : công thức chỉ đưa ra để tham khảo...Hiện nay phương pháp ước lượng có hầu hết trong các phần mềm thống kê : ET, MICRO TSP và SHAZAM, Eviews...

Nếu phần dư là nhiều trắng thì có thể dừng và dùng mô hình đó để dự báo.

3.1.3.5. Kiểm tra mô hình phù hợp nhất

Dựa vào các kiểm định như

- BIC nhỏ (Schwarz criterion được xác định bởi : $n \cdot \text{Log}(\text{SEE}) + K \cdot \text{Log}(n)$)[]
- SEE nhỏ [19]

$$\text{SEE} = \left[\frac{\sum e_i^2}{n-2} \right]^{1/2}$$

- R^2 lớn : $R\text{-squared} = (\text{TSS}-\text{RSS})/\text{TSS}$ [19]

$$\text{TSS} = \sum (Y_i - \bar{Y})^2, \quad \text{RSS} = \sum (e_i^2) = \sum (\hat{Y}_i - Y_i)^2$$

3.1.3.6 Dự báo ngắn hạn mô hình

Dựa vào mô hình được chọn là tốt nhất, với dữ liệu quá khứ tới thời điểm t , ta sử dụng để dự báo cho thời điểm kế tiếp $t+1$.

3.2. Áp dụng

Ứng dụng mô hình ARIMA vào bài toán dự báo chứng khoán của *của Công ty cổ phần Thủy sản Mekong* (Mã CK : AAM)

Sử dụng Phần mềm EVIEWS 5.1 để dự đoán (Ứng dụng của mô hình ARIMA cho bài toán dự đoán chuỗi thời gian).

Quy trình thực nghiệm được tiến hành như đã mô tả ở 2.2.2.

3.2.1. Môi trường thực nghiệm

Môi trường thực nghiệm Eview 5.1 chạy trên hệ điều hành Window XP SP2, máy tính tốc độ 2*2.0 GHz, bộ nhớ 1GB RAM.

3.2.2. Dữ liệu

Chọn loại dữ liệu dự báo: Dữ liệu được lấy từ

<http://www.cophieu68.com/datametastock.php>

Trong đó ta chọn Cổ phiếu có mã MMA để dự đoán, và sử dụng riêng ***Giá đóng cửa.***

Dữ liệu đầu vào là file.CSV or .dat được lấy từ website xuống.

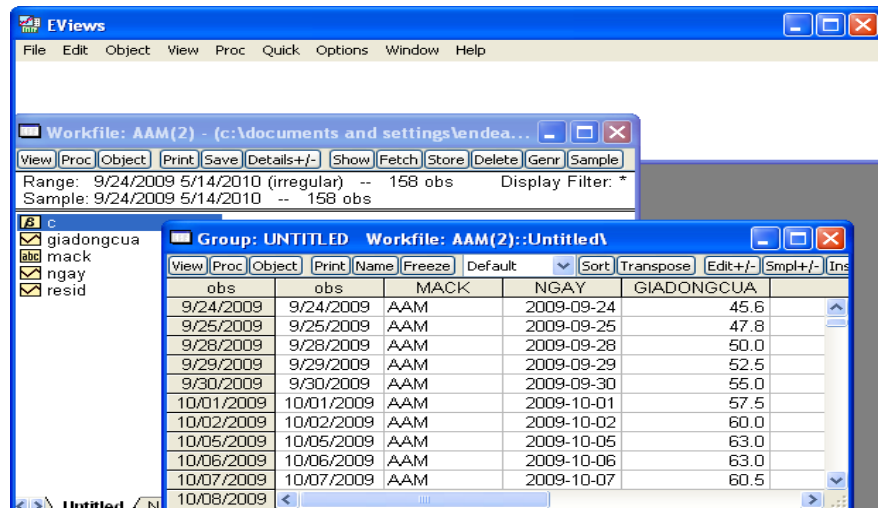
Dữ liệu ở đây có dạng như sau :

MaCK	Ngày	GiaDongCua
AAM	5/14/2010	33.4

AAM	5/13/2010	33.2
AAM	5/12/2010	33.2
AAM	5/11/2010	34.4
AAM	5/10/2010	34.9
AAM	5/7/2010	36.5
...		

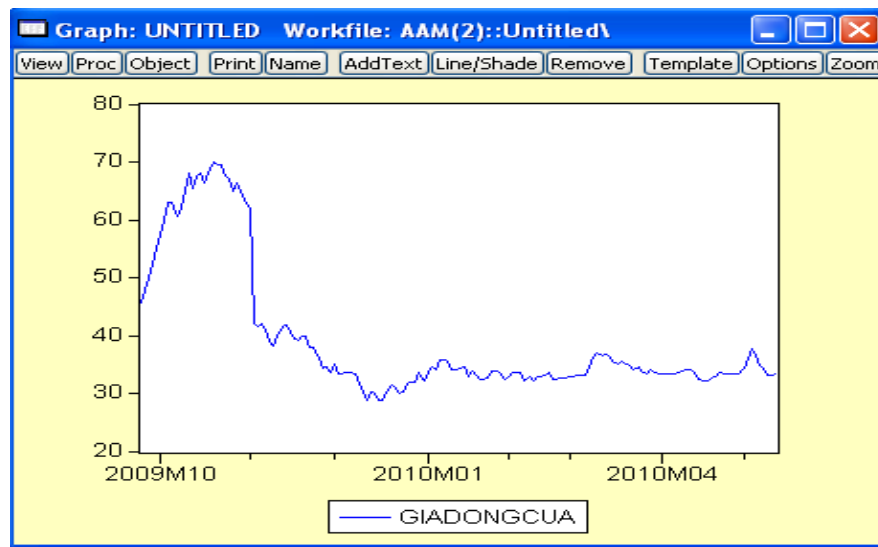
Bảng 1. Dữ liệu đầu vào

Dữ liệu cho quá trình dự báo được bắt đầu từ ngày 24/9/2009 đến ngày 14/5/2010. Ở đây khóa luận chỉ tập trung vào GiaDongCua, và quá trình dự báo sẽ giúp ta xác định được Giá đóng cửa của ngày kế tiếp ngay sau đó.



Hình 15. Chọn GIADONGCUA làm mục tiêu dự báo

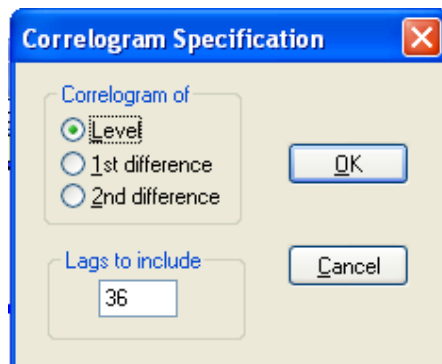
3.2.3. Kiểm tra tính dừng của chuỗi chứng khoán AAM



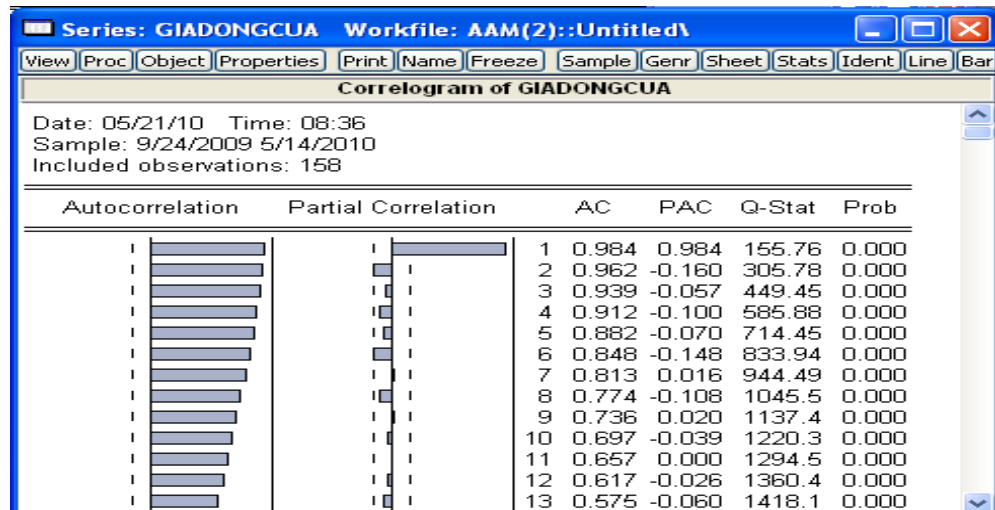
Hình 16. Biểu đồ đóng cửa

3.2.4. Nhận dạng mô hình

Xác định các tham số p , d , q trong ARIMA



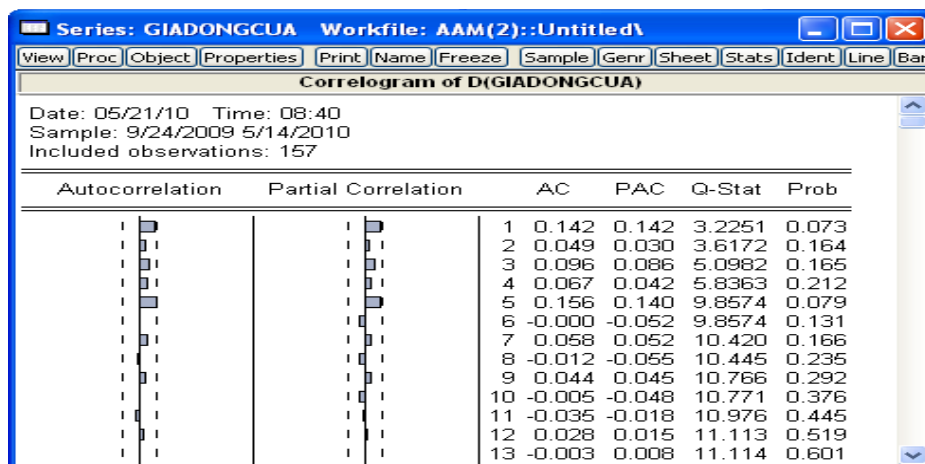
Hình 17. xác định $d = 0,1,2$?



Hình 18. Biểu đồ của SAC và SPAC của chuỗi GIATHAMCHIEU

Nhìn vào hình 3.7, ta thấy biểu đồ hàm tự tương quan ACF giảm dần một cách từ từ về 0. Chuỗi chưa dừng, ta phải sai phân lần 1.

Kiểm tra đồ thị Correlogram của chuỗi sai phân bậc 1.



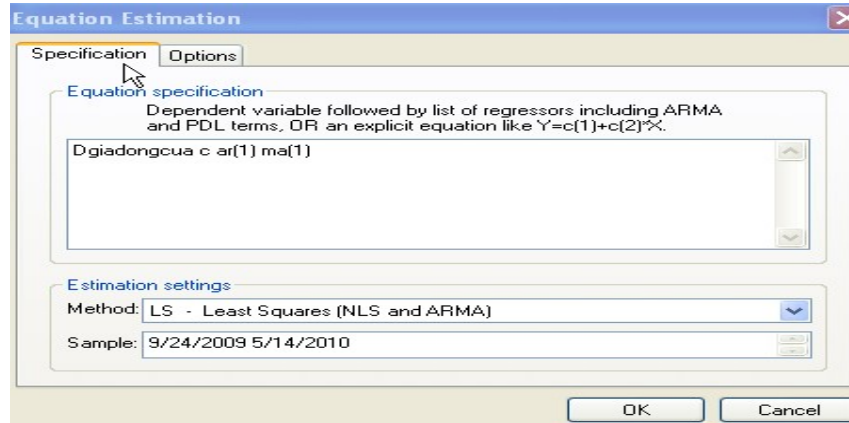
Hình 19. Biểu đồ của SPAC và SAC ứng với d=1

Như vậy sau khi lấy sai phân bậc 1 chuỗi đã dừng: $\rightarrow d=1$, ACF tắt nhanh về 0 sau 1 độ trễ $\rightarrow q=1$, PAC giảm nhanh về 0 sau 1 độ trễ: $\rightarrow p=1$

3.2.5. Ước lượng và kiểm định với mô hình ARIMA

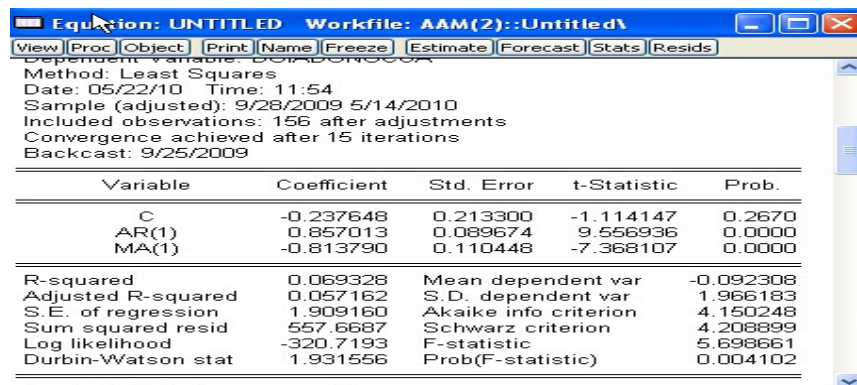
Xây dựng mô hình ARIMA(1,1,1)

Chọn **Quick/Estimate Equation**, sau đó gõ "dgiathamchieu c ar(1) ma(1)",



Hình 20. Ước lượng mô hình ARIMA(1,1,1)

Click OK, kết quả là :

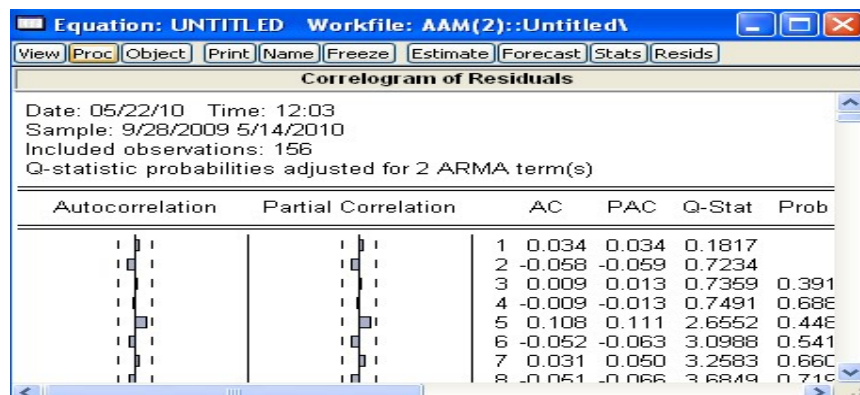


Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.237648	0.213300	-1.114147	0.2670
AR(1)	0.857013	0.089674	9.556936	0.0000
MA(1)	-0.813790	0.110448	-7.368107	0.0000

Statistic	Value	Statistic	Value
R-squared	0.069328	Mean dependent var	-0.092308
Adjusted R-squared	0.057162	S.D. dependent var	1.966183
S.E. of regression	1.909160	Akaike info criterion	4.150248
Sum squared resid	557.6687	Schwarz criterion	4.208899
Log likelihood	-320.7193	F-statistic	5.698661
Durbin-Watson stat	1.931556	Prob(F-statistic)	0.004102

Hình 21. Kết quả mô hình ARIMA(1,1,1)

Chọn **“View/Residual tests/Correlogram-Q-Statistic”**



Hình 22 : Kiểm tra phần dư có nhiễu trắng

Như vậy, sai số của mô hình ARIMA(1,1,1) là một chuỗi dừng và nó có phân phối chuẩn. Sai số này là nhiễu trắng.

Ta có bảng xác định các tiêu chuẩn đánh giá sau khi đã thử với một vài mô hình khác nhau :

Mô hình ARIMA	BIC	Adjusted R ²	SEE
ARIMA(1,0,0)	4.24	0.97	1.967
ARIMA(2,1,1)	4.26	0.004	1.96
ARIMA(1,1,1)	4.20	0.57	1.909
ARIMA(4,2,1)	4.26	0.44	1.957

Bảng 2 : Tiêu chuẩn đánh giá các mô hình ARIMA

3.2.6 Thực hiện dự báo

Tại cửa sổ Equation ấn nút Forecast

Hình 23. Dự báo

Tại Forecast sample : ta chỉnh ngày dự báo : 14/5/2010 – 20/5/2010

Kết quả là :

obs	GIADONGCUAF	GIADONG...	NGAY
4/28/2010	33.40000	33.4	2010-04-28
4/29/2010	33.40000	33.4	2010-04-29
5/04/2010	34.30000	34.3	2010-05-04
5/05/2010	36.00000	36.0	2010-05-05
5/06/2010	37.80000	37.8	2010-05-06
5/07/2010	36.50000	36.5	2010-05-07
5/10/2010	34.90000	34.9	2010-05-10
5/11/2010	34.40000	34.4	2010-05-11
5/12/2010	33.20000	33.2	2010-05-12
5/13/2010	33.20000	33.2	2010-05-13
5/14/2010	33.11907	33.4	2010-05-14
5/15/2010	33.05020	NA	2010-05-15
5/16/2010	32.99160	NA	2010-05-16
5/17/2010	32.94174	NA	2010-05-17
5/18/2010	32.89932	NA	2010-05-18
5/19/2010	32.86322	NA	2010-05-19
5/20/2010	32.83250	NA	2010-05-20

Hình 24. Kết quả của bảng thống kê dự báo.

→ Ta có kết quả dự báo của 3 ngày 14/5/2010 – 20/5/2010

<i>Ngày</i>	<i>Giá thực tế</i>	<i>Giá dự báo</i>	<i>Đánh giá</i>
17/05/2010	33.5	32.94174	-0.55826
18/05/2010	33.2	32.89932	-0.30068
19/05/2010	32.5	32.86322	0.36322
20/05/2010	33.2	32.83250	-0.3675

Bảng 3. Đánh giá dự báo

Qua thực nghiệm dự báo được 4 ngày từ ngày 17/05 – 20/05/2010, chúng ta nhận thấy kết quả đưa ra khá chính xác so với giá thực tế của mã chứng khoán AAM.

Tuy số lượng ngày dự báo thử nghiệm chưa nhiều song có thể nhận định rằng mô hình ARIMA(1,1,1) là khá phù hợp để dự báo mã CK AAM.

Tóm tắt chương 3

Chương 3 giới thiệu về môi trường thực nghiệm phần mềm, dữ liệu đầu vào là giá chứng khoán của công ty với mã AAM (chọn GiaDongCua làm biến dự báo). Khóa luận đã tiến hành từng bước quá trình thi hành dự báo twf dữ liệu như đã nêu ở

chương 2. Đánh giá sơ bộ thành công của mô hình được chọn : Mô hình được chọn dự báo khá chính xác.

KẾT LUẬN

Qua thời gian nghiên cứu để thực hiện khóa luận tốt nghiệp, em đã nắm được quy trình xây dựng mô hình ARIMA cho dữ liệu tài chính và áp dụng mô hình này vào bài toán thực tế - bài toán dự báo tài chính. Những kết quả chính mà khóa luận đã đạt được có thể tổng kết như sau :

- Nghiên cứu một số nội dung lý thuyết cơ bản về chuỗi thời gian, về mô hình ARIMA, về công cụ Eviews để có thể áp dụng được Eviews thi hành mô hình ARIMA trong dự báo tài chính, chứng khoán.
- Nắm được quy trình dùng phần mềm Eviews thi hành mô hình ARIMA cho dữ liệu thời gian thực (với 4 bước cơ bản) tính toán giá trị dự báo dữ liệu tài chính, chứng khoán.
- Thực hiện quy trình sử dụng phần mềm Eviews thi hành mô hình ARIMA cho dữ liệu mã cổ phiếu mã CK AAM để dự báo ngắn hạn giá cổ phiếu.

Bên cạnh những kết quả đã đạt được, còn có những vấn đề mà thời điểm này, khóa luận chưa giải quyết được:

- Áp dụng với chuỗi dữ liệu có tính xu thế.
- Thuật toán để ước lượng cũng như đánh giá còn nhiều hạn chế.
- Đây chỉ là mô hình phân tích kỹ thuật, chưa thể dự báo một cách chính xác, bởi chỉ phụ thuộc vào một biến – Thời gian, trong khi quá trình dự báo phụ thuộc vào nhiều yếu tố.

Những nội dung cần nghiên cứu phát triển để tiếp tục nội dung khóa luận:

- Xây dựng mô hình ARIMA đa biến : chỉ số của giá chứng khoán phụ thuộc vào nhiều biến khác nhau.
- Giải quyết yếu tố xu thế cho chuỗi dữ liệu

TÀI LIỆU THAM KHẢO

Tài liệu tham khảo tiếng Việt

- [1]. Đặng Thị Ánh Tuyết. *Tìm hiểu và ứng dụng một số thuật toán khai phá dữ liệu time series áp dụng trong bài toán dự báo tài chính*. Khóa luận tốt nghiệp đại học hệ chính quy, khoa Công nghệ thông tin – Đại học Công Nghệ - Đại học Quốc Gia Hà nội, 2009.
- [2]. Nguyễn Thị Hiền Nhã. *Sử dụng mô hình ARIMA cho việc giải quyết bài toán dự báo tỷ giá*. Luận văn thạc sĩ tin học, Đại học Khoa Học Tự Nhiên – Đại Học Quốc Gia TP.HCM, 2002.
- [3]. Nguyễn Thị Thanh Huyền, Nguyễn Văn Huân, Vũ Xuân Nam. *Phân tích và dự báo kinh tế*, Đại Học Thái Nguyên, <http://ictu.edu.vn/LinkClick.aspx?fileticket=EKrb8h5MaQ%3D&tabid=212&mid=910>.
- [4]. Damodar N Gujarati. *Kinh tế lượng căn bản*. Chương 21, 22
- [5]. Phùng Thanh Bình. *Hướng dẫn sử dụng Eviews 5.1*

Tài liệu tham khảo tiếng Anh

- [6] Boris Kovalerchuk and Evgenii Vityaev (2001). *Data Mining in Finance: Advances in Relational and Hybrid Methods*, Kluwer Academic Publishers, Boston, Dordrecht - London, 2001.
- [7] Jamie Monogan. *ARIMA Estimation adapting Maximum Likelihood to the special Issues of Time Series*.
- [8] Cao Hao Thi, Pham Phu, Pham Ngoc Thuy. *Application of ARIMA model for testing “serial independence” of stock prices at the HSEC*, The Joint 14th Annual PBFEA and 2006 Annual FeAT Conference, Taipei, Taiwan, July, 2006.
- [9] Robert Yaffee and Monnie McGee. *Time series Analysis and forecasting*.
- [10] Box G E P & Jenkins G M. *Time series analysis : Forecasting and control*. San Francisco, CA: Holden-day, 1970.
- [11] Roy Batchelor. *Box-Jenkins Analysis*. Cass Business School, City of Lodon
- [12]. http://en.wikipedia.org/wiki/Time_series. Time series

- [13] Ramasubramanian V.I.A.S.R.I. *Time series analysis*, Library Avenue, New Delhi-110 012
- [14]. <http://www.pstat.ucsb.edu/faculty/feldman/174-03/lectures/113.pdf>. Sample PACF; Durbin - Levinson algorithm.
- [15]. <http://adt.curtin.edu.au/theses/available/adt-WCU20030818.095457/unrestricted/07Chapter6.pdf>. Chapter six Univariate ARIMA models
- [16]. Ross Ihaka. *Time Series Analysis*, Lecture Notes for 475.726, Statistics Department, University of Auckland, 2005.
- [17]. <http://www.barigozzi.eu/ARIMA.pdf>. *ARIMA estimation theory and applications*
- [18]. <http://www.hkbu.edu.hk/~billhung/econ3600/application/app05/app05.html>. *ARIMA models*.
- [19]. <http://www.stata.com/statalist/archive/2006-06/msg00554.html>. *R-Squared with ARIMA*
- [20]. http://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average: Autoregressive integrated moving average.