ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN TOÁN ỨNG DỤNG VÀ TIN HỌC

——————o0o——————

BÁO CÁO BÀI TẬP LỚN

CHUỖI THỜI GIAN

Đề tài

MÔ HÌNH LAI ARIMA/WAVELETS-ARIMA TRONG DỰ BÁO GIÁ CHỨNG KHOÁN

Giảng viên hướng dẫn: TS NGUYỄN THỊ NGỌC ANH

Nhóm sinh viên thực hiện:

Nguyễn Thị Quý	**20185396**
Hoàng Phương Cúc	**20185332**
Nguyễn Ngọc Thìn	**20185408**

Hà Nội, tháng 1 năm 2022

Mục lục

Danh sách hình vẽ

Danh sách bảng

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ

Phần đánh giá của giảng viên chấm bài:

.....
.....
.....
.....
.....
.....
.....
.....
.....

Hà Nội, ngày..... tháng..... năm.....
Giảng viên chấm bài

Phần đánh giá của giảng viên hướng dẫn:

.....
.....
.....
.....
.....
.....
.....
.....
.....

Hà Nội, ngày..... tháng..... năm.....
Giảng viên hướng dẫn

.

LỜI CẢM ƠN

Để có thể hoàn thành được đề tài này, trước hết chúng em xin gửi lời cảm ơn sâu sắc tới TS.Nguyễn Thị Ngọc Anh,... Do thời gian thực hiện đề tài vẫn còn ngắn và kiến thức của chúng em vẫn còn nhiều hạn chế nên đồ án này vẫn còn nhiều thiếu sót nên mong cô sẽ góp ý để chúng em hoàn thiện hơn trong thời gian tới.

Chúng em xin chân thành cảm ơn!

Hà Nội, ngày 08 tháng 02 năm 2022

Nhóm 2

MỞ ĐẦU

Tổng quan đề tài

Trong cuộc sống hiên đại ngày nay, với các hoạt động riêng tư của mỗi cá nhân, trong các hoạt động của cơ quan Chính phủ, các tập đoàn, các công ty lớn, nhỏ và ngay cả các quốc gia hùng mạnh, việc đoán trước xu thế tương lai đóng vai trò cực kỳ quan trọng trong công tác quản lý, điều hành, hoạch định chính sách. Có nhiều cách cách để dự đoán các xu thế, trong bài báo cáo này em xin trình bày về mô hình phân tích, dự báo chuỗi thời gian (Time serial) một phương pháp toán học, theo đó người ta có thể rút ra được những quy luật của một quá trình được "quan sát" thông qua chuỗi số liệu. Dự báo chứng khoán là một ví dụ điển hình, ta có thể áp dụng các mô hình chuỗi thời gian để dự báo chứng khoán.

Mục tiêu nghiên cứu

Mục tiêu nghiên cứu của bài báo cáo là tìm ra và cải thiện phương pháp phù hợp để có thể dự báo tốt trong dự báo chứng khoán.

Lý do chọn đề tài

Một trong các mô hình phổ biến nhất trong chuỗi thời gian hiện nay chính là mô hình ARIMA, mô hình này đã được giới thiệu bởi Box và Jenkins năm 1970. Ngày nay mô hình đã trở nên phổ biến vì tính dự báo chính xác với dự báo ngắn hạn và nhận được nhiều sự nghiên cứu quan tâm cách cải thiện, nâng cấp và khắc phục các nhược điểm. Báo cáo này chúng em xin phép trình bày về áp dụng mô hình dự báo ARIMA và kết hợp với phép biến đổi Wavelet để dự báo chứng khoán.

Phương pháp nghiên cứu

Trong bài báo cáo này, nhóm chúng em sử dụng phương pháp phân tích và tổng hợp lý thuyết dựa trên tài liệu, bài báo có sẵn về các vấn đề liên quan. Từ đó xây dựng mô hình phù hợp cho thực nghiệm.

Bố cục đồ án

Nội dung bài báo cáo gồm 3 chương chính:

Chương 1: Nêu ra các lý thuyết chung về mô hình chuỗi thời gian cũng như thành phần và ứng dụng, giới thiệu mô hình AR, MA và ARMA; lý thuyết về Wavelet.

Chương 2: Đưa ra mô hình ARIMA sử dụng để dự báo dữ liệu, mô hình ARIMA kết hợp phép biến đổi Wavelet và các độ đo để đánh giá mô hình.

Chương 3: Áp dụng dữ liệu thực tế vào mô hình đã được nêu ở trên, so sánh, đánh giá mô hình và đề xuất giải pháp cải tiến hơn.

Từ khóa: **chuỗi thời gian, ARIMA, Wavelet, dự báo, chứng khoán**.

Chương 1

CƠ SỞ LÍ THUYẾT

1.1 Chuỗi thời gian

1.1.1 Giới thiệu chuỗi thời gian

Trong thời đại 4.0 hiên nay, lĩnh vực máy học đã trở nên rất phổ biến được ứng dụng rất nhiều trong xử lý các bài toán trí tuệ nhân tạo. Trong đó, chuỗi thời gian là một khía cạnh rất quan trọng nhưng thường bị bỏ qua. Chuỗi thời gian được ứng dụng trong nhiều lĩnh vực như dự báo kinh tế, phân tích giá cổ phiếu, ... Chuỗi thời gian rất hữu ích trong việc dự báo cũng như phân tích trong các lĩnh vực thực tế ngày nay song thành phần thời gian lại trở thành khái niệm đầy thách thức để xử lý.

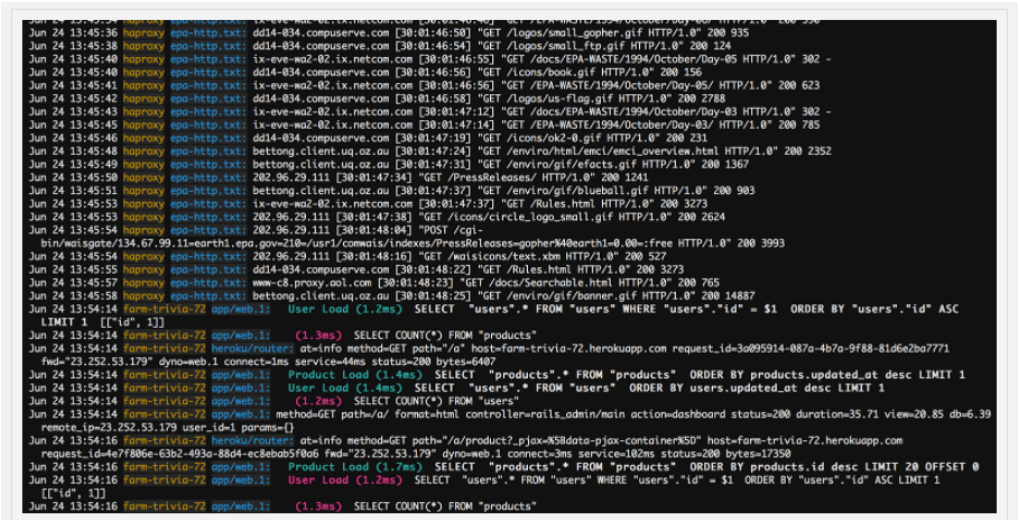
Định nghĩa 1.1. *Chuỗi thời gian* được định nghĩa là chuỗi hoặc tập các điểm dữ liệu được thêm vào theo thời gian, trong đó tập dữ liệu là một tập hợp các quan sát.

Các chỉ số kinh tế, diễn biến sức khỏe của bệnh nhân, chỉ số về thời tiết - tất cả đều là dữ liệu chuỗi thời gian. Dữ liệu chuỗi thời gian cũng có thể là dữ liệu mạng, dữ liệu cảm biến, giám sát hiệu suất ứng dụng và các loại dữ liệu phân tích khác.

Ví dụ: Trong đầu tư, một chuỗi thời gian theo dõi chuyển động của các điểm dữ liệu, chẳng hạn như giá của chứng khoán trong một khoảng thời gian cụ thể với các điểm dữ liệu được ghi lại trong khoảng thời gian đều đặn.

Hình 1.1: Ví dụ về chuỗi thời gian biến động theo khoảng thời gian đều đặn

Mặt khác, dữ liệu chuỗi thời gian có thể được ghi lại bất cứ khi nào xảy ra - bất kể khoảng thời gian nào, chẳng hạn như trong nhật ký. Nhật ký là một sổ đăng ký các sự kiện, quy trình, thông báo và giao tiếp giữa các ứng dụng phần mềm và hệ điều hành. Mỗi tệp thực thi tạo ra một tệp nhật ký nơi mà tất cả các hoạt động được ghi nhận. Dữ liệu nhật ký là một nguồn ngữ cảnh quan trọng để phân loại và giải quyết các vấn đề. Ví dụ, trong mạng, nhật ký sự kiện giúp cung cấp thông tin về lưu lượng mạng, việc sử dụng và các điều kiện khác.



Hình 1.2: Ví dụ về chuỗi thời gian biến động theo khoảng thời gian bất kỳ

Một chuỗi thời gian bao gồm 4 thành phần chính:

- Trend (Xu hướng - T): Thể hiện sự tăng hoặc giảm trong dài hạn của dữ liệu. Xu hướng có thể tăng, giảm, tuyến tính hoặc phi tuyến tính.
- Seasonality (Thời vụ - S): Thể hiện sự thay đổi ngắn hạn của dữ liệu xảy ra do các yếu tố theo mùa. Tính thời vụ có thể dự đoán được bởi nó xuất hiện trong trường hợp mà dữ liệu trải qua những thay đổi thường xuyên.
- Cyclicity (Chu kì - C): Thể hiện biến động dữ liệu trong khoảng thời gian dài hạn (thường là hơn 1 năm).
- Irregularity (Bất thường - I): Thể hiện việc biến động không báo trước hoặc bất thường của một chuỗi thời gian và không thể dự đoán trước được thành phần này.

Những thành phần này kết hợp với nhau trong chuỗi thời gian bằng nhiều cách thức khác nhau, chẳng hạn chuỗi thời gian $\{X_t\}$ được mô tả là :

- $X_t = T * P * S * I$: gọi là mô hình tích.
- $X_t = T + P + S + I$: gọi là mô hình tổng.
- $X_t = T * P * S + I$: gọi là mô hình hỗn hợp.

Do vậy, để phân tích, nghiên cứu hành vi cũng như dự báo biến động của chuỗi thời gian thì cần thiết phải ước lượng các thành phần nói trên trong chuỗi thời gian và cách thức kết hợp chúng với nhau trong chuỗi.

1.1.2 Tính dừng của một chuỗi thời gian và hàm ACF

Một việc quan trọng trong phân tích chuỗi thời gian là việc lựa chọn mô hình cho dữ liệu. Mô hình chuỗi thời gian cho dữ liệu quan sát x_t là mô hình phân phối của một chuỗi biến ngẫu nhiên X_t trong đó x_t là dữ liệu thực. Trong phân tích dữ liệu chuỗi thời gian, một mô hình tốt được đưa ra khi phân tích trên các dữ liệu dừng. Theo Gujarati (2003) một chuỗi thời gian là dừng khi giá trị trung bình, phương sai, hiệp phương sai (tại các độ trễ khác nhau) giữ nguyên không đổi cho dù chuỗi được xác định vào thời điểm nào đi nữa. Chuỗi dừng có xu hướng trở về giá trị trung bình và những dao động quanh giá trị trung bình sẽ là như nhau. Nói cách khác, một chuỗi thời gian không dừng sẽ có giá trị trung bình thay đổi theo thời gian, hoặc giá trị phương sai thay đổi theo thời gian hoặc cả hai. Nói một cách dễ hiểu, chuỗi thời gian $X_t, t = 0, , ...$ được gọi là dừng nếu nó có cùng thuộc tính phân phối với chuỗi thời gian $X_{t+h}, t = 0, 1, ...$ với mỗi số nguyên h.

Định nghĩa 1.2. Giả sử ta có $\{X_t\}$ là một chuỗi thời gian với $E(X_t^2) < \infty$. Ta định nghĩa

kỳ vọng của $\{X_t\}$ như sau:

$$\mu_X(t) = E(X_t).$$

Hiệp phương sai của $\{X_t\}$:

$$\gamma_X(r,s) = Cov(X_r, X_s) = E[(X_r - \mu_X(r))(X_s - \mu_X(s))],$$

với mọi $r, s \in \mathbb{Z}$.

Định nghĩa 1.3. Giả sử rằng $\{X_t\}$ là một chuỗi thời gian, $\{X_t\}$ là một chuỗi dừng yếu nếu:
(i) $\mu_X(t)$ không phụ thuộc vào t.
(ii) $\gamma_X(t+h,t)$ không phụ thuộc vào t chỉ phụ thuộc vào h.

- Chuỗi dừng chặt $\{X_t\}$ được xác định bởi điều kiện $(X_1, ..., X_n$ và $(X_{1+h}, ..., X_{n+h}$ có cùng phân phối với mọi h nguyên và $n > 0$. Ta có thể dễ dàng kiểm tra nếu $\{X_t\}$ là chuỗi dừng chặt và $E(X_t^2) < \infty$ với mọi t.
- Trong định ngĩa trên hàm $\gamma_X(t+h,t)$ được gọi là hàm tự hiệp phương sai với độ trễ h.

Định nghĩa 1.4. Cho chuỗi thời gian dừng X_t.
Hàm tự hiệp phương sai (ACVF) của $\{X_t\}$ với độ trễ h là $\gamma_X(h) = Cov(X_{t+h}, X_t)$.
Hàm tự tương quan của $\{X_t\}$ với độ trề h là $\rho_X(h) = \dfrac{\gamma_X(h)}{\gamma_X(0)} = Cor(X_{t+h}, X_t)$.

Định nghĩa 1.5. Nếu $\{X_t\}$ là chuỗi ngẫu nhiên không tương quan, với mỗi kỳ vòng bằng không và phương sai δ^2 được gọi là nhiễu trắng (White noise) và kí hiệu là $\{X_t\} \sim WN(0, \delta^2)$.

1.1.3 AR, MA

Quá trình $MA(q)$

Định nghĩa 1.6. Quá trình $MA(q)$ được định nghĩa nếu chuỗi $\{X_t\}$ được biểu diễn dưới dạng:

$$X_t = Z_t + \theta_1 Z_{t-1} + ... + \theta_q Z_{t-q},$$

trong đó, $\{Z_t\} \sim WN(0, \sigma^2)$ và $\theta_1, \theta_2, ..., \theta_q$ là các hằng số.

Tính chất:

- $MA(q)$ là một quá trình dừng có $E(X_t) - 0$.
- Hàm ACVF: $\gamma(h) = \begin{cases} \sigma^2 \displaystyle\sum_{j=0}^{q-|h|} \theta_j \theta_{j+h}, & |h| \le q, \\ 0, & |h| > q. \end{cases}$ với $\theta_0 = 1$.
- $MA(q)$ là một quá trình tuyến tính.

Quá trình tuyến tính

Định nghĩa 1.7. Quá trình tuyến tính được định nghĩa nếu chuỗi thời gian $\{X_t\}$ được biểu diễn dưới dạng:

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}, \qquad \forall t \in \mathbb{Z}, \tag{1.1}$$

với $\{Z_t\} \sim WN(0, \sigma^2)$ và $\{\psi_j\}$ là một chuỗi các hằng số sao cho $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$.

Tính chất:

- Quá trình tuyết tính là $MA(q)$ nếu: $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$.
- Ta có: $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty \Rightarrow E(Z_t) \leq \sigma \rightarrow E(X_t) < \infty$.
- $\psi(B)$ là một bộ lọc tuyến tính đưa ra một chuỗi dừng,i.e.,

$$\psi(B) = \sum_{j=-\infty}^{\infty} \psi_j B^j \rightarrow X_t = \psi(B) Z_t.$$

Quá trình $AR(p)$

Định nghĩa 1.8. Quá trinh $AR(p)$ được định nghĩa nếu $\{X_t\}$ là một chuỗi dừng được biểu diễn dưới dạng:

$$X_t - \phi_1 X_{t-1} - ... - \phi_p X_{t-p} = Z_t, \tag{1.2}$$

với $\{Z_t\} \sim WN(0, \sigma^2)$.

Tính chất:

$\{X_t\}$ có PACF: $\alpha(h) = \begin{cases} \phi(p) & h = p \\ 0, & h > p \\ \phi_{hh} & 1 \leq h < p. \end{cases}$ với ϕ_{hh} là phần tử cuối cùng của $\phi_h = \Gamma_h^{-1}\gamma_h$. Trong đó, $\Gamma_h = [\gamma(i-j)]_{i,j=1}^h$ và $\gamma_h = [\gamma(1), \gamma(2), ..., \gamma(h)]'$.

Quá trình $AR(1)$**:**

$$X_t = \phi X_{t-1} + Z_t, \quad t = 0, \pm 1, \pm 2, ...$$

- $|\phi| < 1$.
- $E(X_t) = 0$.
- $\gamma_X(h) = \gamma_X(-h)$.
- $\rho(h) = \dfrac{\gamma_X(h)}{\gamma_X(0)} = \phi^h$.

- $\gamma_X(0) = \dfrac{\sigma^2}{(1-\phi^2)}$.

1.1.4 Mô hình ARMA

Mô hình ARMA

Định nghĩa 1.9. $\{X_t\}$ là chuỗi ARMA(p,q) nếu $\{X_t\}$ là chuỗi dừng và với mọi t :

$$X_t - \phi_1 X_{t-1} - ... - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + ... + \theta_q Z_{t-q} \tag{1.3}$$

ở đó: $\{Z_t\} \approx WN(0, \sigma^2)$ và đa thức $1 - \phi_1 z - ... - \phi_p z^p$ và $(1 + \theta_1 z + ... + \theta_q z^q)$ không có nhân tố chung.

Định nghĩa 1.10. Mô hình ARMA(q,p) được gọi là nhân quả, nếu chuỗi $\{X_t\}$ có thể biểu diễn dưới dạng tuyến tính một phía:

$$X_t = \sum_{j=0}^{\infty} \Psi_j w_{t-j} = \Psi(B) w_t \tag{1.4}$$

trong đó: $\Psi(B) = \sum_{j=0}^{\infty} \Psi_j B^j$, và $|\sum_{j=0}^{\infty} \Psi_j| < \infty$, $\Psi_0 = 1$

Định nghĩa 1.11. Định lí tồn tại và duy nhất
Phương trình $\{X_t\}$ 1.3 có nghiệm (và cũng là nghiệm dừng duy nhất) khi và chỉ khi:

$$\phi(z) = 1 - \phi_1 z - ... - \phi_p z^p \neq 0 \quad \text{với mọi } \ |z| = 1 \tag{1.5}$$

Định nghĩa 1.12. Quá trình ARMA(p,q) của chuỗi $\{X_t\}$ là quan hệ nhân quả hoặc hàm nhân quả của $\{Z_t\}$, nếu tồn tại hằng số $\{\Psi_i\}$ sao cho $|\sum_{j=0}^{\infty} \Psi_j| < \infty$ và

$$X_t = \sum_{j=0}^{\infty} \Psi_j Z_{t-j} \quad \text{với mọi } \ t \tag{1.6}$$

Nói cách khác, nó tương đương với điều kiện:

$$\phi(z) = 1 - \phi_1 z - ... - \phi_p z^p \quad \text{với mọi } \ |z| \leq 1 \tag{1.7}$$

Chuỗi ngẫu nhiên Ψ_j được xác định bởi quan hệ $\Psi_j = \sum_{j=0}^{\infty} \Psi_j z^j = \frac{\theta(z)}{\phi(z)}$, nghĩa là ta có:

$$(1 - \phi_1 z - ... - \phi_p z^p)(\psi_0 + \psi_1 z + ...) = 1 + \theta_1 z + ... + \theta_q z^q$$

Từ đó, tính toán các hệ số cho $z^j, j = 0, 1, ...$, ta tìm được:

$$\begin{aligned} 1 &= \psi_0 \\ \theta_1 &= \psi_1 - \psi_0\phi_1 \\ \theta_2 &= \psi_2 - \psi_1\phi_1 - \psi_0\phi_2 \\ &.... \end{aligned}$$

Tổng quát:

$$\psi_j - \sum_{k=1}^{p} \phi_k \psi_{j-k} = \theta j, \quad j = 0, 1, ...$$

Định nghĩa 1.13. Quá trình ARMA(p,q) của chuỗi $\{X_t\}$ là quá trình đảo ngược (invertible) nếu nó tồn tại hẳng số $\{\pi_j\}$ sao cho $\sum_{j=0}^{\infty} |\pi_j| < \infty$ và

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j} \text{với mọi } t \tag{1.8}$$

Nói cách khác, nó tương đương với điều kiện:

$$\theta(z) = 1 + \theta_1 z + ... + \theta_q z^q \neq 0 \text{với mọi } |z| \leq 1 \tag{1.9}$$

1.2 Wavelets

1.2.1 Giới thiệu về wavelets

Chuỗi Walevet của một hàm tích phân bình phương f được biểu diễn tương tự như biểu diễn chuỗi Fourier. Biểu diễn chuỗi Fourier của một hàm lấy làm cơ sở trực giao nhằm mở rộng nó thành hàm điều hòa cơ bản e^{inx}, trong khi biểu diễn chuỗi Wavelet các bộ cơ sở trực chuẩn khác nhau dựa trên $\Psi_{jk}(x)$ thỏa mãn một số điều kiện được sử dụng.

Nếu f là hàm tích phân bình phương xác định trên đoạn $[0, \pi]$ thì biểu diễn chuỗi Fourier của f là:

$$f(x) = \sum_{n=-\infty}^{\infty} c_n e^{inx} \tag{1.10}$$

trong đó hằng số c_n được xác định bằng

$$c_n = \frac{1}{2} \int_0^{2\pi} f(x) e^{inx} \mathrm{d}x. \tag{1.11}$$

Đặc điểm của biểu diễn chuỗi Fourier là một cơ sở trực chuẩn w_n trong đó f là hàm mở rộng, được tạo ra bởi sự biến đổi tương tự của hàm đơn $w(x) = e^{ix}$ mà $w_n(x) = w(nx$.

Đối với biến đổi chuỗi Wavelet của một hàm, chúng ta mở rộng hàm đó theo cơ sở trực chuẩn Ψ_n khác với cơ sở lượng giác cơ bản. Mỗi một trong số các cơ sở này được tạo ra bởi phép tịnh tiến và biến đổi tương tự của chuỗi Wavelet đơn ban đầu như cơ sở lượng giác cơ bản được tạo ra từ biến đổi tương tự của hàm e^{ix}.

Chuỗi Wavelet con Ψ_{jk} thu được từ Ψ bởi phép biến đổi tương tự như sau

$$\Psi_{jk}(x) = 2^{\frac{j}{2}}\Psi(2^j x - k). \tag{1.12}$$

Hệ số $2^{\frac{j}{2}}$ là hệ số mở rộng. Dưới điều kiện phù hợp của Ψ, tập hợp Ψ_{jk} là cơ sở trực chuẩn của không gian hàm tích phân bình phương. Khi đó, biểu diễn chuỗi Wavelet của hàm f là

$$f(x) = \sum_{j,k} f_{jk}\Psi_{jk}(x), \tag{1.13}$$

trong đó f_{jk} là hệ số Wavelet của f với kỳ vọng của cơ sở Ψ_{jk} mà được thể hiện bởi:

$$f_{jk} = \int_{-\infty}^{\infty} f(x)\Psi_{jk}(x)\mathrm{d}x. \tag{1.14}$$

Ví dụ cơ bản của cơ sở Wavelet là cơ sở Haar được tạo ra bởi Ψ như sau:

$$\Psi(x) = \begin{cases} 1 & 0 \le x < 1/2 \\ -1 & 1/2 \le x < 1 \\ 0 & x \in (-\infty, 0) \cup (1, \infty) \end{cases}$$

Hàm Wavelet có một vài lợi ích hơn thay vì dùng hàm điều hòa cơ bản e^{inx} trong việc biểu diễn một hàm số. Các Wavelet ở cấp j trong Ψ_{jk} phát hiện các thành phần của một tập số cụ thể trong hàm được phân tích.

1.2.2 Phép biến đổi wavelets rời rạc

Từ việc phân tích các điểm dữ liệu, Wavelet cung cấp biểu diễn các hàm tạo bởi tập dữ liệu. Cho véc tơ $y = (y_0, ..., y_{2^n-1})$ có kích thước 2^n, y được xét trong $[0, 1)$ xác định bởi:

$$f(x) = y_k \quad x \in [k/2^n, (k+1)/2^n]. \tag{1.15}$$

Hàm f là tích phân bình phương và có phân rã Wavelet dạng:

$$f(x) = c_{00}\Phi(x) + \sum_j \sum_k d_{jk}\Psi_{jk}(x), \tag{1.16}$$

trong đó Φ là hàm mở rộng.
Phép biến đổi Fourier rời rạc của chuỗi $a = (a_0, ..., a_{N-1})$ là chuỗi $b = (b_0, ..., b_{N-1})$ như sau:

$$b_j = \sum_{t=0}^{N-1} a_t e^{-i(2\pi j/N)t} \quad j = 0, ..., N-1. \tag{1.17}$$

Chuỗi b là biến đổi tuyến tính của a cho phép tìm kiếm tập dữ liệu trong miền tần số thay vì miền thời gian. b là tập các hệ số Fourier của phép biến đổi Fourier rời rạc a. Tương tự, bằng phép biến đổi Wavelet rời rạc với hàm Wavelet Ψ_{jk}, sự biến đổi wavelet rời rạc của một véc tơ dữ liệu x có kích thước $N = 2^n$ là một véc tơ dữ liệu khác có cùng kích thước với x cho phép chúng ta xem xét tập dữ liệu x trong miền địa phương và tần số cục bộ thay vì trong miền thời gian. Véc tơ d là tập các hệ số của biến đổi chuỗi wavelet x với kỳ vọng của cơ sở Wavelet đã chọn. Biến đổi này là tuyến tính và trực giao, và được mô tả bởi ma trận trực giao $WNxN$.

Các phần tử của d là các hệ số Wavelet của x với kỳ vọng từ cơ sở Ψ_{jk}, d có thể dược viết:

$$d = (c_{00}, d_{00}, d_{10}, d_{20}, ..., d_{n-1,2^{n-1}-1}), \tag{1.18}$$

trong đó c_{00} là hệ số của hàm mở rộng Φ. d_{00} là hệ số bậc 0, hai hệ số bậc 1 là d_{10} và d_{11} và tổng quát có 2^j hệ số bậc j là $d_{j0}, d_{j1}, ..., d_{j,2^j-1}$. Hệ số cuối cùng là hệ số bậc $(n-1)$.

Cuối cùng, ta định nghĩa biểu đồ tỷ lệ của d là bản sao của biểu đồ chu kì trong phân tích Fourier. Nếu d là véc tơ hệ số của biến đổi Wavelet rời rạc của x ($d = Wx$), giá trị của d ở cấp j là:

$$E(j) = \sum_{k=0}^{2^j-1} d_{jk}^2 \quad j = 0, ..., n-1. \tag{1.19}$$

Biểu đồ tỷ lệ của d là một véc tơ:

$$(c_{00}^2, E(0), E(1), ...E(n-1)).$$

Biểu đồ tỷ lệ của phép biến đổi Wavelet rời rạc là một chuỗi thời gian được sử dụng để phân rã chuỗi thành chu kì của các tần số khác nhau.

1.2.3 Phân rã chuỗi thời gian thành chu kì

Ta có $x = (x_t)$ là một tập dữ liệu có kích thước là 2^n. Mục tiêu của thuật toán là phân tích x thành hai tập dữ liệu $y = (y_t)$ và $z = (z_t)$ [1] có cùng số chiều với x:

$$x = y + z,$$

trong đó mỗi phần tử của tập dữ liệu y, z phản ánh sự dao động của x ở các tần số khác nhau. Với phân tích wavelet của x có bậc j bé thì hầu hết hệ số d_{jk} lớn, tức là với một khoảng thời gian dài thì phần tử chu kỳ tuần hoàn của x sẽ thấp. Khoảng thời gian xét không nhất thiết là một hằng số, bởi vì một số phần tử d_{jk} của các bậc j này có thể nhỏ. Mặc khác, với bậc j cao, hầu hết các hệ số d_{jk} lớn trong một khoảng thời gian ngắn, chu kỳ tuần hoàn của x lớn. Theo những phân tích chuỗi thời gian trong kinh tế, bài toán thông thường sẽ tìm thành phần mùa trong khoảng thời gian là 12 tháng và với xu hướng dài hạn. Do đó, việc tìm hai đỉnh trong biểu đồ vô hướng các hệ số wavelet d của một tập dữ liệu kinh tế x là hợp lí. Phân tách d thành $d^{(1)}$ và $d^{(2)}$ dùng cho các phần phân tách biển đổi wavelet ngược W^{-1} từ đó hai thành phần của $x = (x_t)$ được xác định. Cụ thể hơn, giả sử rằng trong các biểu đồ vô hướng của d. hai đỉnh tại mức $j_1 < j_2$. Chúng ta chia tập $\{0, 1, 2, ..., n-1\}$ thành hai tập con $A = \{0, 1, ..., j\}$ và $B = \{j+1, ..., n\}$ sao cho các mức trong tập A nằm trong khoảng xung quanh j_1, và mức của tập B nằm trong khoảng xung quanh j_2.
Khi đó, $d^{(1)}$ và $d^{(2)}$ được xác định như sau:

$$d^{(1)} = (c_{00}, d_{00}, d_{10}, d_{11}, ..., d_{j0}, ..., d_{j,2^j-1}, 0, ..., 0)$$
$$d^{(2)} = (0, ..., d_{j+1,1}, ..., d_{j+1,2^{j+1}-1}, ..., d_{n-1,2^{n-1}-1}),$$

và x sẽ được phân tách thành:

$$y = W^{-1}d^{(1)}$$
$$z = W^{-1}d^{(2)}.$$

Thông thường, y đại diện cho xu hướng dài hạn của tập dữ liệu kinh tế. Các thành phần này bằng cách nào đó liên quan đến xu hướng dài hạn của các dữ liệu kinh tế liên quan, tất cả được định nghĩa là chu kỳ kinh doanh của nền kinh tế. z thường đại diện cho thành phần mùa 12 tháng của chuỗi thời gian kinh tế. Điều đó dễ dàng hơn cho việc dự báo y và z một cách riêng biệt hơn là dự báo cả chuỗi x, Bằng cách xây dựng này, nó có thành phần chu kỳ cụ thể: z phải là chu kỳ 12 tháng và y trơn. Tất cả phản ánh xu hướng dài hạn của chuỗi x.

Chương 2

MÔ HÌNH VÀ PHƯƠNG PHÁP TIẾP CẬN

2.1 Mô hình ARIMA

Như chúng ta đã biết, mô hình ARMA là mô hình quan trọng cho các chuỗi thời gian dừng. Thực tế, không phải chuỗi nào cũng dừng. Tổng quát hơn của vấn đề này, ta có mô hình ARIMA. Hay còn được biết đến như là hướng tiếp cận Box-Jenkins.

Định nghĩa 2.1. Nếu d là một số nguyên không âm, khi đó $\{X_t\}$ là một chuỗi ARIMA(p,d,p) nếu $Y_t := (1-B)^d X_t$ là chuỗi nhân quả của chuỗi ARMA(p,q).

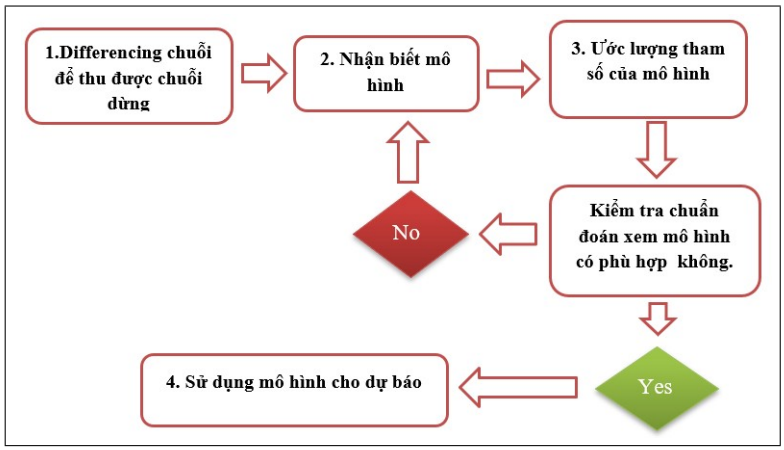
Nghĩa là $\{X_t\}$ thỏa mãn một phương trình sai phân có dạng:

$$\phi^* X_t = \phi(B)(1-B)^d X_t = \theta(B) Z_t, \qquad Z_t \approx WN(0, \sigma^2) \tag{2.1}$$

trong đó: $\phi(z)$ và $\theta(z)$ tương ứng là các đa thức bậc p và q, $\phi(z)$ và $|z| \leq 1$. Đa thức $\phi^*(z)$ có bậc d bằng 0 tại $z = 1$. Chuỗi $\{X_t\}$ là chuỗi dừng khi và chỉ khi d = 0, và khi đó nó trở thành chuỗi ARMA(p,q).

Khi $d \geq 1$ chúng ta có thể thêm một xu hướng đa thức tùy ý bậc $(d-1)$ tùy ý vào $\{X_t\}$ mà không vi phạm phương trình sai phân 2.1. Mô hình ARIMA hữu ích khi biểu diễn dữ liệu có trend. Tuy nhiên, nó cũng có thể thích hợp với mô hình dũ liệu không có trend. Ngoại trừ khi $d = 0$, thì giá trị trung bình của $\{X_t\}$ không được xác định bởi phương trình 2.1 mà nó có thể bằng 0. Với $d \geq 1$, phương trình xác định các thuộc tính bậc 2 của $\{(1-B)^d X_t\}$ nhưng không xác định các thuộc tính của $\{X_t\}$. Ước lượng các tham số ϕ, θ, σ^2 sẽ dựa trên các quan sát khác nhau của $\{(1-B)^d X_t\}$.

Lược đồ phương pháp ARIMA



Hình 2.1: Lược đồ phương pháp ARIMA

- Bước 1: Xác định chuỗi dừng
  Để có được mô hình chuỗi thời gian theo hướng tiếp cận Box-Jenkins thì chuỗi phải là chuỗi dừng.
  Một số phương pháp đưa chuỗi thời gian thành chuỗi dừng. Thông thường người ta sử dụng toán tử sai phân (differencing):chuyển đổi chuỗi thành một chuỗi thời gian mới trong đó các giá trị là sự khác biệt giữa các giá trị liên tiếp.
  Một số bậc sai phân thường dùng:
  (Sai phân bậc 1) $\delta Xt = X_t - X_{t-1}$
  (Sai phân bậc 2) $\delta^2 X_t = (\delta X_t - \delta X_{t-1}) = X_t - 2X_{t-1} + X_{t-2}$
  Ngoài ra còn một số cách tiếp cận khác nhằm áp dụng trên các mô hình chuỗi thời gian khác nhau. Việc lựa chọn phương pháp sao cho phù hợp, ta dựa trên một số đặc điểm như sau:

  - Nếu chuỗi thời gian là chuỗi ngẫu nhiên thì sử dụng toán tử sai phân.
  - Nếu là chuỗi xác định (có xu hướng xác định hoặc chu kì mùa): sử dụng hồi quy.
  - Kiểm tra nếu phương sai thay đổi theo thời gian thì làm cho phương sai không đổi với log hoặc căn bậc hai.
  - Loại bỏ xu hướng với sai phân bậc 1 (bậc 2), làm trơn hóa hoặc sai phân để loại bỏ thành phần mùa.
  - Nếu dữ liệu là theo mùa: loại bỏ thành phần mùa bằng phương pháp trung bình trượt, sai phân hoặc làm trơn hóa.

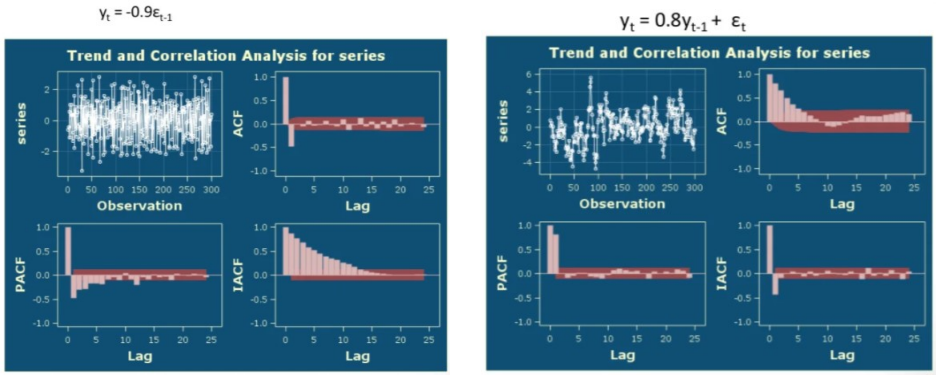
- Bước 2: Lựa chọn các tham số d, p, q phù hợp.
  Các tham số p,d,q có thể được tìm thấy bằng cách sử dụng đồ thị ACF và PACF xác định thông qua tính chất của nó đối với mô hình (bảng 2.1). Nếu cả ACF và PACF đều giảm dần, điều đó chỉ ra rằng chúng ta cần làm cho chuỗi thời gian dừng với giá trị d hợp lí.

Quá trình	MA(q)	AR(p)	ARMA(p,q)
ACF	cắt bỏ	bằng 0 nếu h>p	cắt bỏ
PACF	bằng 0 nếu h>p	cắt bỏ	cắt bỏ

Bảng 2.1: Tính chất của hàm ACF và PACF

Từ đó ta thấy: hàm ACF thường dùng để xác định bậc của quá trình MA, hàm PACF dùng để xác định bậc của quá trình AR.
Các ACF và PACF của các quá tình AR(p) và MA(q) có các dạng trái ngược; trong trường hợp AR(p), ACF giảm theo cấp số nhân hay theo số mũ nhưng PACF đạt tới giới hạn qua một độ trễ nhất định, trái lại hiện tượng này đối nghịch đối với quá trình MA(q).

MA(1) AR(1)



Một ví dụ biểu diễn ACF, PACF của MA(1) và AR(1)

- Ước lượng tham số, kiểm tra chuẩn đoán.
  Sau khi đã xác định được mô hình ARIMA, bước kế tiếp là ước lượng các thông số của các số hạng tự hồi quy và trung bình trượt trong mô hình thông qua phương pháp bình phương sai số nhỏ nhất. Bằng cách tối thiểu hóa hàm tổng bình phương sai lệch:

$$\textbf{min} \qquad \sum_t \epsilon_t^2 \tag{2.2}$$

$$\textbf{min} \qquad \sum_{t=2}^{T} (y_t - \phi y_{t-1})^2 \tag{2.3}$$

Hiện nay, công việc này có thể sủ dụng tự động bằng một số phần mềm thống kê.
Sau khi đã lựa chọn mô hình ARIMA cụ thể và ước lượng các tham số của nó, ta tìm hiểu xem mô hình lựa chọn có phù hợp với dữ liệu ở mức chấp nhận hay không. Một chuẩn đoán đơn giản từ cấc phần dư là tính ACF và PACF trên các phần dư cho tới một độ trễ nhất định, sau đó kiểm tra tính tự tương quan hay tự tương quan riêng từng phần trên các phần dư là thuần nhất ngẫu nhiên hay tương quan. Tù đó, quyết định xem mô hình là phù hợp hay chưa phù hợp.

- Bước 4: Dự báo
  Một trong số các lý do về tính phổ biến của phương pháp lập mô hình ARIMA là thành công của nó trong dự báo. Trong nhiều trường hợp, các dự báo thu được từ phương pháp này tin cậy hơn so với các phương pháp kinh tế lượng truyền thống, đặc biệt là trong dự báo ngắn hạn.
  Dự báo bước - One-step ahead: Mỗi mô hình đều hỗ trợ dự báo 1 bước. Dự báo trước một bước là cần thiết để tính toán lỗi mô hình trong quá trình ước lượng. Các dựa báo trước 1 bước được tính toán tuần tự cho từng điểm dữ liệu bằng cách sử dụng các trạng thái xu hướng và mức tính toán cho điểm hiện tại và các trạng thái theo mùa trước. Sai số dự báo được tính bằng cách trừ đi giá trị dự báo tại điểm trúoc đó cho giá trị quan sát tại điểm hiện tại.
  Dự báo h bước - h-step ahead: dự báo k bước được sử dụng để đưa ra dự đoán cho bất kì số lượng giá trị tương lai nào theo sau dữ liệu chuỗi thời gian được quan sát. Thông thường, dự báo càng gần giá trị quan sát hiện tại thì càng cho dự báo chính xác.
  Khoảng tin cậy - Confidence Intervals: Khoảng tin cậy cung cấp mức độ không chắc chắn của mỗi giá trị dự báo. Các giá trị này thường trở nên rộng hơn trong tương lai, bởi vì dự báo xa hơn sẽ kém tin cậy hơn. Giới hạn độ tin cậy cung cấp thông tin chi tiết có liên quan về hành vi trong tương lai của chuỗi thời gian được quan sát. Tính toán giới hạn tin cậy dựa trên phương sai tổng thể của sai số dự báo được ước tính dựa trên dữu liệu quan sát và yếu tố phụ thuộc vào mô hình chỉ định và số bước từ điểm quan sát cuối cùng.

2.2 Mô hình Wavelets ARIMA

Các bước trong mô hình lai Wavelets ARIMA[2]:

- Phát hiện các hàm wavelet đơn giản và bậc của thành phần phân rã, sau đó tiến hành phân rã wavelet cho chuỗi thời gian gốc để có được thành phần hệ số tần cao và hệ số tần thấp.
- Xác định các ngưỡng của tần số cao hệ số wavelet và tần số thấp với tiêu chí nghiêm

ngặt và tiêu chí tương ứng, sử dụng hàm ngưỡng mềm để thực hiện xử lí ngưỡng trên các hệ số wavelet.

- Wavelets ngược để tạo ra được chuỗi thời gian khử nhiễu.
- Thực hiện mô hình dự báo ARIMA trên chuỗi khử nhiễu để thu được kết quả.

2.3 Độ đo đánh giá mô hình

Để đánh giá mức độ phù hợp của mô hình với dữ liệu đầu vào thì ta có đưa ra một hàm hợp lý để đánh giá chúng. Trong thống kê , hàm Likelihood (thường được gọi đơn giản là hợp lý) đo lường mức độ phù hợp của mô hình thống kê với một mẫu dữ liệu cho các giá trị đã cho của các tham số chưa biết . Nó được hình thành từ phân phối xác suất chung của mẫu, nhưng chỉ được xem và sử dụng như một hàm của các tham số, do đó xử lý các biến ngẫu nhiên là cố định tại các giá trị quan sát được.

Hàm Log-Likelihood mô tả một siêu mặt có đỉnh, nếu nó tồn tại, biểu thị sự kết hợp của các giá trị tham số mô hình nhằm tối đa hóa xác suất vẽ mẫu thu được. Quy trình lấy các đối số này của ước lượng hợp lí cực đại tối đa được gọi là ước lượng khả năng tối đa , để thuận tiện cho việc tính toán thường được thực hiện bằng cách sử dụng logarit tự nhiên của khả năng, được gọi là hàm Log-Likelihood. Công thức để tính Likelihood bằng tay là quá phức tạp, tuỳ từng model mà có cách tính khác nhau.

Giả sử các tham chiếu dựa trên khả năng xảy ra, đối với N đối tượng và một số cụm K cố định là:

$$L(p_1, ..., p_k, \mu_1, \sum_1, ..., \mu_K, \sum_K) = \prod_{j=1}^{N} (f_{\text{Mix}}(x_j|\mu_1, \sum_1, ..., \mu_K, \sum_K))$$
$$= \prod_{j=1}^{N} (\sum_{k=1}^{K} p_k \frac{1}{(2\pi)^{p/2}\left|\sum_k\right|^{1/2}} \exp(-\tfrac{1}{2}(x-\mu_k)' \sum_k^{-1} (x-\mu_k)))$$

trong đó trong đó các xác suất p_k, các vector trung bình μ_k, ma trận hiệp phương sai Σ_k là không biết. Các phép đo cho các đối tượng khác nhau được coi là các quan sát độc lập và phân bố giống hệt nhau từ sự phân bố hỗn hợp.

Thực tế khi đánh giá, lựa chọn một mô hình phù hợp thì thường xác định thông qua 2 chỉ số cơ bản là AIC và BIC. Cụ thể

- AIC - Akaike information criteriom:
  Là một công cụ mạnh mẽ ước tính sai số dự đoán tương đối của các mô hình thống kê cho tập dữ liệu nhất định được xây dựng bởi Hirotugu Akaike năm 1974. Nó đưa ra một tập hợp các mô hình cho dữ liệu, AIC ước tính độ hiệu quả của mỗi mô hình, so sánh với từng mô hình khác.
  Giả sử ta có một mô hình thống kê của một số dữ liệu, k là tham số ước lượng mô hình.

Phép $\hat{L}$ là giá trị lớn nhất của hàm likehood cho mô hình. Khi đó giá trị AIC cho mô hình được tính như sau:

$$AIC = 2k - 2ln(\hat{L}) \tag{2.4}$$

- BIC - Bayesian information criterion:
  Cũng như AIC, BIC cũng là một tiêu chí để lựa chọn mô hình trong một tập hữu hạn các mô hình, nhưng nó phạt các tham số nhiều hơn do độ phức tạp của nó. BIC được phát triển bởi Gideon E. Schwarz, và được định nghĩa:

$$BIC = kln(n) - 2ln(\hat{L}) \tag{2.5}$$

  Một điều lưu ý rằng, tiêu chí BIC xác định theo khung xác suất Bayes nghĩa là việc lựa chọn các mô hình ứng viên bao gồm một mô hình đúng cho tập dữ liệu, thì xác suất BIC chọn mô hình thực sẽ tăng theo kích thước của tập dữ liệu huấn luyện.

AIC/BIC hữu dụng cho sự so sánh sự hiệu quả của 2 mô hình. Giá trị càng nhỏ thì mô hình càng tốt.

Bên cạnh đó, để xác định giá trị dự báo nào là tốt, thường sử dụng một số độ đo đánh giá sau:

- Mean absolute error:

$$\text{MAE} = \sum_{i=1}^{n} \frac{|Y_i - \hat{Y}_i|}{n} \tag{2.6}$$

- Mean absolute percent error:

$$\text{MAPE} = \frac{100}{n} \sum_{i=1}^{n} \frac{|Y_i - \hat{Y}_i|}{Y_i} \tag{2.7}$$

- Mean square error:

$$\text{MSE} = \sum_{i=1}^{n} \frac{(Y_i - \hat{Y}_i)^2}{n} \tag{2.8}$$

- Root mean square error:

$$\text{RMSE} = \sqrt{MSE} \tag{2.9}$$

Chương 3

THỰC NGHIỆM

3.1 Phát biểu bài toán

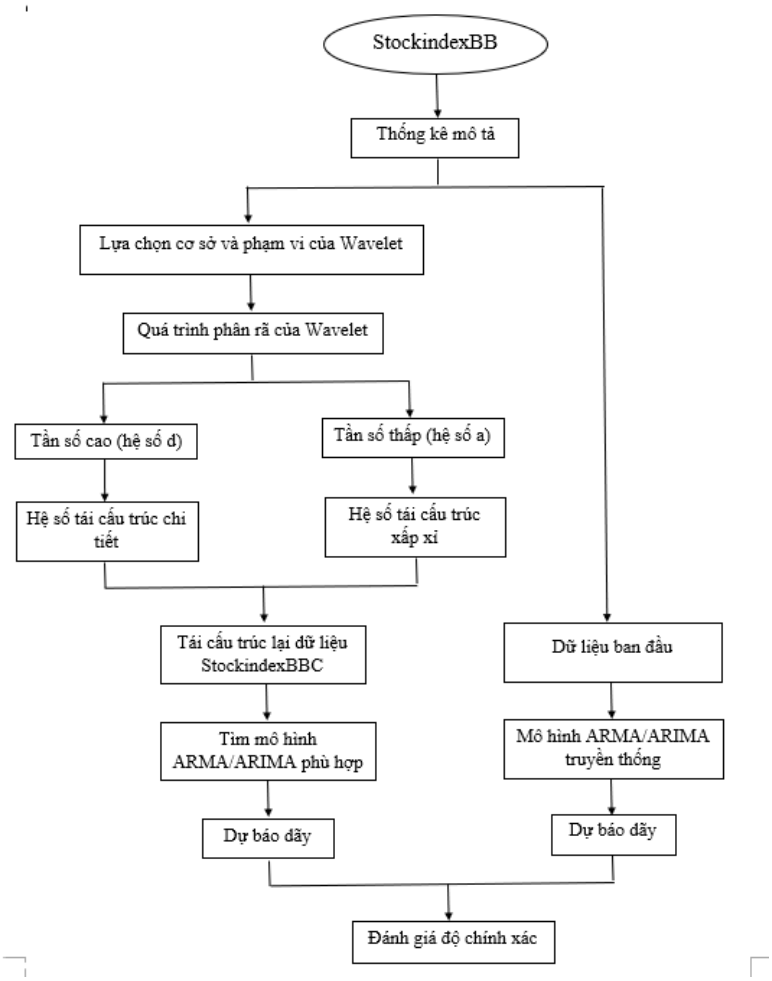
Những năm gần đây xu thế hội nhập phát triển kinh tế ngày càng trở nên quan trọng, nền chứng khoán của thị trường Việt Nam được đánh giá và kỳ vọng với những tiềm năng và tài nguyên dồi dào. Trên nền chứng khoán hiện nay, rất nhiều phương pháp thuật toán được áp dụng để có thể dự báo ngắn hạn chính xác giá chứng khoá những ngày kế tiếp dựa vào những dữ liệu đã có trước, từ đó giúp các nhà đầu tư làm chủ và nắm bắt được cơ hội. Thông quan nghiên cứu tìm hiểu, nhóm chúng em nhận thấy phương pháp áp dụng mô hình ARIMA trong việc phân tích và dự báo chứng khoán đã và đang đưa lại kết quả khá tốt. Trong bài báo cáo này, nhóm chúng em sẽ áp dụng kết hợp mô hình ARIMA và Wavelet với mong muốn cải thiện thuật toán đã có.

3.2 Mô tả dữ liệu

Ở đây, nhóm chúng em lấy dữ liệu chứng khoán từ công ty cổ phần chứng khoán BBC trong khoảng thời gian từ 01/01/2021 - 30/10/2021 để áp dụng thuật toán từ đó phân tích và dự báo kết quả. Sau đó so sánh với thời gian thực để có thấy sự khác nhau giữa việc dùng thuần túy mô hình dự báo ARIMA và việc kết hợp giữa ARIMA và Wavelet.
Bộ dữ liệu gồm hai cột: DATE (Ngày tháng năm) và CLOSE (Giá lúc đóng sàn).

3.3 Phương pháp thực nghiệm

Sơ đồ thuật toán



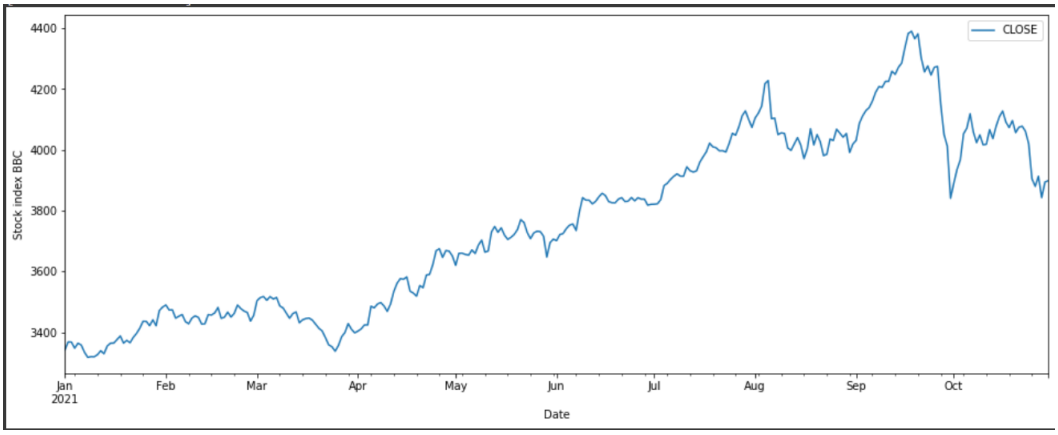
Hình 3.1: Sơ đồ thuật toán

Với mục tiêu là dự báo giá ngắn hạn giá chứng khoán, trong thực nghiệm của mình chúng em sử dụng 2 mô hình:

- Mô hình 1: Áp dụng mô hình ARIMA dự báo trên chuỗi gốc.
- Mô hình 2: Áp dụng Wavelets chuyển đổi làm nhiễu chuỗi thời gian, sau đó dự báo bằng mô hình ARIMA.

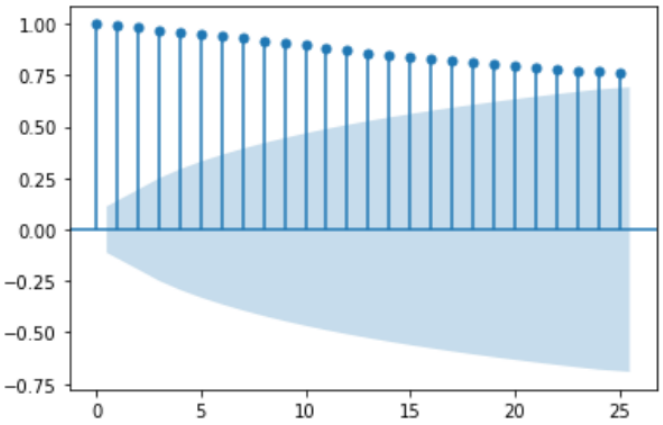
3.3.1 Mô hình 1: Sử dụng thuần túy mô hình dự báo ARIMA

- **Mô hình dữ liệu ban dầu**



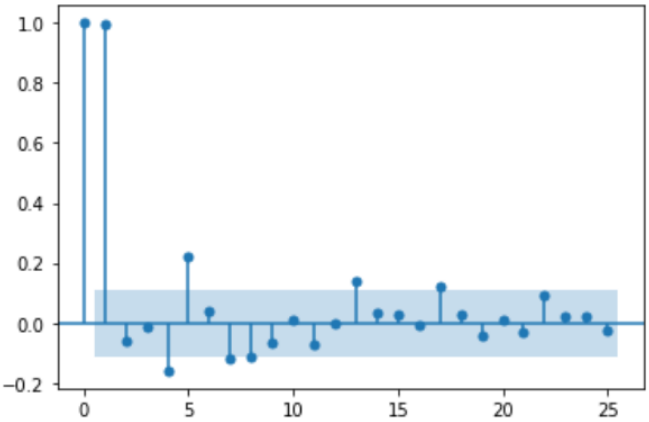
Hình 3.2: Mô hình dữ liệu ban dầu

- **ACF**



Hình 3.3: ACF

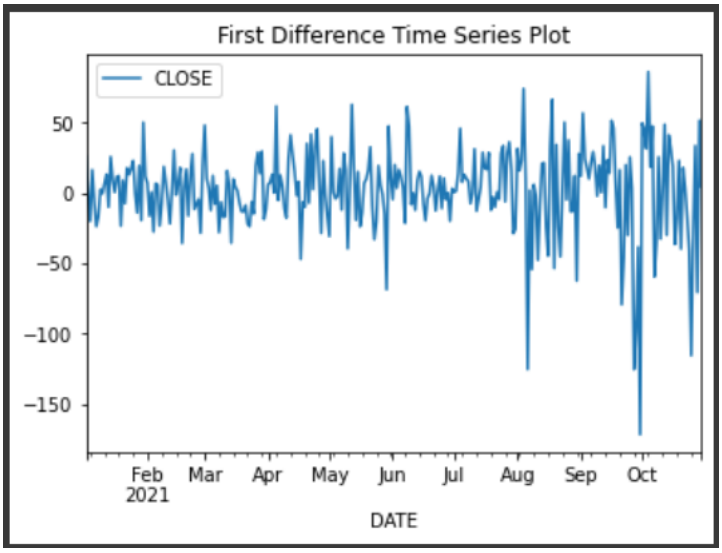
- **PACF**



Hình 3.4: PACF

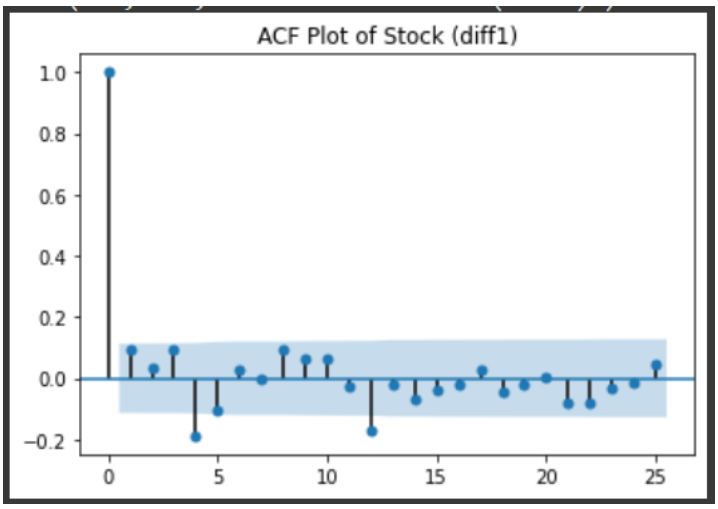
Nhìn vào đồ thị của ACF và PACF đã vẽ ở trên, nhận thấy đây không là một chuỗi dừng, nhóm đã sử dụng phương pháp toán tử sai phân (differencing) bậc 1 để đưa dãy về thành chuỗi dừng.
Dữ liệu sau khi đã xử lý:

- **Mô hình dữ liệu sau sai phân một lần**



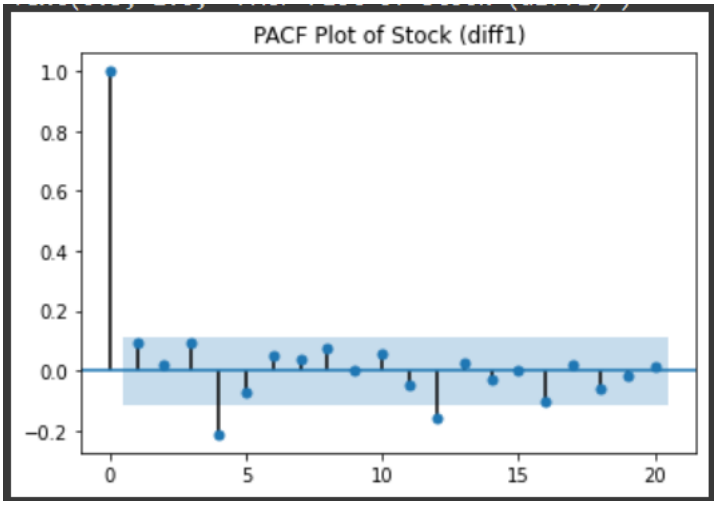
Hình 3.5: Mô hình dữ liệu sau sai phân một lần

- **ACF**



Hình 3.6: ACF sau sai phân 1 lần

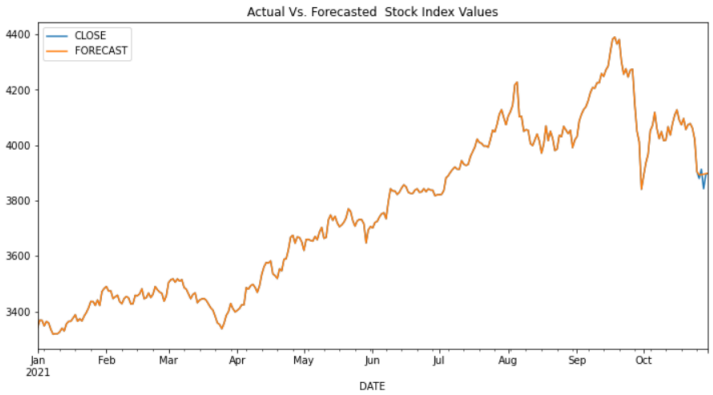
- **PACF**



Hình 3.7: PACF sau sai phân 1 lần

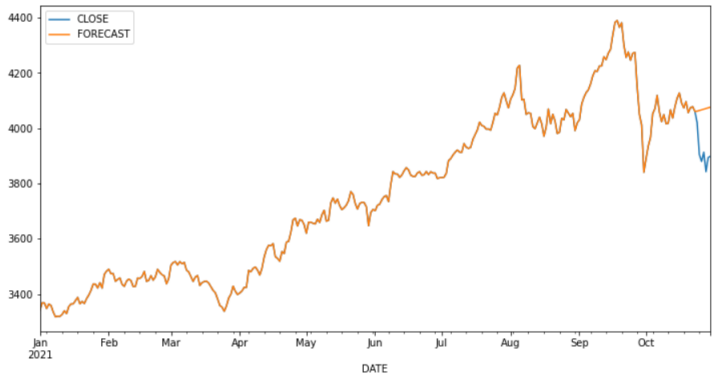
Tiếp đó, ta sẽ tìm một mô hình ARIMA(p,d,q) phù hợp với dữ liệu đã có được. Ở đây chúng em sử dụng thư viện *pmdarima* hỗ trợ việc tìm mô hình ARIMA phù hợp dựa vào chỉ số AIC. Chỉ số AIC càng nhỏ thì mô hình đó càng phù hợp với dữ liệu.
Sau khi chạy chương trình thì mô hình ARIMA(1,1,0) là mô hình phù hợp dùng để dự báo cho dữ liệu hiện có.
Kết quả sau khi đã dự báo:

- Dự báo 5 ngày



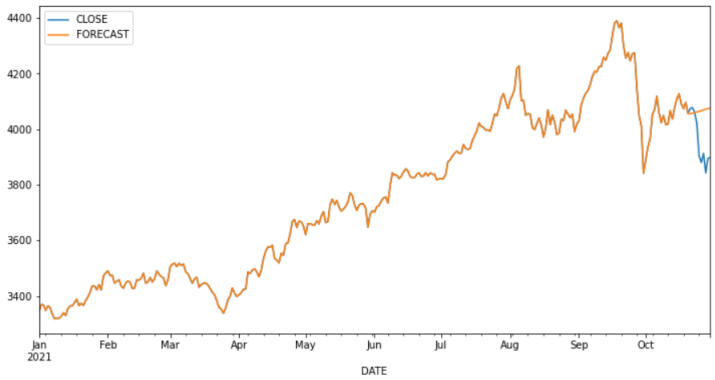
Hình 3.8: Dự báo 5 ngày

- Dự báo 7 ngày



Hình 3.9: Dự báo 7 ngày

- Dự báo 10 ngày

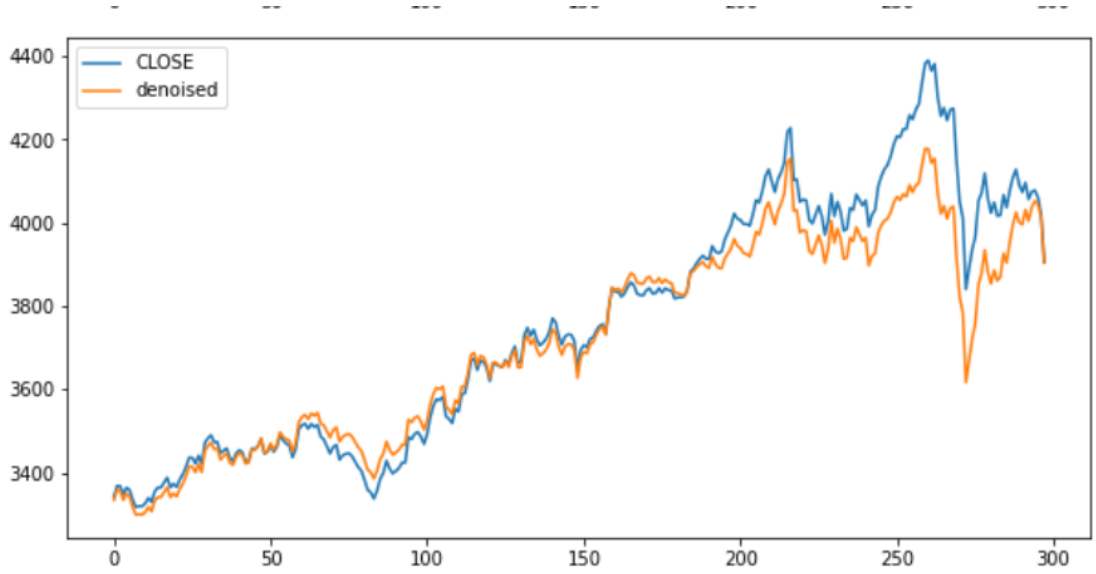


Hình 3.10: Dự báo 10 ngày

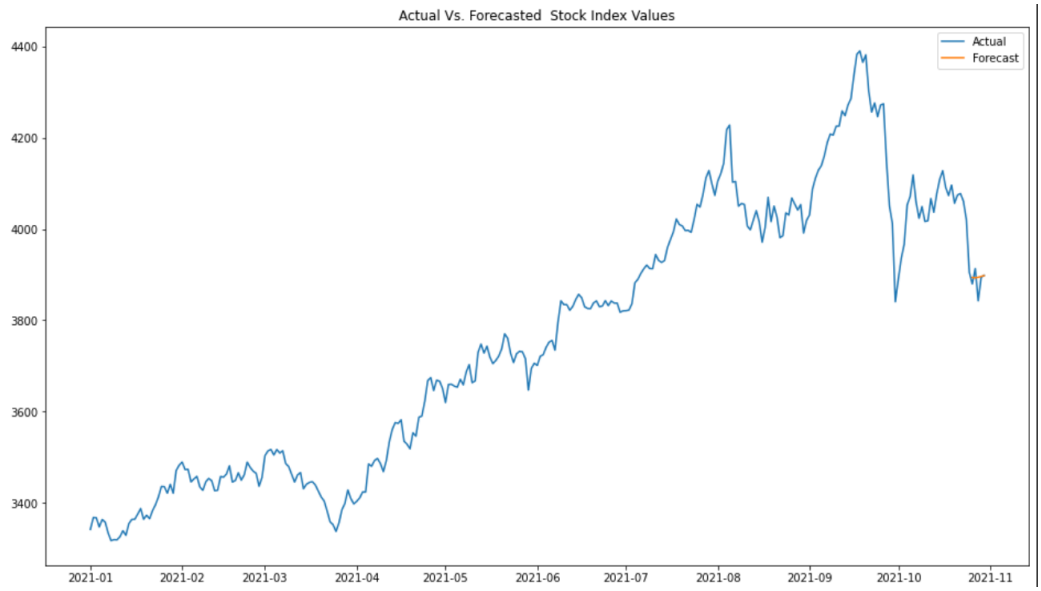
3.3.2 Mô hình 2: Kết hợp ARIMA và Wavelet

Dãy dữ liệu ban đầu nhóm sẽ áp dụng phép biến đổi DWT có bộ lọc là *db4* và cấp độ *level 9* để tách chuỗi gốc thành 1 chuỗi xấp xỉ A_1 và 9 chuỗi thành phần chi tiết $D_8, D_7, D_6, D_5, D_4, D_3, D_2, D_1, D_0$.

Sau đó thực hiện biến đổi tái cấu trúc trên mỗi chuỗi con, thực hiện biến đổi IDWT ngược ta thu được chuỗi tái cấu trúc khử nhiễu (denoised). Tương tự, mô hình ARIMA được áp dụng cho chuỗi thời gian khử nhiễu nhằm đưa ra kết quả dự báo. Kết quả sẽ được so sánh, đánh giá với mô hình ARIMA truyền thống sẽ được trình bày trong bảng3.2 .



Hình 3.11: Tái cấu trúc xấp xỉ với dữ liệu ban đầu



Hình 3.12: Dự báo 5 ngày

3.4 Kết quả thực nghiệm

Bảng kết quả

Date	Dữ liệu thực	ARIMA dự báo	ARIMA, Wavelet dự báo
26/10	3879.893	3879.893	3893.9528
27/10	3913.2689	3913.2689	3893.0981
28/10	3842.7155	3842.7155	3894.3396
29/10	3894.0498	3894.0498	3896.0717
30/10	3898.4977	3898.4977	3897.9151

Bảng 3.1: Bảng so sánh kết quả dự báo 5 ngày

Đánh giá mô hình

	ARIMA	Wavelet with ARIMA
MSE	665.902	654.855
RMSE	25.805	25.590
MAE	17.776	17.693
MAPE	0.460 %	0.469 %

Bảng 3.2: Đánh giá mô hình

Sau khi nghiêm cứu, tìm hiệu và thử nghiệm nhóm chúng em có một số kết luận như sau:

- Dự báo bằng mô hình ARIMA hay kết hợp ARIMA và Wavelet thì số ngày dự báo càng gần sẽ mang lại kết quả tốt hơn.
- Thuật toán Wavelets giúp phân rã dãy dữ liệu ban đầu thành nhiều dãy con đơn giản hơn, từ đó giúp giảm thời gian xử lí và giảm khối lượng tính toán trong quá trình dự báo.
- Kết quả khi áp dụng mô hình kết hợp tốt hơn so với kết quả khi dùng thuần túy mô hình ARIMA.
- Trên thực tế, mô hình lai Wavelets - ARIMA cho kết quả dự báo tốt hơn các mô hình ARIMA truyền thống [3], [2]. Tuy nhiên, trong thực nghiệm của nhóm, kết quả lại tương đồng. Lí giải cho điều này, có thể do dữ liệu lựa chọn chưa tối ưu, hay lựa chọn tham số mô hình chưa hiệu quả.

KẾT LUẬN

Hiện nay có rất nhiều phương pháp và mô hình được áp dụng để dự báo chứng khoán, tài chính, kinh tế,... Nhưng hầu hết trong các yêu cầu dự báo ngắn hạn thì mô hình dự báo ARIMA cho một kết quả đáng tin cậy nhất trong các mô hình dự báo. Tuy nhiên, khi sử dụng mô hình ARIMA để dự báo, chúng ta cần một số lượng lớn dữ liệu quan sát và không thể đưa các yếu tố thay đổi có ảnh hưởng đến biến số cần dự báo của thời kỳ cần dự báo vào mô hình. Như vậy có thể nói xây dựng mô hình ARIMA theo phương pháp luận Box-Jenkins có tính chất nghệ thuật hơn là khoa học, hơn nữa kỹ thuật và khối lượng tính toán khá lớn nên đòi hỏi phải có phần mềm kinh tế lượng chuyên dùng.

Khi gặp một số lượng lớn dữ liệu và phức tạp, khi đó ta có thể áp dụng phương pháp Wavelet để phân rã dãy dữ liệu thành nhiều dãy con đơn giản hơn nhiều, sau đó áp dụng mô hình ARIMA phù hợp lần lượt cho mỗi dãy con, từ đó dự báo và truy xuất ngược lại dự báo cho dãy ban đầu. Phương pháp này giúp chúng ta tiết kiệm thời gian xử lí và làm giảm khối lượng tính toán đáng kể.

Bên cạnh đó phương pháp vẫn còn nhiều chỗ chưa xử lí tốt trong dự báo và không có tính linh hoạt cao như một số mô hình khác. Cần có sự lựa chọn thích hợp các mô hình trong từng trường hợp dự báo để đạt kết quả tốt nhất và đảm bảo tính thời gian thực của dữ liệu.

Tài Liệu Tham Khảo

[1] M. Arino, "Time series forecasts via wavelets: An application to car sales in the spanish market. institute of statistics & decision sciences," *Duke University*, 1995.

[2] Z. Wang and Y. Lou, "Hydrological time series forecast model based on wavelet denoising and arima-lstm," in *2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, IEEE, 2019, pp. 1697–1701.

[3] H. Zhang, S. Zhang, P. Wang, Y. Qin, and H. Wang, "Forecasting of particulate matter time series using wavelet analysis and wavelet-arma/arima model in taiyuan, china," *Journal of the Air & Waste Management Association*, vol. 67, no. 7, pp. 776–788, 2017.

[4] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

[5] A. K. Jain, J. Mao, and K. M. Mohiuddin, "Artificial neural networks: A tutorial," *Computer*, vol. 29, no. 3, pp. 31–44, 1996.