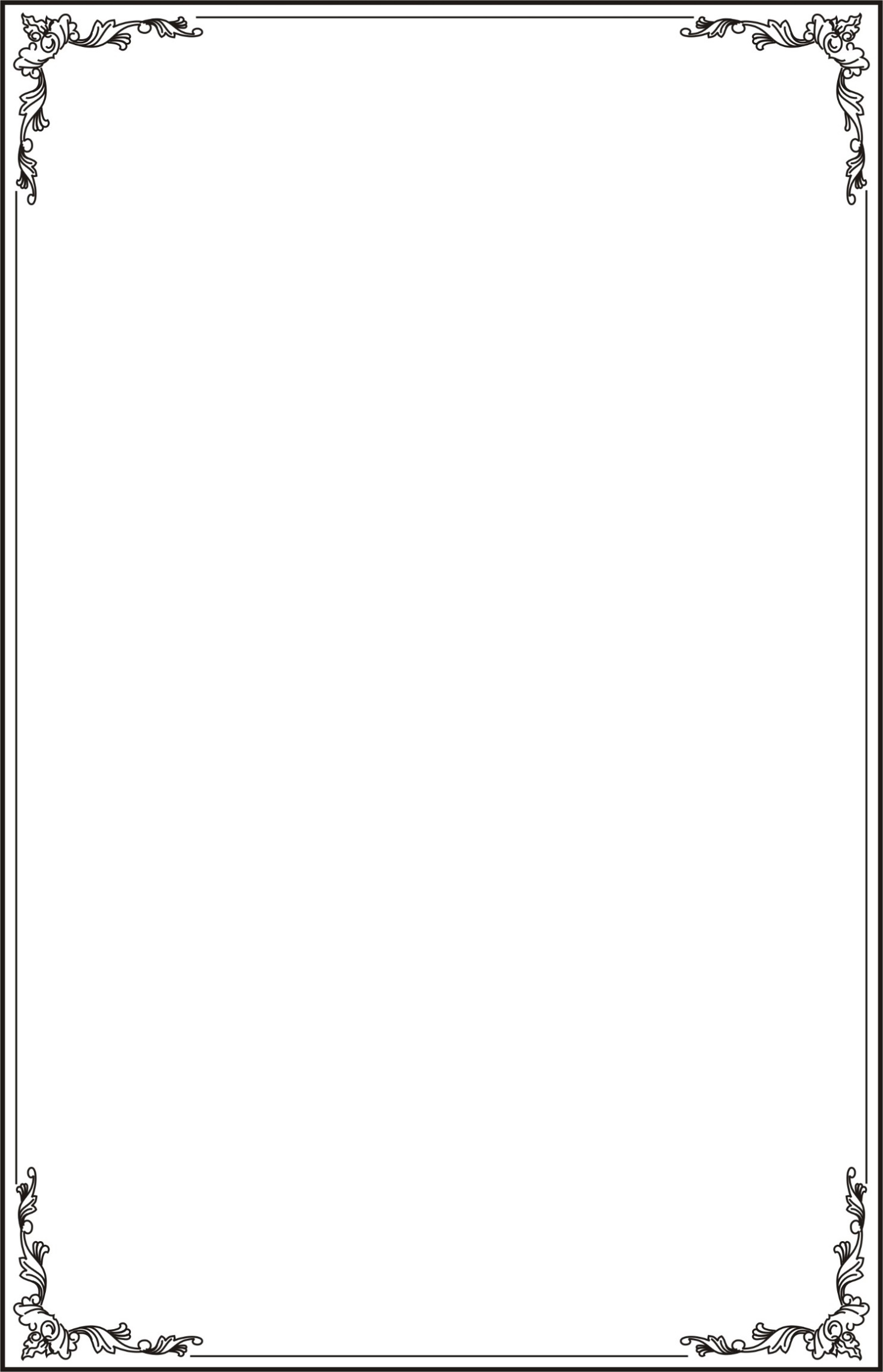
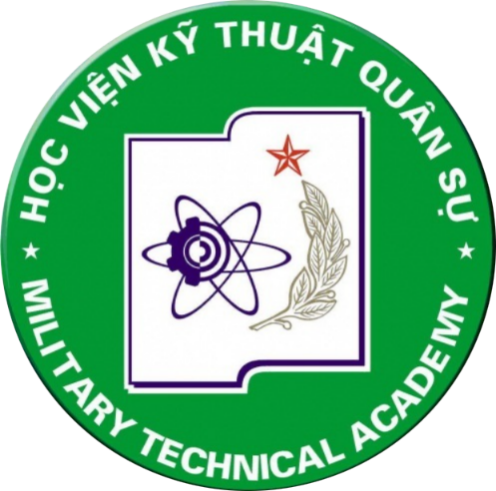
**HỌC VIỆN KỸ THUẬT QUÂN SỰ**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO ĐỒ ÁN 1**

**Đề tài: Xây dựng nền tảng Low-Code AI hỗ trợ dự báo chuỗi thời gian**

**Giảng viên : Ts. Phạm Trường Sơn**

**Học viên thực hiện : Hà Mạnh Duy (Lớp BĐATTT)**

**Nguyễn Thị Nga (Lớp BĐATTT)**

Hà Nội, Tháng 05/2023

# MỤC LỤC

# DANH MỤC HÌNH VẼ

# LÍ DO CHỌN ĐỀ TÀI

# TỔNG QUAN VỀ DỮ LIỆU CHUỖI THỜI GIAN (TIME SERIES DATA)

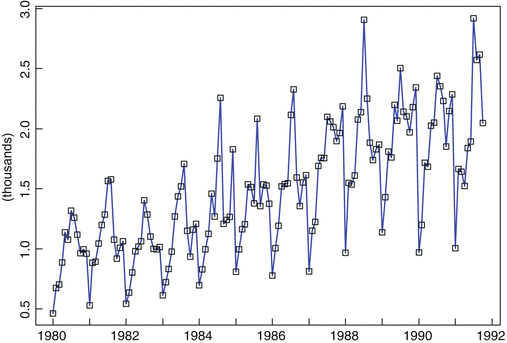
## Khái niệm chuỗi thời gian (Time series)

Một chuỗi thời gian là tập hợp các quan sát thu được thông qua phép đo lặp lại theo thời gian, có nghĩa là được quan sát trong một khoảng thời gian. Thông thường, các quan sát có thể trong toàn bộ khoảng thời gian, được lấy mẫu ngẫu nhiên trong một khoảng thời gian hoặc tại các thời điểm cố định.

Ứng dụng của chuỗi thời gian trải khắp các ngành công nghiệp khác nhau như: quan sát hoạt động sóng điện trong não, đo lượng mưa, dự báo giá cổ phiếu, theo dõi doanh số bán lẻ hàng năm, người đăng ký hàng tháng, nhịp tim mỗi phút,...  
Dữ liệu chuỗi thời gian ở khắp mọi nơi, vì thời gian là thành phần của mọi thứ mà chúng ta có thể nhận biết được.

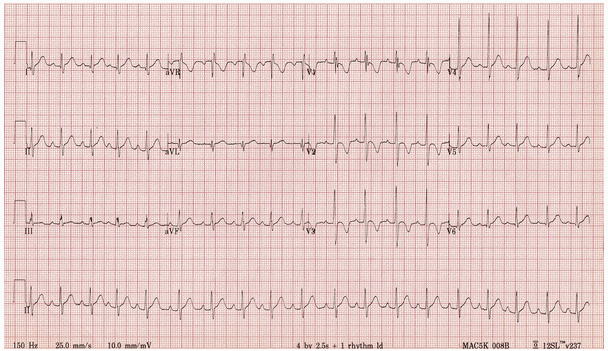
Số liệu chuỗi thời gian đề cập đến một phần dữ liệu được theo dõi theo thời gian tăng dần. Chẳng hạn, một số liệu có thể đề cập đến số lượng hàng tồn kho đã được bán trong một cửa hàng từ ngày này sang ngày khác.

Dưới đây là ví dụ cho thấy doanh số bán rượu vang đỏ hàng tháng (tính bằng kilolit) của các nhà sản xuất rượu vang Úc từ tháng 1 năm 1980 đến tháng 10 năm 1991. (Hình 1.1) [1]



Ở đây, các quan sát thu được theo thời gian là tháng. Nhìn vào biểu đồ, có vẻ như doanh số bán hàng có xu hướng tăng và theo mô hình seasonal với mức cao nhất vào tháng 7 và thấp nhất vào tháng 1.

Một ví dụ quen thuộc khác của dữ liệu chuỗi thời gian là theo dõi sức khỏe bệnh nhân, chẳng hạn như trong điện tâm đồ (ECG), theo dõi hoạt động của tim để cho biết liệu nó có hoạt động bình thường hay không.

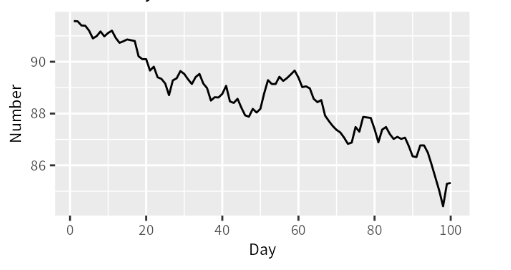


## Các thành phần của Time series

Time series có các thành phần cơ bản như sau:

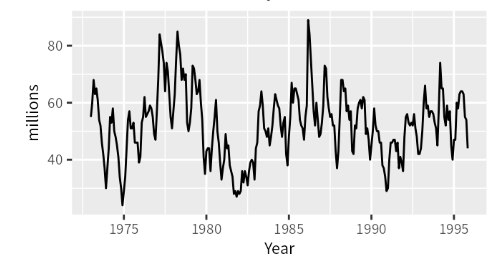
* **Trend**: thành phần xu hướng (tăng hay giảm) của dữ liệu trong một khoảng thời gian dài. Điều này có thể phát hiện thông qua độ dốc của dữ liệu trên biểu đồ.

Dưới đây là ví dụ cho thấy kết quả từ thị trường Chicago trong 100 ngày giao dịch liên tiếp vào năm 1981 với xu hướng giảm rõ rệt.



* **Seasonality** (mùa vụ, thời vụ): dữ liệu bị ảnh hưởng theo thước đo thời gian như thời gian trong năm hoặc ngày hoặc tuần. Trong trường hợp này, dữ liệu có chu kỳ lặp đi lặp lại theo thước đo thời gian kể trên với một tần suất cố định.

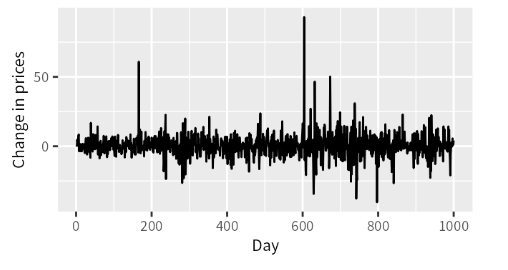
Ví dụ dưới cho thấy doanh số bán nhà ở hàng tháng ở Mỹ cho thấy tính thời vụ theo mỗi năm



* **Cyclic:** thành phần chu kỳ (dài hạn): biểu thị bằng sự tăng, giảm của dữ liệu chuỗi thời gian xoay quanh xu hướng. Thành phần Cyclic thể hiện rõ nhất trong lĩnh vực kinh tế.

Ở hình 1.5, thành phần Cyclic cũng được thể hiện rõ ràng với thời gian khoảng 6-10 năm.

* **Residual:** Đây là thành phần ngược với chu kỳ, Residual chỉ sự thay đổi bất thường của giá trị trong chuỗi thời gian và thường không dự đoán được sự thay đổi của dữ liệu.



Ví dụ trên là sự thay đổi hàng ngày trong giá cổ phiếu đóng cửa của Google. Ta thấy rằng không có xu hướng, tính thời vụ hoặc theo chu kỳ thể hiện trong biểu đồ trên. Có những biến động ngẫu nhiên dường như không thể dự đoán được.

## Tính ổn định của chuỗi thời gian

Một tính chất rất quan trọng của chuỗi thời gian là tính ổn định (Stationarity). Tính ổn định trong chuỗi thời gian đề cập đến thuộc tính của chuỗi bao gồm: kì vọng hay giá trị trung bình (mean), phương sai (variance) và hiệp phương sai (covariance) không đổi theo thời gian.

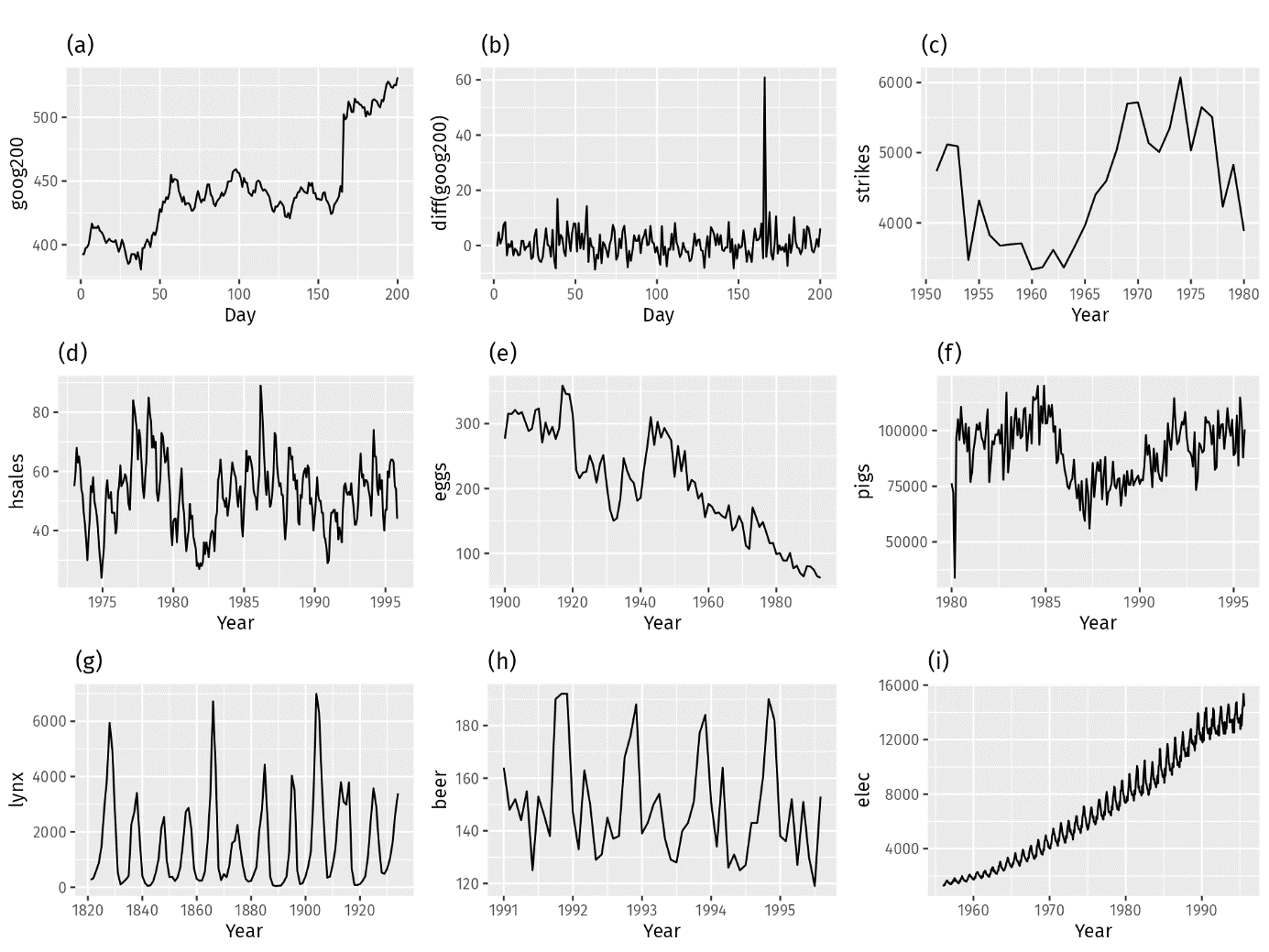
Xét chuỗi thời gian . Trung bình chuỗi kí hiệu là , là giá trị kì vọng, là phương sai, là hiệp phương sai ứng với 2 chuỗi .

Ta có các công thức như sau: [3]

|  |
| --- |
|  |
| = |
|  |

### Kiểm tra tính ổn định của chuỗi thời gian

Chuỗi thời gian ổn định là chuỗi mà các thuộc tính của nó không phụ thuộc vào thời gian mà chuỗi đó được quan sát. Do đó, chuỗi thời gian có xu hướng hoặc có tính thời vụ, không ổn định - xu hướng và tính thời vụ sẽ ảnh hưởng đến giá trị của chuỗi thời gian tại các thời điểm khác nhau.  
Mặt khác, white noise series là đứng yên - không quan trọng khi ta quan sát nó, nó sẽ trông giống nhau tại bất kỳ thời điểm nào.  
Một số trường hợp có thể gây nhầm lẫn - một chuỗi thời gian có hành vi theo chu kỳ (nhưng không có xu hướng) là ổn định. Điều này là do các chu kỳ không có độ dài cố định, vì vậy trước khi quan sát chuỗi, chúng ta không thể chắc chắn đâu sẽ là đỉnh và đáy của các chu kỳ.

Nói chung, một chuỗi thời gian cố định sẽ không có mô hình dự đoán được trong dài hạn. Biểu đồ thời gian sẽ cho thấy chuỗi gần như nằm ngang (mặc dù có thể xảy ra một số hành vi theo chu kỳ), với phương sai không đổi.  


Tính thời vụ rõ ràng quy ra chuỗi (d), (h) và (i). Các xu hướng và mức độ thay đổi quy định chuỗi (a), (c), (e), (f) và (i). Phương sai tăng cũng loại trừ (i). Điều đó chỉ để lại (b) và (g) là chuỗi ổn định.

Khi một chuỗi thời gian không ổn định, việc áp dụng các phương pháp phân tích và mô hình hóa trực tiếp có thể không mang lại kết quả chính xác hoặc đáng tin cậy. Trong trường hợp này, thường cần thực hiện các biến đổi hoặc điều chỉnh để đạt được tính ổn định. Các biến đổi như chuyển đổi chia, loại bỏ xu hướng, loại bỏ chu kỳ hoặc ước lượng phương sai không đồng nhất có thể được sử dụng để biến đổi chuỗi thời gian không ổn định thành một chuỗi thời gian ổn định hơn để tiếp tục phân tích và dự đoán.

### Biến đổi về chuỗi ổn định

# Chuyển đổi chia (Transformation): Áp dụng phép biến đổi chia (transformation) lên chuỗi thời gian để giảm đi sự biến động và đưa nó về gần với một trạng thái ổn định hơn. Các biến đổi phổ biến bao gồm:

* Logarithmic transformation: Áp dụng hàm logarithm lên chuỗi thời gian.
* Difference transformation: Lấy sự khác biệt giữa các giá trị liên tiếp trong chuỗi thời gian.
* Square root transformation: Áp dụng hàm căn bậc hai lên chuỗi thời gian.

# Xóa bỏ xu hướng (Detrending): Nếu chuỗi thời gian có xu hướng tăng/giảm, ta có thể loại bỏ xu hướng đó để đưa nó về một trạng thái ổn định hơn. Có thể sử dụng phương pháp hồi quy tuyến tính để ước lượng và xóa bỏ xu hướng.

# Kiểm tra và loại bỏ sự chu kỳ (Seasonality): Nếu chuỗi thời gian có sự chu kỳ, ta có thể loại bỏ thành phần chu kỳ để đạt được tính ổn định. Phương pháp phổ biến để loại bỏ sự chu kỳ là phương pháp khử trừ mùa vụ (seasonal differencing) bằng cách lấy sự khác biệt giữa các giá trị cùng một mùa vụ trong chuỗi thời gian.

# Kiểm tra và loại bỏ phương sai không đồng nhất (Heteroscedasticity): Nếu chuỗi thời gian có phương sai không đồng nhất, ta có thể áp dụng phép biến đổi như logarit hoặc chia cho căn bậc hai để đạt được tính ổn định hơn.

# CÁC MÔ HÌNH DỰ BÁO CHUỖI THỜI GIAN

## Tổng quan về dự báo chuỗi thời gian (Time Series Forecasting)

Dự báo chuỗi thời gian hiểu theo một cách đơn giản có nghĩa là dự báo hoặc dự đoán giá trị tương lai (ví dụ: giá cổ phiếu) trong một khoảng thời gian. Đó là quá trình phân tích dữ liệu chuỗi thời gian bằng cách sử dụng số liệu thống kê và mô hình hóa để đưa ra dự đoán và cung cấp thông tin cho việc ra quyết định chiến lược. Không phải lúc nào dự đoán cũng chính xác và khả năng dự đoán có thể thay đổi nhiều, đặc biệt là khi xử lý các biến thường dao động trong dữ liệu chuỗi thời gian cũng như các yếu tố nằm ngoài tầm kiểm soát của chúng ta. Tuy nhiên, dự báo cái nhìn sâu sắc về kết quả nào có nhiều khả năng hoặc ít có khả năng xảy ra hơn so với các kết quả tiềm năng khác. Thông thường, chúng ta có dữ liệu càng toàn diện thì dự báo càng chính xác.

Dự báo chuỗi thời gian là một trong những kỹ thuật khoa học dữ liệu phổ biến được sử dụng trong nhiều ngành công nghiệp và lĩnh vực khoa học . Một số trường hợp sử dụng dự báo chuỗi thời gian:

* Trong lĩnh vực tài chính, kinh doanh, marketing: dự báo nhu cầu về định giá; dự báo doanh số bán hang, lợi nhuận và các chỉ số kinh doanh để giúp đưa ra kế hoạch sản xuất, kinh doanh phù hợp; dự báo xu hướng tiêu dùng, xu hướng thị trường..
* Trong y tế: dự báo xu hướng dịch bệnh; mô hình bệnh lây lay; nhận dạng mẫu…
* Dự báo thời tiết.
* Phát hiện gian lận: xác định các đợt tăng đột biến bất thường, lệch đáng kể so với cách nhìn của cả trend và seasonality.

Khi thực hiện dự báo chuỗi thời gian, có khá nhiều yếu tố liên quan chúng ta cần phải xem xét, quan trọng nhất cần quan tâm các yếu tố sau:

* *Số lượng dữ liệu:* nhiều dữ liệu hơn thường mang lại cơ hội hơn cho phân tích dữ liệu, thử nghiệm và điều chỉnh mô hình cũng như độ trung thực của mô hình. Từ đó mô hình sẽ dữ báo tốt hơn.
* *Chất lượng dữ liệu*: dữ liệu đòi hỏi cần thỏa mãn một số yêu cầu như: không có dữ liệu trùng lặp, định dạng dữ liệu được tiêu chuẩn hóa và dữ liệu được thu thập một cách nhất quán hoặc theo định kỳ.
* *Trends:*  là một trong những yếu tố quan trọng mô tả rằng chắc chắn có chuỗi thời gian xu hướng tăng hoặc giảm trong một khoảng thời gian nhất định.
* *Seasonality*: nghĩa là có những khoảng thời gian riêng biệt khi dữ liệu chứa những điểm bất thường nhất quán. Hiểu đơn giản là trong khi dự đoán dữ liệu chuỗi thời gian, có một số tháng trong một miền cụ thể có giá trị đầu ra ở mức cao nhất so với các tháng khác.

Ví dụ, nhìn vào dữ liệu của một công ty du lịch nào đó, rõ ràng dữ liệu vào những khoảng thời gian mùa lễ hội sẽ cao hơn rất nhiều so với các khoảng thời gian khác.

* *Sự kiện bất ngờ:* các sự kiện này luôn có thể xảy ra, chúng thể hiện sự nhiễu trong dữ liệu lịch sử và không dự đoán được. Sự kiện bất ngời xảy ra một cách linh hoạt nên việc nắm bắt trở nên khá khó khăn. Chúng ta cần xem xét các sự kiện này khi tạo mô hình dự báo.

## Mô hình ARIMA

## Mô hình RNN

## Meachine learning trong dự báo chuỗi thời gian

# TỔNG QUAN VỀ LOW-CODE AI

# XÂY DỰNG NỀN TẢNG LOW-CODE AI HỖ TRỢ DỰ BÁO CHUỖI THỜI GIAN

# DANH MỤC TÀI LIỆU THAM KHẢO

[1] Peter J. Brockwell, Richard A. Davis. “Introduction to Time Series and Forecasting”. Third Edition 2016

[2] Chi-Chen Wang. “A comparision study between fuzzy time series model and ARIMA model for forecasting Taiwan export”. Expert Systems with Applications, vol.38, no.8, pp.9296-9304, 2011

[3] Pham Dinh Khanh – Model AR regression. Octorber 15, 2018

Trong lớp statsmodels.tsa.arima.model.ARIMA, các tham số được sử dụng để xác định và tùy chỉnh mô hình ARIMA (Autoregressive Integrated Moving Average). Dưới đây là giải thích cho các tham số trong lớp ARIMA:

1. endog: Đây là chuỗi dữ liệu đầu vào (endogenous variable) mà chúng ta muốn phân tích và mô hình hóa. Chuỗi dữ liệu này phải là một mảng một chiều hoặc một Series/Pandas DataFrame.
2. exog (tùy chọn): Đây là ma trận dữ liệu exogenous (hoặc các biến không phụ thuộc) mà chúng ta muốn bao gồm trong mô hình. Nếu không có, mặc định là None.
3. order: Đây là bộ ba (p, d, q) xác định các tham số của thành phần ARIMA chính:

* p là số lượng lags (số lượng giá trị tự hồi quy) cho thành phần autoregressive (AR).
* d là số lần lấy tích phân để biến đổi chuỗi dữ liệu thành chuỗi dừng (stationary). Thông thường, nếu d > 0, mô hình sẽ áp dụng phép lấy sai phân (difference) để loại bỏ xu hướng trong dữ liệu.
* q là số lượng lags (số lượng giá trị trung bình chuyển động) cho thành phần moving average (MA).

1. seasonal\_order: Đây là bộ tứ (P, D, Q, s) xác định các tham số cho thành phần ARIMA mùa vụ (seasonal ARIMA):

* P là số lượng lags cho thành phần autoregressive mùa vụ.
* D là số lần lấy tích phân mùa vụ.
* Q là số lượng lags cho thành phần moving average mùa vụ.
* s là số lượng quý trình trong một chu kỳ mùa vụ (ví dụ: s = 4 cho dữ liệu quý).

1. trend: Đây là yếu tố xu hướng có thể có các giá trị sau:

* 'c' (hằng số): Mô hình có một hằng số.
* 'nc' (không có hằng số): Mô hình không có hằng số.
* 't' (linear trend): Mô hình có một xu hướng tuyến tính.
* 'ct' (constant plus linear trend): Mô hình có một hằng số và một xu hướng tuyến tính.

1. enforce\_stationarity: Đặt giá trị True để đảm bảo chuỗi dữ liệu đã xác định là dừng (stationary) theo cách thức đã được chỉ định. Nếu chuỗi dữ liệu không phải dừng và enforce\_stationarity được đặt là True, mô hình sẽ áp dụng phép lấy sai phân (difference) để biến đổi dữ liệu thành dữ liệu dừng trước khi xây dựng mô hình.
2. enforce\_invertibility: Đặt giá trị True để đảm bảo rằng mô hình ARMA đã xây dựng là có khả năng đảo ngược (invertible). Mô hình ARMA có thể được coi là có khả năng đảo ngược nếu tất cả các giá trị tự hồi quy (AR) và giá trị trung bình chuyển động (MA) nằm trong miền chấp nhận được của mô hình ARMA.
3. concentrate\_scale: Đặt giá trị True để tối ưu hóa phương sai thay đổi dựa trên quy tắc sigma2 (phương sai) trong quá trình ước lượng. Nếu False, ước lượng sẽ được thực hiện theo phương pháp MLE (Maximum Likelihood Estimation).
4. trend\_offset: Đây là tham số thay đổi vị trí xu hướng trong chuỗi dữ liệu. Giá trị mặc định là 1, tức là xu hướng sẽ bắt đầu từ vị trí đầu tiên của chuỗi dữ liệu.
5. dates: Đây là một mảng các ngày thể hiện thời gian tương ứng với từng điểm dữ liệu trong chuỗi. Nếu không được cung cấp, mặc định là None.
6. freq: Đây là xác định tần suất thời gian của chuỗi dữ liệu. Nếu không được cung cấp, mặc định là None.