MÔN: CÁC MÔ HÌNH DỰ BÁO

Mô hình dự báo ARIMA dự báo dịch COVID-19

20 December, 2022

Abstract

Trong báo cáo này nhóm 2 sẽ trình bày về mô hình ARIMA áp dụng vào bộ dữ liệu COVID-19 sau đó trực quan dữ liệu COVID-19 theo TS. Và cuối cùng là xây dựng mô hình ARIMA dự báo xu hướng dịch COVID-19 trong một khoảng thời gian cụ thể.

1 TỔNG QUAN VỀ LÝ THUYẾT VÀ THAO TÁC DỮ LIỆU CHUỖI THỜI GIAN

1.1 Chuỗi thời gian

1.1.1 Khái niệm

Là chuỗi các điểm dữ liệu được đo theo từng khoảng thời gian liền nhau, khoảng cách giữa các lần đo là bằng nhau. (Lưu ý rằng dữ liệu mà ta đang nói ở đây phải là tập các biến ngẫu nhiên - stochastic process).

Ví dụ:

- Giá mở cửa của chỉ số ngoại hối EURUSD, được thống kê vào 0 giờ 0 phút hàng ngày
- Nhiệt độ trung bình theo ngày.

1.1.2 Đặc điểm và tính chất

Điều làm cho dữ liệu chuỗi thời gian khác biệt với các dữ liệu khác là phân tích có thể chỉ ra cách các biến số thay đổi theo thời gian. Nói cách khác, thời gian là một biến số quan trọng vì nó cho biết dữ liệu điều chỉnh như thế nào trong suốt quá trình của các điểm dữ liệu cũng như kết quả cuối cùng. Nó cung cấp một nguồn thông tin bổ sung và một thứ tự phụ thuộc giữa các dữ liệu.

Phân tích chuỗi thời gian thường yêu cầu một số lượng lớn các điểm dữ liệu để đảm bảo tính nhất quán và độ tin cậy. Tập hợp dữ liệu mở rộng đảm bảo bạn có cỡ mẫu đại diện và phân tích đó có thể loại bỏ dữ liệu nhiễu. Nó cũng đảm bảo rằng bất kỳ xu hướng hoặc mẫu nào được phát hiện đều không phải là ngoại lệ và có thể giải thích cho sự khác biệt. Ngoài ra, dữ liệu chuỗi thời gian có thể được sử dụng để dự báo—dự đoán dữ liệu trong tương lai dựa trên dữ liệu lịch sử.

- **Xu hướng** nghĩa là về trung bình, các phép đo có xu hướng tăng (hoặc giảm) theo thời gian
- **Có tính thời vụ** nghĩa là có một mô hình cao và thấp lặp lại thường xuyên liên quan đến thời gian theo lịch như mùa, quý, tháng, ngày trong tuần, v.v.
- **Ngoại lệ** với dữ liệu chuỗi thời gian, các giá trị ngoại lệ của cách xa dữ liệu khác của bạn.
- **Chu kỳ** hoặc khoảng thời gian dài hạn không liên quan đến các yếu tố thời vụ
- **Phương sai** không đổi theo thời gian hay phương sai không đổi
- **Thay đổi đột ngột** nào đối với cấp độ của chuỗi hoặc phương sai

1.2 Một số khái niệm về Time Series

1.2.1 Sai phân(difference)

a) Khái niệm

- Là phương pháp để biến đổi một chuỗi time series thành chuỗi dừng (stationary), để loại bỏ xu hướng (trend), hay sự tự tương quan (auto-correlation).
- Sai phân là một phương pháp chuyển đổi tập dữ liệu chuỗi thời gian. Nó có thể được sử dụng để loại bỏ sự phụ thuộc của chuỗi vào thời gian, cái gọi là sự phụ thuộc vào thời gian. Điều này bao gồm các cấu trúc như xu hướng và tính thời vụ.
- Sai phân được thực hiện bằng cách trừ quan sát trước đó khỏi quan sát hiện tại:

$$Difference(t) = Observation(t) - Observation(t-1)$$

- Đảo ngược quy trình là cần thiết khi một dự đoán phải được chuyển đổi trở lại tỷ lệ ban đầu.
- Quá trình này có thể được đảo ngược bằng cách thêm quan sát ở bước thời gian trước vào giá trị chênh lệch:

$$Inverted(t) = Differenced(t) + Observation(t-1)$$

- Bằng cách này, có thể tính toán một loạt các chênh lệch và chênh lệch đảo ngược.

b) Một số tính chất trong sai phân

1. Chêch lệch độ trễ

- Lấy sự khác biệt giữa các quan sát liên tiếp được gọi là chênh lệch $lag - 1$.
- Sự khác biệt về độ trễ có thể được điều chỉnh để phù hợp với cấu trúc thời gian cụ thể.
- Đối với chuỗi thời gian có thành phần theo mùa, độ trễ có thể được coi là khoảng thời gian (độ rộng) của tính thời vụ.

2. Lệnh chêch lệnh

- Một số cấu trúc thời gian có thể vẫn tồn tại sau khi thực hiện thao tác phân biệt, chẳng hạn như trong trường hợp xu hướng phi tuyến tính.
- Như vậy, quá trình khác biệt có thể được lặp đi lặp lại nhiều lần cho đến khi loại bỏ tất cả sự phụ thuộc về thời gian.
- Số lần mà sự khác biệt được thực hiện được gọi là thứ tự khác biệt.

3. Tính chêch lệnh

- Chúng tôi có thể phân biệt tập dữ liệu theo cách thủ công.

• Điều này liên quan đến việc phát triển một chức năng mới để tạo ra một bộ dữ liệu khác biệt.

• Hàm sẽ lặp qua một chuỗi được cung cấp và tính toán các giá trị khác nhau ở khoảng thời gian hoặc độ trễ đã chỉ định.

4. Khác biệt để loại bỏ xu hướng

• Trong phần này, chúng ta sẽ xem xét việc sử dụng biến đổi chênh lệch để loại bỏ một xu hướng.

• Một xu hướng làm cho một chuỗi thời gian không cố định bằng cách tăng mức độ.

• Điều này có tác dụng thay đổi giá trị chuỗi thời gian trung bình theo thời gian.

5. Khác biệt để loại bỏ tính thời vụ

• Trong phần này, chúng ta sẽ xem xét việc sử dụng phép biến đổi khác biệt để loại bỏ tính thời vụ.

• Sự thay đổi theo mùa, hay tính thời vụ, là những chu kỳ lặp lại thường xuyên theo thời gian.

• Một mẫu lặp lại trong mỗi năm được gọi là biến thể theo mùa, mặc dù thuật ngữ này được áp dụng chung hơn cho các mẫu lặp lại trong bất kỳ khoảng thời gian cố định nào.

• Có nhiều loại tính thời vụ. Một số ví dụ rõ ràng bao gồm; thời gian trong ngày, hàng ngày, hàng tuần, hàng tháng, hàng năm, v.v.

• Do đó, việc xác định xem có thành phần thời vụ nào trong vấn đề chuỗi thời gian của bạn hay không là chủ quan.

• Cách tiếp cận đơn giản nhất để xác định xem có khía cạnh nào của tính thời vụ hay không là vẽ biểu đồ và xem xét dữ liệu của bạn, có thể ở các tỷ lệ khác nhau và có thêm các đường xu hướng.

1.2.2 Độ dừng(stationary)

a) Khái niệm

• Một chuỗi thời gian có tính dừng khi các giá trị **mean**, **variance**, **autocorrelation** không thay đổi theo thời gian.

• Với hầu hết các phương pháp thống kê dự báo, ta đều phải đảm bảo tính dừng của chuỗi dữ liệu vì thế việc kiểm tra tính dừng là rất quan trọng.

b) Một số tính chất trong độ dừng

1. Chuỗi thời gian cố định

• Các quan sát trong một chuỗi thời gian dừng không phụ thuộc vào thời gian.

• Chuỗi thời gian cố định hay chuỗi thời gian dừng nếu chúng không có xu hướng hoặc hiệu ứng theo mùa.

• Số liệu thống kê tóm tắt được tính toán trên chuỗi thời gian nhất quán theo thời gian, chẳng hạn như giá trị trung bình hoặc phương sai của các quan sát.

• Khi một chuỗi thời gian đứng yên, việc lập mô hình có thể dễ dàng hơn. Các phương pháp lập mô hình thống kê giả định hoặc yêu cầu chuỗi thời gian là cố định.

2. Chuỗi thời gian không cố định

• Các quan sát từ chuỗi thời gian không cố định cho thấy các hiệu ứng theo mùa, xu hướng và các cấu trúc khác phụ thuộc vào chỉ số thời gian.

• Số liệu thống kê tóm tắt như giá trị trung bình và phương sai thay đổi theo thời gian, cung cấp sự thay đổi trong các khái niệm mà một mô hình có thể cố gắng nắm bắt.

• Các phương pháp phân tích và dự báo chuỗi thời gian cổ điển liên quan đến việc làm cho dữ liệu chuỗi thời gian không cố định trở nên ổn định bằng cách xác định và loại bỏ các xu hướng cũng như loại bỏ các hiệu ứng tĩnh.

3. Tạo chuỗi dữ liệu cố định

• Bạn có thể kiểm tra xem chuỗi thời gian của mình có dừng hay không bằng cách xem biểu đồ đường của chuỗi theo thời gian.

• Dấu hiệu của các xu hướng rõ ràng, tính thời vụ hoặc các cấu trúc có hệ thống khác trong chuỗi là các chỉ số của một chuỗi không cố định.

• Một phương pháp chính xác hơn là sử dụng một bài kiểm tra thống kê, chẳng hạn như bài kiểm tra Dickey-Fuller.

• Nếu có xu hướng và tính thời vụ rõ ràng trong chuỗi thời gian của mình, thì hãy lập mô hình các thành phần này, loại bỏ chúng khỏi các quan sát, sau đó huấn luyện các mô hình trên phần dư.

• Nếu chúng khớp một mô hình cố định với dữ liệu, thì có thể cho rằng dữ liệu đó là của một quy trình cố định. Vì vậy, bước đầu tiên của trong quá trình phân tích là kiểm tra xem có bất kỳ bằng chứng nào về xu hướng hoặc tác động theo mùa hay không và nếu có thì hãy loại bỏ chúng.

• Các phương pháp chuỗi thời gian thống kê và thậm chí cả các phương pháp học máy hiện đại sẽ được hưởng lợi từ tín hiệu rõ ràng hơn trong dữ liệu.

1.2.3 Nhiễu trắng(white-noise)

a) Khái niệm

• Nhiễu trắng là một khái niệm quan trọng trong phân tích và dự báo chuỗi thời gian.

• Một chuỗi thời gian t được gọi là nhiễu trắng nếu nó thỏa mãn kì vọng bằng 0, các giá trị với các độ trễ khác nhau không có hiện tượng tự tương quan và phương sai sai số không đổi.

• Do kì vọng và phương sai không đổi nên chúng ta gọi phân phối của nhiễu trắng là phân phối xác định (identical distribution).

• Nhiễu trắng là một thành phần ngẫu nhiên thể hiện cho yếu tố không thể dự báo của model do nó không có tính qui luật.

b) Tại sao nhiễu trắng quan trọng?

• Nó quan trọng vì hai lý do chính:

- Khả năng dự đoán: Nếu chuỗi thời gian của bạn là nhiễu trắng, thì theo định nghĩa, nó là ngẫu nhiên. Bạn không thể mô hình hóa nó một cách hợp lý và đưa ra dự đoán.
- Chẩn đoán mô hình: Chuỗi lỗi từ mô hình dự báo chuỗi thời gian lý tưởng nhất phải là nhiễu trắng.

• Dữ liệu chuỗi thời gian dự kiến sẽ chứa một số thành phần nhiễu trắng bên trên tín hiệu do quy trình cơ bản tạo ra.

• Ví dụ:

$$y(t) = tnhiu(t) + nhiu(t)$$

• Khi các dự đoán đã được thực hiện bởi một mô hình dự báo theo chuỗi thời gian, chúng có thể được thu thập và phân tích. Một loạt các lỗi dự báo lý tưởng nhất là nhiễu trắng.

• Khi lỗi dự báo là nhiễu trắng, điều đó có nghĩa là tất cả thông tin tín hiệu trong chuỗi thời gian đã được mô hình khai thác để đưa ra dự đoán. Tất cả những gì còn lại là những dao động ngẫu nhiên không thể mô hình hóa được.

• Một dấu hiệu cho thấy các dự đoán của mô hình không phải là nhiễu trắng khi một dấu hiệu cho thấy có thể có những cải tiến hơn nữa đối với mô hình dự báo. [1]

1.3 KỸ THUẬT PHÂN TÍCH TIME SERIES

1.3.1 Trend

Trend: thành phần này chỉ ra xu hướng tổng quan của dữ liệu theo thời gian: lên hoặc xuống, tăng hoặc giảm

```
library(fpp)
```

```
## Warning: package 'fpp' was built under R version 4.2.2
```

```
## Loading required package: forecast
```

```
## Warning: package 'forecast' was built under R version 4.2.2
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
## Loading required package: fma
```

```
## Warning: package 'fma' was built under R version 4.2.2
```

```
## Loading required package: expsmooth
```

```
## Warning: package 'expsmooth' was built under R version 4.2.2
```

```
## Loading required package: lmtest
```

```
## Warning: package 'lmtest' was built under R version 4.2.2
```

```
## Loading required package: zoo
```

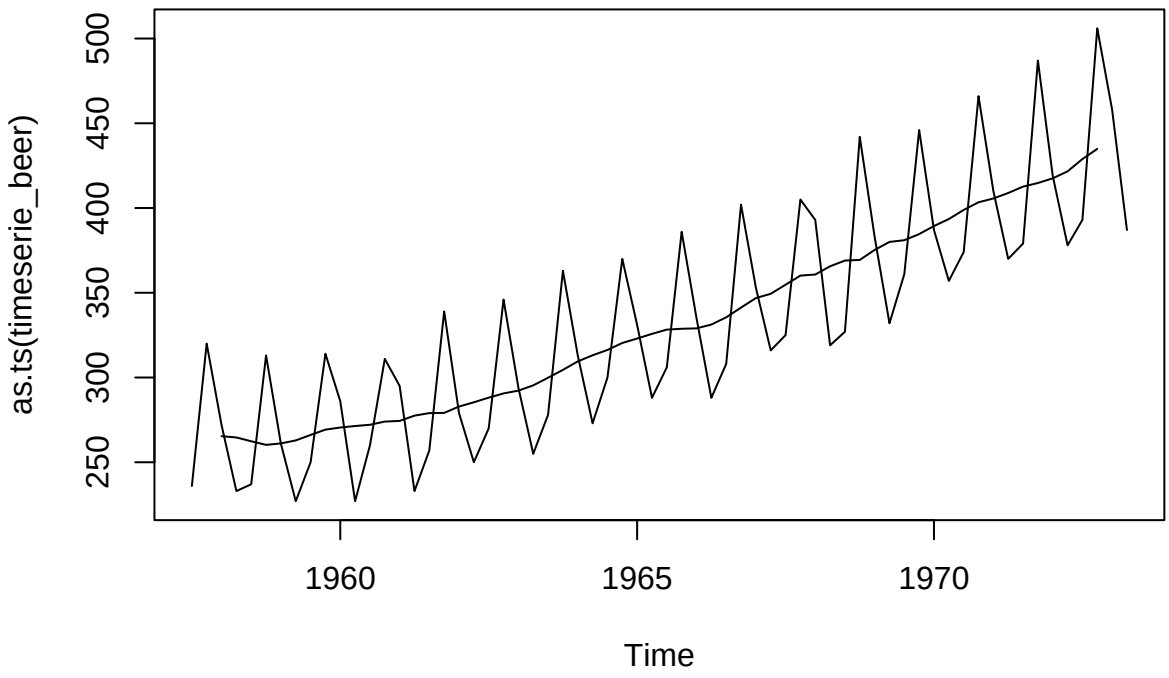
```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

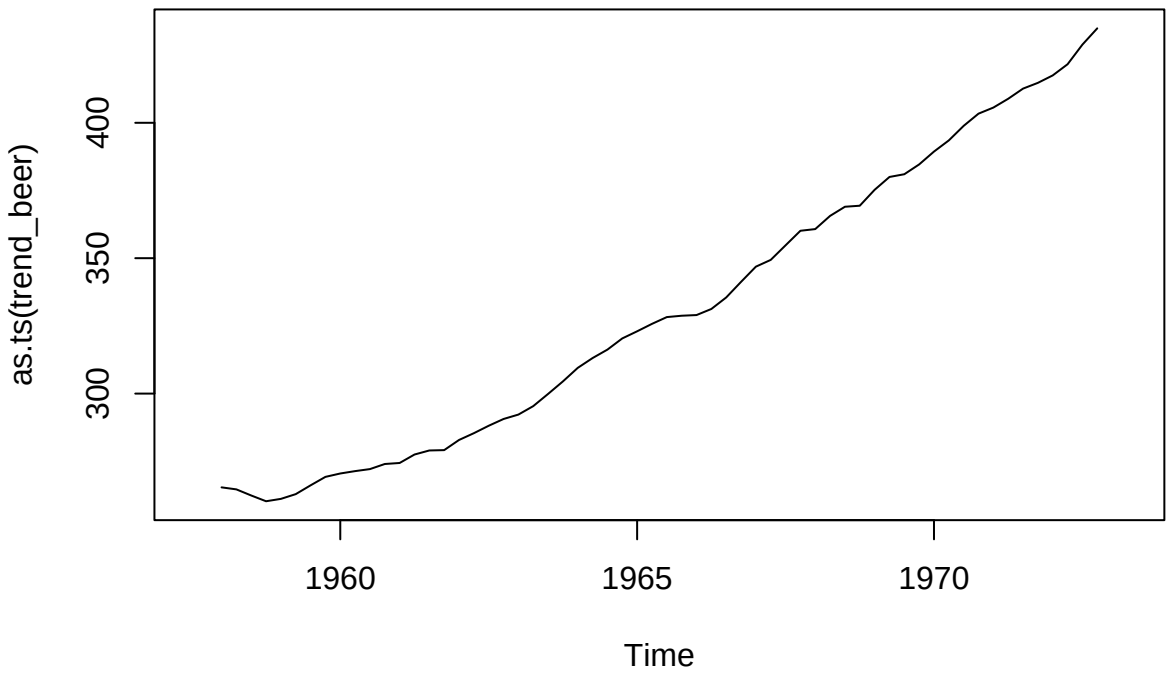
```
## Loading required package: tseries
```

```
## Warning: package 'tseries' was built under R version 4.2.2
```

```
data(ausbeer)
timeserie_beer = tail(head(ausbeer, 17*4+2),17*4-4)
plot(as.ts(timeserie_beer))

library(forecast)
trend_beer = ma(timeserie_beer, order = 4, centre = T)
plot(as.ts(timeserie_beer))
lines(trend_beer)
```



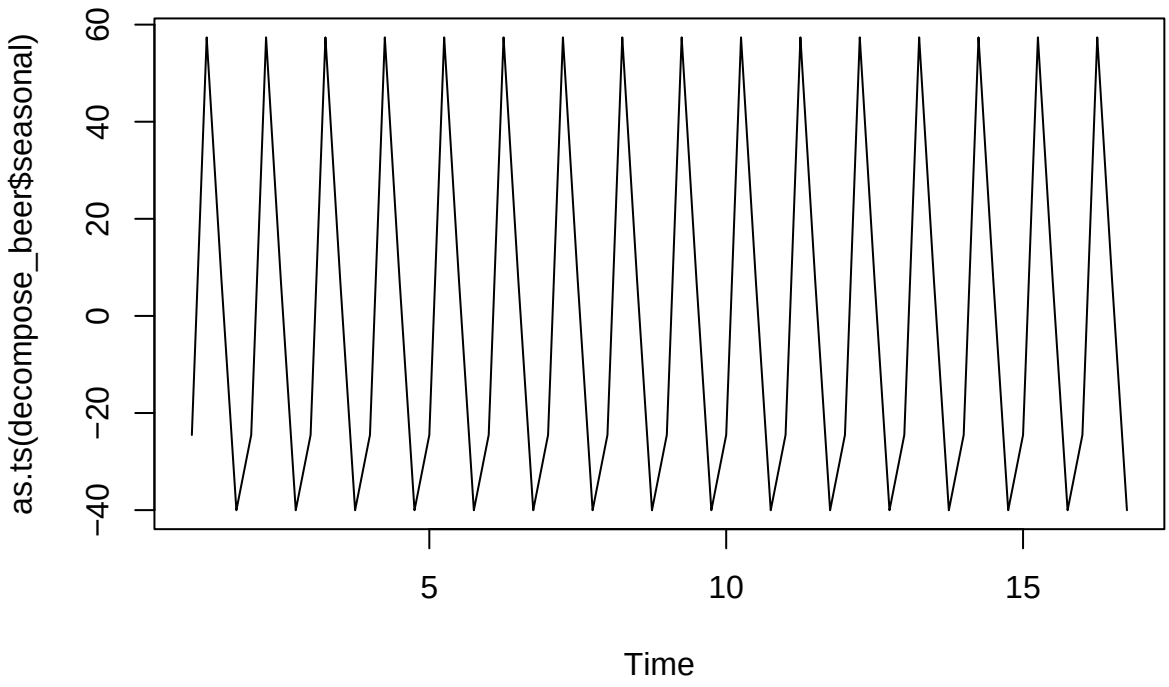

```
plot(as.ts(trend_beer))
```



1.3.2 Seasonality

Seasonality: thành phần chỉ ra các xu hướng theo mùa vị, chỉ ra các pattern theo tháng, theo quý

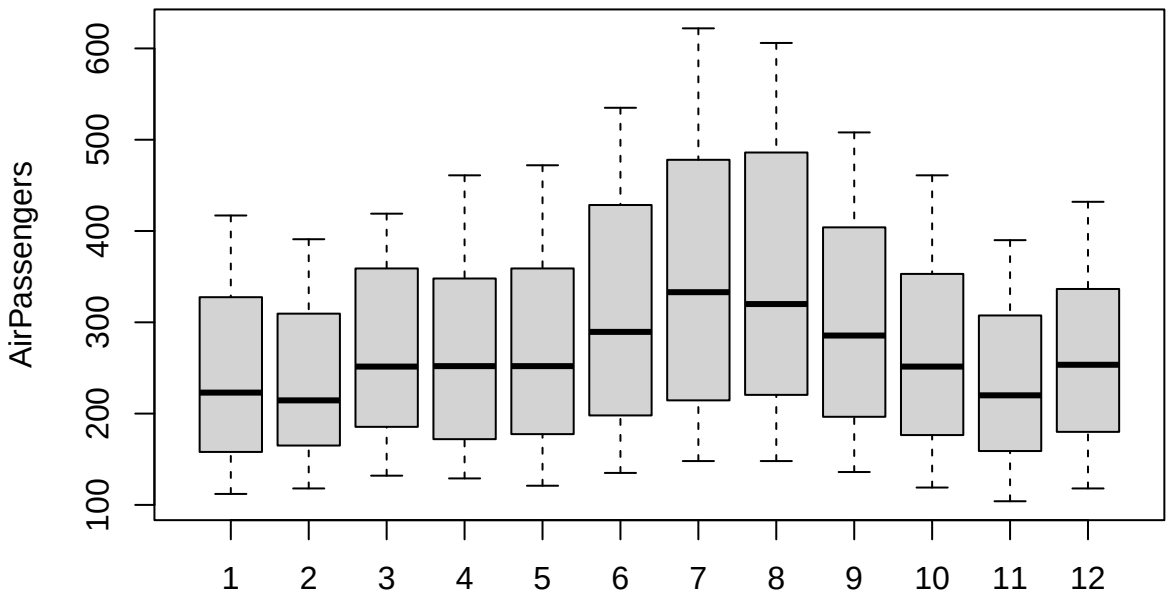
```
ts_beer = ts(timeserie_beer, frequency = 4)
decompose_beer = decompose(ts_beer, "additive")

plot(as.ts(decompose_beer$seasonal))
```



1.3.3 Cycle

Cycle: thành phần chu kỳ có sự vận động trong khoảng thời gian dài (nhiều năm)

```
boxplot(AirPassengers~cycle(AirPassengers, xlab="Date", ylab = "Passenger Numbers (1000's)", main = "Mo
```

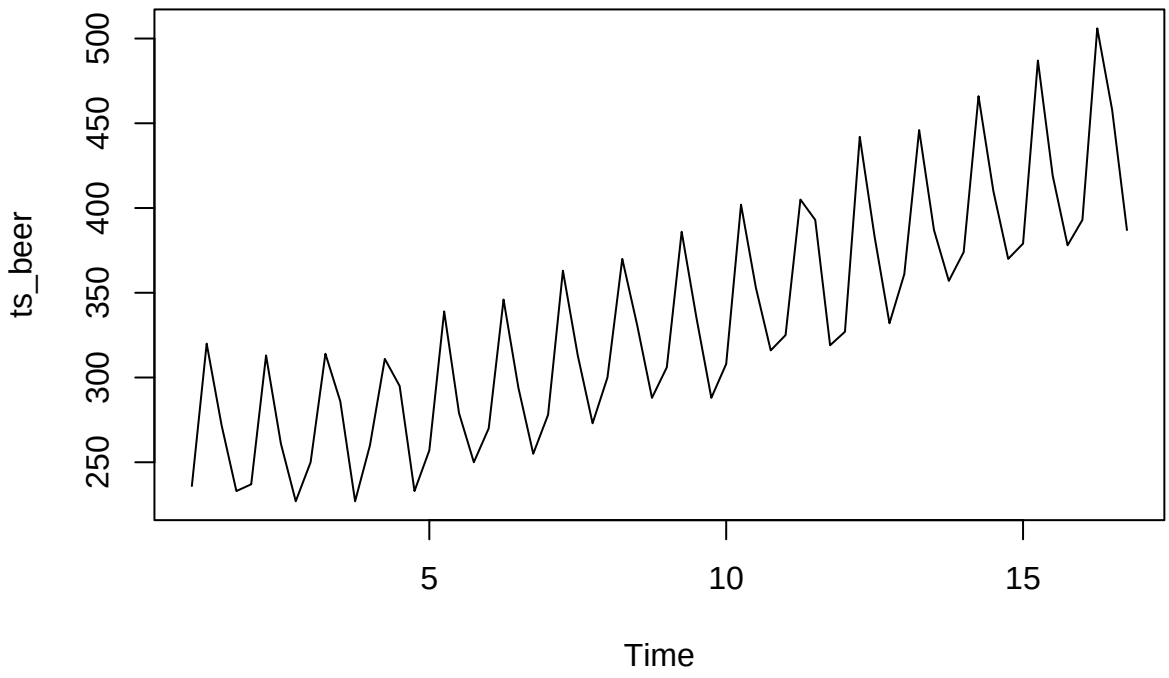


s, xlab = "Date", ylab = "Passenger Numbers (1000's)", main = "Monthly air passengers b

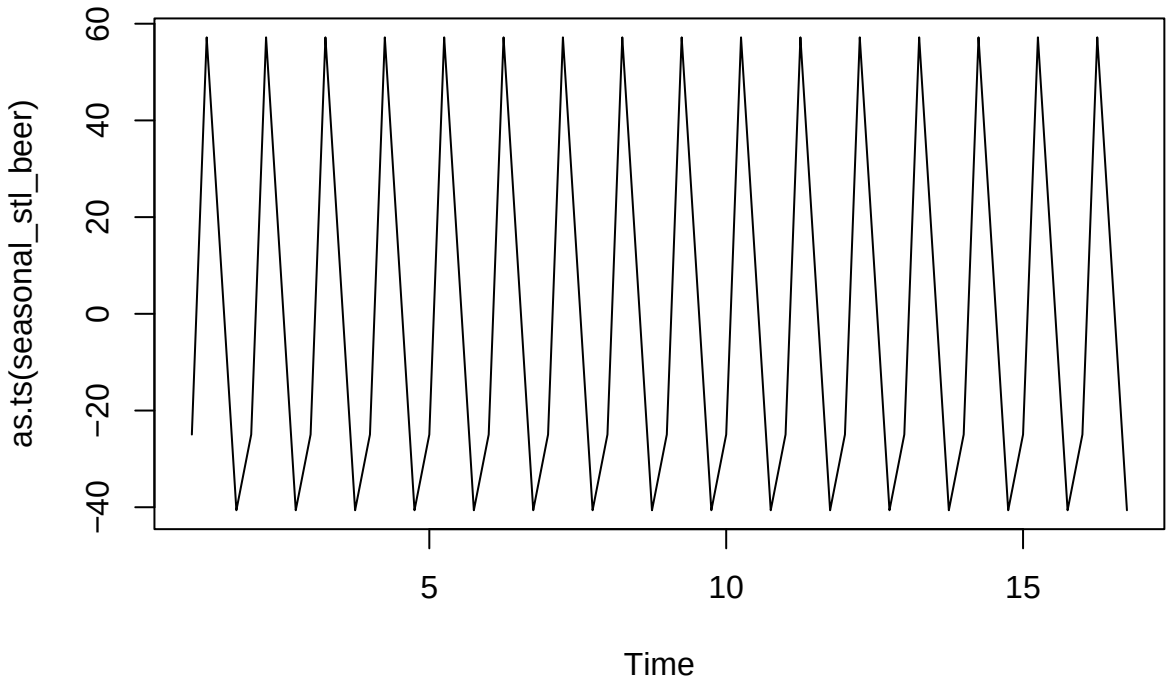
1.3.4 Irregular remainder

Irregular remainder: thành phần nhiễu còn lại sau khi trích xuất hết các thành phần ở trên, nó chỉ ra sự bất thường của các điểm dữ liệu [2]

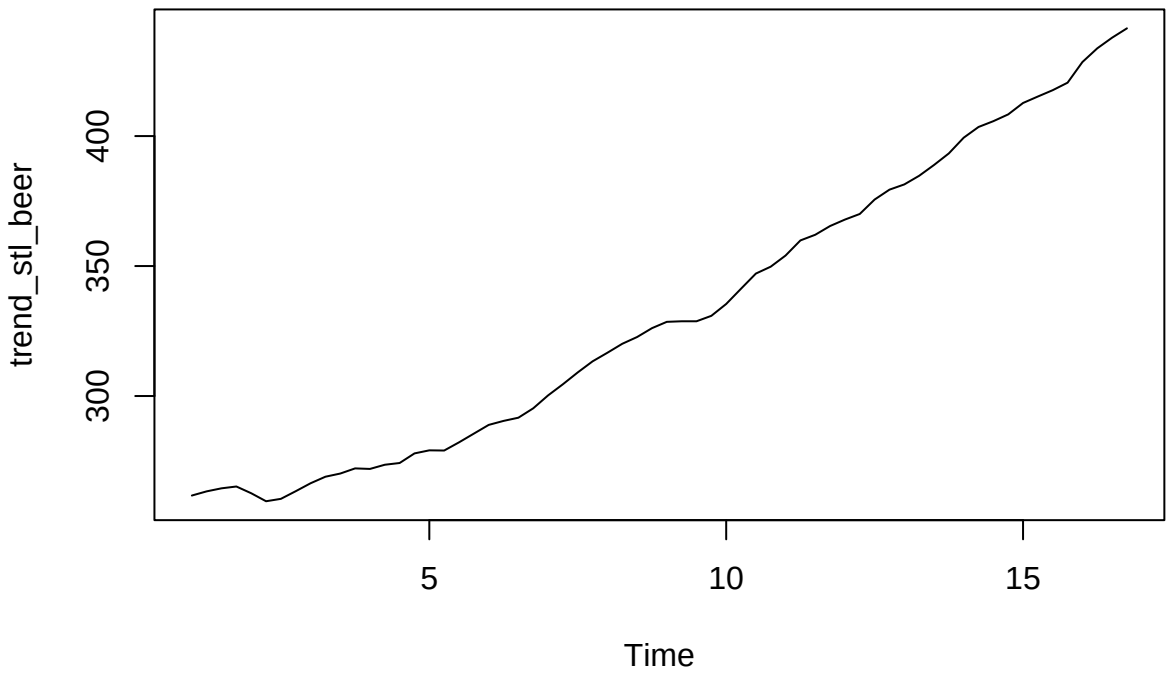
```
ts_beer = ts(timeserie_beer, frequency = 4)
stl_beer = stl(ts_beer, "periodic")
seasonal_stl_beer   <- stl_beer$time.series[,1]
trend_stl_beer     <- stl_beer$time.series[,2]
random_stl_beer  <- stl_beer$time.series[,3]

plot(ts_beer)
```



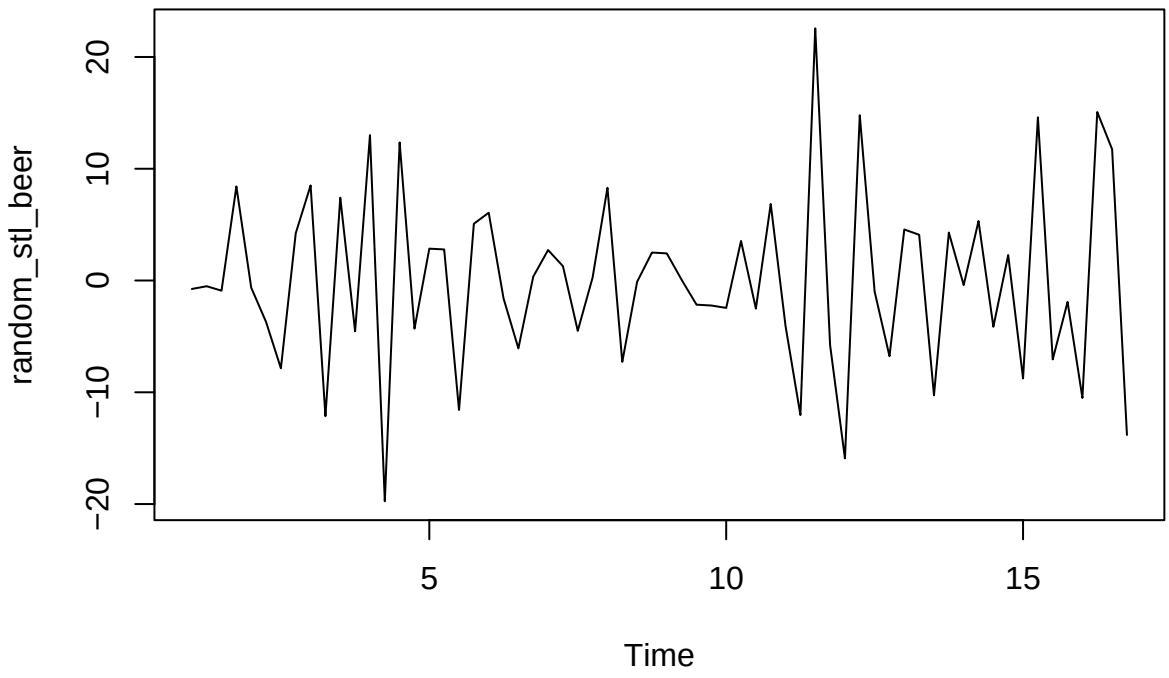

```
plot(as.ts(seasonal_stl_beer))
```



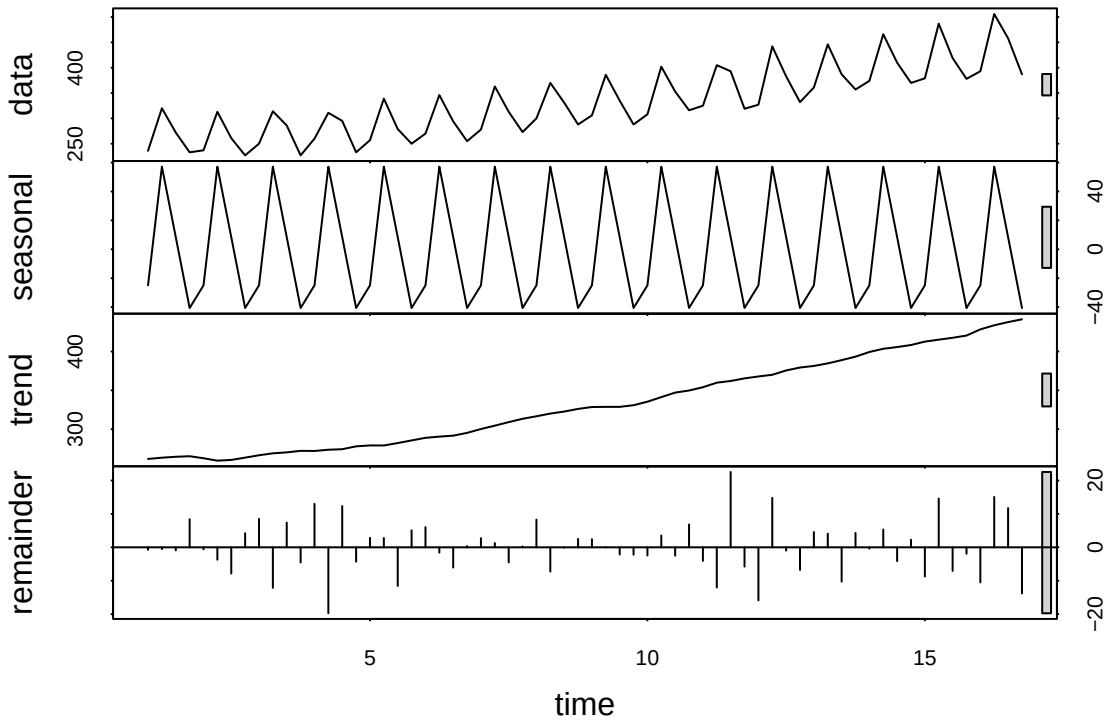

```
plot(trend_stl_beer)
```




```
plot(random_stl_beer)
```




```
plot(stl_beer)
```



2 LÝ THUYẾT VỀ MÔ HÌNH ARIMA

2.1 Mô hình ARIMA dùng trong dự báo

2.1.1 Tổng quan về 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 – Knowledge Discovery and Data 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 thủa 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.

• 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 năm 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.

2.1.2 Phân loại dự báo

• 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 gian 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.

• Mô hình ARIMA (AutoRegressive Integrate Moving Average) do Box-Jenkins đề nghị năm 1976, 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.

2.2 Tính dừng

2.2.1 Khái niệm

• Dữ liệu của bất kỳ chuỗi thời gian nào đều có thể được coi là được tạo ra từ một quá trình ngẫu nhiên và một tập hợp dữ liệu cụ thể có thể được coi là một kết quả cá biệt của quá trình ngẫu nhiên đó. Hay nói các khác, có thể xem quá trình ngẫu nhiên là tổng thể và một tập hợp dữ liệu cụ thể là một mẫu có được của tổng thể đó. Một tính chất của quá trình ngẫu nhiên được các nhà phân tích về chuỗi thời gian đặc biệt quan tâm và xem xét kỹ lưỡng là Tính dừng. Một quá trình ngẫu nhiên Yt được coi là dừng nếu kỳ vọng, phương sai và hiệp phương sai tại cùng một độ trễ của nó không đổi theo thời gian.

Cụ thể, Y_t được gọi là dừng nếu:

- Trung bình: $E(y_t) = (\forall t)$ (1)
- Phương sai: $Var(y_t) = E(y_t{}^-)^2 = \sigma^2(\forall t)$ (2)
- Đồng phương sai: $Cov(y_t, y_{t+k}) = E[(y_t{}^-)(y_{t+k}{}^-)] = \gamma_k(\forall t)$ (3)

Điều kiện thứ 3 có nghĩa là hiệp phương sai giữa y_t và y_{t+k} chỉ phụ thuộc vào độ trễ về thời gian (k) giữa hai thời đoạn chứ không phụ thuộc vào thời điểm t.

Ví dụ:

$$Cov(y_2, y_7) = Cov(y_{10}, y_{15}) = Cov(y_{30}, y_{35}) = ... = Cov(y_t, y_{t+5}).$$

Nhưng $Cov(y_t, y_{t+5})$ có thể khác $Cov(y_t, y_{t+6})$... Quá trình ngẫu nhiên y_t được coi là không dừng nếu nó vi phạm ít nhất một trong ba điều kiện trên.

2.2.2 Hậu quả của chuỗi không dừng

Trong mô hình hồi quy cổ điển, ta giả định rằng sai số ngẫu nhiên có kỳ vọng bằng không, phương sai không đổi và chúng không tương quan với nhau. Với dữ liệu là các chuỗi không dừng, các giả thiết này bị vi phạm, các kiểm định t, F mất hiệu lực, ước lượng và dự báo không hiệu quả hay nói cách khác phương pháp OLS không áp dụng cho các chuỗi không dừng.

Điển hình là hiện tượng hồi quy giả mạo: nếu mô hình tồn tại ít nhất một biến độc lập có cùng xu thế với biến phụ thuộc, khi ước lượng mô hình ta có thể thu được các hệ số có ý nghĩa thống kê và hệ số xác định R2 rất cao. Nhưng điều này có thể chỉ là giả mạo, R2 cao có thể là do hai biến này có cùng xu thế chứ không phải do chúng tương quan chặt chẽ với nhau.

Theo Gujarati (2003), nếu một chuỗi thời gian là không dừng, ta chỉ có thể nghiên cứu hành vi của nó trong phạm vi thời gian đang xem xét. Lúc này chúng ta chỉ xem xét được những tình tiết của hiện tại và quá khứ chứ không dự báo được cho tương lai vì chuỗi không dừng biến động một cách không hội tụ, tức giá trị quá khứ lúc này tác động đến giá trị hiện tại một cách vô hạn và không bao giờ kết thúc. Khi xây dựng các mô hình dự báo, chúng ta giả định rằng các xu hướng hiện tại và quá khứ giữ nguyên chiều hướng vận động cho tương lai. Do vậy, nếu thực hiện mô hình hóa trên một chuỗi dữ liệu không dừng thì việc dự báo cho tương lai rất khập khiễng và dường như không có ý nghĩa. Trong thực tế, phần lớn các chuỗi thời gian đều là chuỗi không dừng, kết hợp với những hậu quả trình bày trên đây cho thấy tầm quan trọng của việc xác định một chuỗi thời gian là có tính dừng hay không

2.2.3 Kiểm định tính dừng

a. Dựa trên đồ thị của chuỗi thời gian

Một cách trực quan chuỗi y_t có tính dừng nếu như đồ thị $y = f(t)$ cho thấy trung bình và phương sai của quá trình y_t không đổi theo thời gian. Ngược lại, nếu nhìn vào đồ thị của một chuỗi theo thời gian mà trung bình của nó có xu hướng tăng hoặc giảm theo từng thời kỳ thì lúc này có thể suy đoán rằng điều kiện một bị vi phạm (điều kiện trung bình không đổi theo thời gian), nên chuỗi đó là không dừng. Phương pháp này cho ta cái nhìn trực quan, đánh giá ban đầu về tính dừng của chuỗi thời gian. Tuy nhiên, với những chuỗi thời gian có xu hướng không rõ ràng, phương pháp này trở nên khó khăn và đôi khi không chính xác.

b. Dựa trên lược đồ tự tương quan và tự tương quan riêng phần

*Tự tương quan (ACF)

Một cách kiểm định đơn giản tính dừng là dùng hàm tự tương quan (ACF - Autocorelation Function), với độ trễ k, ký hiệu bằng ρ_k, được xác định như sau:

$$ACF(k) = \rho_k = \frac{cov(y_t, y_{t-k})}{Var(y_t)}$$

Đại lượng ρ_k không có đơn vị đo, giá trị nằm trong khoảng từ -1 đến 1, là hệ số tương quan giữa y_t và y_{t-k}.Ví dụ: ρ_1 là hệ số tương quan giữa y_t và y_{t-1}, ρ_2 là hệ số tương quan giữa y_t và y_{t-2},... Nếu chúng ta vẽ đồ thị ρ_k theo các độ trễ k, thì đồ thị này sẽ cho ra một lược đồ tương quan tổng thể (gọi đó là lược đồ ACF). Bartlett's đã chỉ ra rằng nếu một chuỗi là ngẫu nhiên và dừng, thì các hệ số tự tương quan ρ_k sẽ có phân phối xấp xỉ chuẩn với kỳ vọng toán bằng 0 và phương sai 1/n với n khá lớn, ρ_k ~ N(0, 1/n).

Ta cần kiểm định giả thiết:

$H_0 : \rho\_k = 0$ (chuỗi dừng)

$H_1 : \rho_k \neq 0$ (chuỗi không dừng)

Ta có: $Z = \frac{x-\mu}{\sigma} = \frac{\rho_k}{SE(\rho_k)} = \frac{\rho_k}{\frac{1}{\sqrt{n}}}$

Nếu $\rho_k \in (\frac{-Z_{a/2}}{\sqrt{n}}, \frac{Z_{a/2}}{\sqrt{n}})$ thì chấp nhận giả thiết H_0 với mức ý nghĩa α. Giá trị của các chỉ số Z tra trong bảng đã được tính toán sẵn. Với độ tin cậy 95%, khoảng tin cậy của ρ_k là $\pm\frac{1.96}{\sqrt{n}}$. Nếu $\rho_k \in \left(-\frac{1.96}{\sqrt{n}}, +\frac{1.96}{\sqrt{n}}\right)$ ta chấp nhận giả thiết H_0, tức chuỗi đang xét là một chuỗi dừng, ngược lại nếu ρ_k không thuộc khoảng này, ta bác bỏ H_0 (với mức ý nghĩa 5%).

***Tự tương quan riêng phần (PACF)**

Các hệ số tự tương quan ρ_k (k 2) phản ánh mức độ kết hợp tuyến tính của y_t và y_{t+k}. Tuy nhiên, mức độ kết hợp giữa hai biến còn có thể do một số biến khác gây ra. Trong trường hợp này là ảnh hưởng từ các biến $y_{t-1}, ..., y_{t-k+1}$. Do đó để đo độ kết hợp riêng rẽ giữa y_t và y_{t-k} ta sử dụng hàm tương quan riêng PACF với hệ số tương quan riêng ρ_{kk} được ước lượng theo công thức đệ quy của Durbin:

$$PACF(k) = \rho_{kk} = \frac{\rho_k - \sum_{j=1}^{k-1} r_{k-1,j}\rho_{k-j}}{1 - \sum_{j=1}^{k-1} r_{k-j,j}\rho_j}$$

Hệ số tương quan riêng phần đo lường mối quan hệ giữa hai biến khi tất cả những biến khác giữ nguyên không đổi. Nếu chuỗi dừng thì các ρ_{kk} cũng có phân phối chuẩn N(0,1/n), do đó kiểm định giả thiết đối với ρ_{kk} tương tự như đối với ρ_k

***Kiểm định đồng thời** + Box-Pierce đã đưa ra kiểm định về sự đồng thời bằng không của các hệ số tương quan: $H_0 : \rho_1 = \rho_2 = ... = \rho_k = 0$ (chuỗi đang xét là dừng)

H_1:tồn tại ít nhất một $\rho_k \neq 0$ (chuỗi đang xét là chưa dừng)

Giả thiết H_0 được kiểm định bằng thống kê Q:

$$Q = n\sum_{k=1}^{m} \rho_k^2$$

Với n: kích thước mẫu, m: độ dài của trễ; ta có $Q \sim x^2(m)$. Nếu với một mức ý nghĩa xác định thì ta bác bỏ giả thiết H_0, tức ít nhất phải có một giá trị $\rho_k \neq 0$,lúc này có thể suy đoán chuỗi đang xét chưa dừng và ngược lại.

- Một dạng khác của Q là thống kê Ljung-Box (LB): Với giả thiết H_0, H_1 tương tự như thống kê Q ở trên, ta có tiêu chuẩn kiểm định như sau:$LB = n(n+2)\sum_{k=0}^{m} \frac{\rho_k^2}{n-k}$

Với $LB \sim Z^2(m)$. Nếu $LB > Z^2(m)$ ta bác bỏ giả thiết H_0, tương tự lúc này có thể suy đoán chuỗi đang xét là chưa dừng và ngược lại.

Thống kê LB được xem là tốt hơn so với thống kê Q đối với các mẫu số nhỏ. Với Eviews, ta dễ dàng có được các giá trị của LB với các độ trễ khác nhau (cột Q- Stat) và xác suất nhỏ nhất để giả thiết H0 bị bác bỏ (cột Prob).

c. Kiểm định nghiệm đơn vị (Unit root test)

***Nhiễu trắng**

Tính dừng là một giả định yếu hơn giả định phân phối chuẩn, tuy nhiên hồi quy với chuỗi thời gian có tính dừng sẽ cho ta các thống kê đáng tin cậy, chỉ cần số quan sát tăng lên thì độ tin cậy sẽ càng lớn. Do vậy sai số u_t không nhất thiết phải tuân theo phân phối chuẩn, miễn là mẫu quan sát đủ lớn. Thay vào đó u_t được giả định là **"nhiễu trắng"**.

Giả sử có phương trình sau: $y_t = u_t$. Nếu u_t là một nhiễu trắng thì nó có trung bình bằng 0, phương sai không đổi và hiệp phương sai giữa hai u_t bằng không. Nhiễu trắng là một trường hợp đặc biệt của chuỗi dừng. Các điều kiện này hàm ý rằng chúng ta không thể dự báo được nhiễu trắng từ những giá trị trung bình trong quá khứ của chính nó. Nếu u_t còn có tự tương quan thì điều này có nghĩa là còn có những thông tin ẩn chứa trong ut mà chúng ta có thể khai thác để cải thiện mô hình hồi quy.

***Bước ngẫu nhiên**

Nếu $y_t = y_{t-1} + u_t$ với u_t là nhiễu trắng, thì y_t được gọi là bước ngẫu nhiên. Ta có: $y_1 = y_0 + u_1$

$y_2 = y_1 + u_2 = y_0 + u_1 + u_2$

$y_t = y_0 + u_1 + u_2 + ... + u_t$

Do y_0 là hằng số, các ut độc lập với nhau, phương sai không đổi và bằng σ^2 nên: $Var(y_t) = t\sigma^2$ (thay đổi theo t). Điều này chứng tỏ u_t là chuỗi không dừng và y_t được gọi là bước ngẫu nhiên.

***Kiểm định nghiệm đơn vị Dickey – Fuller (Unit root test)**

Một tiêu chuẩn khác để kiểm định tính dừng là kiểm định nghiệm đơn vị, được giới thiệu bởi Dickey (1976), Dickey & Fuller (1979); kiểm định này được sử dụng phổ biến trong nghiên cứu thay vì dùng lược đồ tương quan vì có tính học thuật và chuyên nghiệp hơn. Xét mô hình sau với u_t là nhiễu trắng: $y_t = \rho y_{t-1} + u_t$ (1.1)

Nếu $\rho = 1$ thì y_t là bước ngẫu nhiên và không dừng. Do đó để kiểm định tính dừng của y_t ta kiểm định giả thiết:

$H_0 : \rho = 1$ (y_t là chuỗi không dừng) $H_1 : \rho \neq 1$ (y_t là chuỗi dừng) Chúng ta biến đổi phương trình (1.1) thành: $y_t - y_{t-1} = \rho y_{t-1} + u_t = (\rho - 1)y_{t-1} + u_t$

$\Delta Y_{t-1} = \sigma y_{t-1} + u_t$

Như vậy, các giả thiết có thể viết lại: $H_0 : \sigma = 0$ (y_t là chuỗi không dừng) $H_1 : \sigma \neq 0$ (y_t là chuỗi dừng)

Ở đây ta không thể sử dụng kiểm định t vì y_t có thể là chuỗi không dừng, hay Dickey và Fuller cho rằng các giá trị t của hệ số y_{t-1} sẽ không tuân theo xác suất Student mà theo xác suất τ (tau statistic), kiểm định thống kê còn được gọi là kiểm định Dickey – Fuller. Ta sử dụng tiêu chuẩn kiểm định như sau:

$$\tau = \frac{\rho}{se(\rho)}$$

Có phân phối theo quy luật DF. Nếu $\tau > \tau_\alpha$ ta bác bỏ giả thiết H_0 và kết luận chuỗi y_t là một chuỗi dừng và ngược lại.Tiêu chuẩn DF cũng được áp dụng cho các mô hình sau: $\Delta y_t = \delta y_{t-1} + u_t$ (1) $\Delta y_t = \beta y_{t-1} + \delta y_{t-1} + u_t$ (2) $\Delta y_t = \beta y_{t-1} + \beta y_{t-2} + \delta y_{t-1} + u_t$ (3)

Với giả thiết $H_0 : \gamma_0$ (chuỗi dừng). Nếu u_t có tự tương quan nghĩa là Δy_t phụ thuộc vào các Δy_{t-i} trong quá khứ như $\Delta y_{t-1}, \Delta y_{t-2}, ...$ thì ta cải biến mô hình (3) thành mô hình:

$$\Delta y_t = \beta_1 + \beta_2 + \delta y_{t-1} + \sum_{i=0}^{m} \alpha_i \Delta y_{t-i} + u_t$$

Lúc này kiểm định DF như phương trình trên được gọi là kiểm định DF mở rộng (ADF – Augmented Dickey – Fuller Test), áp dụng cho chuỗi thời gian có bậc tự hồi quy cao hơn và quy trình tiến hành việc kiểm định là hoàn toàn tương tự.

###Biến đổi chuỗi dừng thành chuỗi không dừng

Nếu một chuỗi thời gian không có yếu tố dừng, chúng ta phải biến đổi nó thành dừng trước khi xây dựng mô hình ARIMA, phương pháp là lấy sai phân cấp d với d=1 hoặc d=2,… Cụ thể xét bước ngẫu nhiên: $y_t = y_{t-1} + u_t$ với u_t là nhiễu trắng.Ta lấy sai phân cấp 1 của $y_t : D(y_t) = y_t - y_{t-1} = u_t$. Trong trường hợp này $D(y_t)$ là chuỗi dừng vì u_t là nhiễu trắng.

Trường hợp tổng quát, với mọi chuỗi thời gian nếu sai phân cấp 1 của Y_t chưa dừng ta tiếp tục lấy sai phân cấp 2, 3… Các nghiên cứu đã chứng minh luôn tồn tại một giá trị d xác định để sai phân cấp d của Y_t là chuỗi dừng. Khi đó Y_t được gọi là liên kết bậc d, ký hiệu là I(d).

Sai phân cấp d được lấy như sau:

- Sai phân cấp 1: $D(y_t) = y_t - y_{t-1}$
- Sai phân cấp 2: $D(D(y_t) = D^2(y_t) = (y_t - y_{t-1}) - (y_{t-1})$ ….
- Sai phân cấp d: $D(D^{d-1}(y_t))$

Nếu y_t ở dạng lôgarít thì $D(y_t)$ sẽ phản ánh phần trăm thay đổi của y_t so với thời kì trước đó, lúc này nếu y_t là giá chứng khoán thì $D(y_t)$ chính là tỷ suất sinh lợi tại thời điểm t. [3]

2.3 Quá trình hồi quy(AR), trung bình trượt(MA) và các mô hình tự hồi quy tích hợp trung bình trượt(ARMA)

2.3.1 Quá trình tự hồi quy (AR)

• Ý tưởng mô hình AR(p), tức Auto regression là hồi quy số liệu của nó trong quá khứ ở những chu kì trước. Đây là thành phần tự hồi qui bao gồm tợp hợp các độ trễ của biến hiện tại. Độ trễ bậc p chính là giá trị lùi về quá khứ p bước thời gian của chuỗi. Độ trễ dài hoặc ngắn trong quá trình AR phụ thuộc vào tham số trễ p. Cụ thể, quá trình của chuỗi được biểu diễn như bên dưới:

$$Y_t = a_0 + a_1 Y_{(t-1)} + a_2 Y_{(t-2)} + ... + a_p Y_{(t-p)} + u_t \tag{1}$$

Trong đó:

- Y_t: Quan sát dừng hiện tại;
- $Y_{(t-1)}$: quan sát dừng ở thời điểm trong quá khứ;
- a_0, a_1, a_2: các tham số phân tích hồi quy;
- u_t:sai số dự báo ngẫu nhiên của giai đoạn hiện tại, với u_t giá trị trung bình được mong đợi bằng 0.sai số dự báo ngẫu nhiên của giai đoạn hiện tại, với ut giá trị trung bình được mong đợi bằng 0.

Hàm tuyến tính Y_t là của chuỗi quan sát dừng những thời điểm trong quá khứ:$Y_{(t-1)}$, $Y_{(t-2)}$,…

Khi phân tích hồi quy Y_t theo các giá trị trong chuỗi thời gian, chuỗi dừng có độ trễ, chúng ta sẽ được mô hình AR. Số quan sát dừng ở các thời điểm quá khứ được sử dụng trong mô hình tự hồi quy là bậc p của mô hình AR. Nếu sử dụng 2 quan sát dừng ở quá khứ, ta có mô hình tương quan bậc hai AR(2).

- Mô hình AR(1):$Y_t = a_0 + a_1 Y_{(t-1)} + u_t$
- Mô hình AR(2): $Y_t = a_0 + a_1 Y_{(t-1)} + a_2 Y_{(t-2)} + u_t$

2.3.2 Mô hình trung bình trượt MA (MA)

- Về mặt ý tưởng thì đó chính là quá trình hồi qui tuyến tính của giá trị hiện tại theo các giá trị hiện tại và quá khứ của sai số nhiễu trắng (white noise error term) đại diện cho các yếu tố shock ngẫu nhiên, những sự thay đổi không lường trước và giải thích bởi mô hình.Qúa trình trung bình trượt được hiểu là quá trình dịch chuyển hoặc thay đổi giá trị trung bình của chuỗi theo thời gian. Do chuỗi của chúng ta được giả định là dừng nên quá trình thay đổi trung bình dường như là một chuỗi nhiễu trắng. Qúa trình moving average sẽ tìm mối liên hệ về mặt tuyến tính giữa các phần tử ngẫu nhiên.
- Hàm tuyến tính Y_t phụ thuộc vào các biến sai số dự báo quá khứ và hiện tại. Mô hình trung bình trượt là một trung bình trọng số của những sai số mới nhất.

$$Y_t = b_0 + u_t + b_1 u_{(t-1)} + b_2 u_{(t-2)} + ... + \ b_q u_{(t-q)} \tag{2}$$

Trong đó: - Y_t: Quan sát dừng hiện tại;

- u_t: sai số dự báo;
- $u_{(t-1)}$,$u_{(t-1)}$,...: sai số dự báo quá khứ;
- b_0, b_1, b_2: giá trị trung bình của Y_t và các hệ số bình quân di động;
- q: bậc của MA.;

Mô hình AR(1):$Y_t = b_0 + u_t + b_1 u_{(t-1)}$

Mô hình AR(2):$Y_t = b_0 + u_t + b_1 u_{(t-1)} + b_2 u_{(t-2)}$

2.3.3 Quá trình tự hồi quy tích hợp trung bình trượt (ARMA)

Để biểu diễn sơ đồ Y không chỉ riêng AR hoặc MA mà có thể kết hợp cả hai, sự kết hợp ta được mô hình ARMA, còn gọi là mô hình trung bình trượt tự hồi quy.

Y_t là quá trình $ARMA(1,\ 1)$ nếu Y có thể biểu diễn dưới dạng sau với(u là nhiễu trắng):

$$Y_t = a_0 + a_1 Y_{(t-1)} + u_t + b_0 + b_1 u_{(t-1)} \tag{3}$$

Viết lên một cách tổng quát hơn, Y_t là quá trình $ARMA(p,\ q)$ nếu Y có thể biểu diễn dưới dạng sau với(u là nhiễu trắng):

$$Y_t = a_0 + a_1 Y_{(t-1)} + a_2 Y_{(t-2)} + ... + a_p Y_{(t-p)} + u_t + b_0 + b_1 u_{(t-1)} + b_2 u_{(t-2)} + ... + b_q u_{(t-q)} \tag{4}$$

2.3.4 Quá trình tự hồi quy, đồng liên kết, trung bình trượt (ARIMA))

Intergrated: Là quá trình đồng tích hợp hoặc lấy sai phân. Yêu cầu chung của các thuật toán trong time series là chuỗi phải đảm bảo tính dừng. Hầu hết các chuỗi đều tăng hoặc giảm theo thời gian. Do đó yếu tố tương quan giữa chúng chưa chắc là thực sự mà là do chúng cùng tương quan theo thời gian. Khi biến đổi sang chuỗi dừng, các nhân tố ảnh hưởng thời gian được loại bỏ và chuỗi sẽ dễ dự báo hơn. Để tạo thành chuỗi dừng, một phương pháp đơn giản nhất là chúng ta sẽ lấy sai phân. Một số chuỗi tài chính còn qui đổi sang logarit hoặc lợi suất. Bậc của sai phân để tạo thành chuỗi dừng còn gọi là bậc của quá trình đồng tích hợp (order of intergration).

Như vậy về tổng quát thì ARIMA là mô hình kết hợp của 2 quá trình tự hồi qui và trung bình trượt. Dữ liệu trong quá khứ sẽ được sử dụng để dự báo dữ liệu trong tương lai. Trước khi huấn luyện mô hình, cần chuyển hóa chuỗi sang chuỗi dừng bằng cách lấy sai phân bậc 1 hoặc logarit. Ngoài ra mô hình cũng cần tuân thủ điều kiện ngặt về sai số không có hiện tượng tự tương quan và phần dư là nhiễu trắng. Đó là lý thuyết của kinh tế lượng. Còn trong học máy thì chỉ cần quan tâm đến làm sao để lựa chọn một mô hình có sai số dự báo là nhỏ nhất.

Một chuỗi thời gian có thể tuân theo nhiều mô hình khác nhau. Tuy nhiên, cả ba mô hình AR, MA, ARMA đều yêu cầu chuỗi phải có tính dừng. Nhưng thực tế có nhiều chuỗi thời gian không có tính dừng. Vậy làm thế nào để áp dụng được các mô hình trong thực tế ? Câu trả lời ở đây là sử dụng phương pháp lấy sai phân biến đổi một chuỗi không dừng thành chuỗi dừng, trước khi sử dụng mô hình ARMA.

Nếu chuỗi Y_t có đồng liên kết bậc d trên mô hình $ARMA(p,\ q)$ cho chuỗi sai phân bậc d, thì chúng ta có mô hình $ARMA\,(p,\ d,\ q)$. Với bậc tự hồi quy p, số lần lấy sai phân d để chuỗi Y_t được xác định là chuỗi dừng, bậc trung bình trượt q (p và q là bậc tương ứng của chuỗi dừng).

- Trong mô hình $ARMA\,(p,\ d,\ q)$, khi d=0 và q=0 thì ta có AR(p)

- Trong mô hình $ARMA\,(p,\ d,\ q)$, khi d=0 và p=0 thì ta có AR(q)

Với $ARMA\,(1,\ 1,\ 1)$ nghĩa là Y_t có sai phân bậc 1 là một chuỗi dừng. Chuỗi sai phân dừng này có thể biểu diễn dưới dạng $ARMA(1,\ 1)$ với u là nhiễu trắng.

$$\Delta Y_t = a_0 + a_1 Y_{(t-1)} + a_0 u_t + a_1 u_{(t-1)} \tag{5}$$

Như vậy, xác định được các giá trị p, d, q ta sẽ mô hình hóa được chuỗi ARIMA.

Ta thấy, mô hình ARIMA chỉ sử dụng các giá trị trong quá khứ của chuỗi chứ không dùng thêm biến độc lập khác. [1]

3 ÁP DỤNG MÔ HÌNH ARIMA TRONG DỰ ĐOÁN DỮ LIỆU COVID-19

3.1 Tìm hiểu về dữ liệu COVID-19

3.1.1 Thông tin về dữ liệu

Dữ liệu được nhóm lấy ở trang:

https://api.covid19india.org

Đây là dữ liệu COVID-19 được ghi nhận tại Ấn độ, dữ liệu được ghi nhận hàng ngày đến 30-10-2021. Từ sau đó dữ liệu đã được dữ án ngừng theo dõi. Nhóm đã thực hiện việc áp dụng mô hình ARIMA trong dự đoán tổng số ca nhiễm từng ngày dựa trên dữ liệu được ghi nhận

Nhóm thực hiện công việc import các thư viện cần thiết để xử lý dữ liệu

```
import warnings
#warnings.filterwarnings('ignore')
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
#%matplotlib inline
from scipy.stats import *
import matplotlib.pyplot as plt  # Matlab-style plotting
import seaborn as sns # informative statistical graphics.
import statsmodels.api as sm #for ARIMA and SARIMAX
import datetime
from datetime import timedelta
```

Số liệu về COVID sẽ được mô tả thông qua dòng lệnh dưới đây:

```
#importing dataset for API
train = pd.read_csv('https://api.covid19india.org/csv/latest/state_wise_daily.csv')
train['Date'] = pd.to_datetime(train['Date'], format="%d-%b-%y")
train.tail()
```

```
##            Date    Date_YMD     Status     TT  AN  ...  TR  UP  UT   WB  UN
## 1786 2021-10-30  2021-10-30  Recovered  14672   1  ...   8   9   6  880   0
## 1787 2021-10-30  2021-10-30   Deceased    445   0  ...   0   0   0   13   0
## 1788 2021-10-31  2021-10-31  Confirmed  12907   0  ...  12   6   5  914   0
## 1789 2021-10-31  2021-10-31  Recovered  13152   0  ...   2   6   9  913   0
## 1790 2021-10-31  2021-10-31   Deceased    251   0  ...   0   0   0   15   0
##
## [5 rows x 42 columns]
```

Định dạng dữ liệu input là dữ liệu wide

Vậy chúng ta có thể thấy các thuộc tính như sau:

- Date: định dạng thời gian (Năm-tháng-ngày)
- Status: Thuộc tính dữ liệu với hàng ngang tương ứng
- TT: Tổng số ca nhiễm
- Các columns đổ sang bên phải: là số liệu từng địa phương của Ấn độ

3.1.2 Tiền xử lý dữ liệu COVID-19

Vậy, dựa vào các thuộc tình trên, chúng ta bỏ đi các status của "Tử vong và hồi phục". Chúng ta chỉ quan tâm đến tổng số ca nhiễm trong ngày trên phạm vi toàn quốc của Ấn Độ. Ta thu được kết quả là bảng dữ liệu chứa tổng số ca nhiễm của Ấn Độ theo ngày

```
cols = ['AN','AP',  'AR',   'AS',   'BR',   'CH',   'CT',   'DD',   'DL',   'DN',   'GA',   'GJ',   'HP
train.drop(cols, axis=1, inplace=True)
train = train.set_index('Status')
train.drop(['Recovered','Deceased'],inplace=True)
train = train.reset_index()
train.drop(["Status"],axis=1,inplace= True)
train_df = train
train_df.head()
```

```
##         Date    Date_YMD  TT
## 0 2020-03-14  2020-03-14  81
## 1 2020-03-15  2020-03-15  27
## 2 2020-03-16  2020-03-16  15
## 3 2020-03-17  2020-03-17  11
## 4 2020-03-18  2020-03-18  37
```

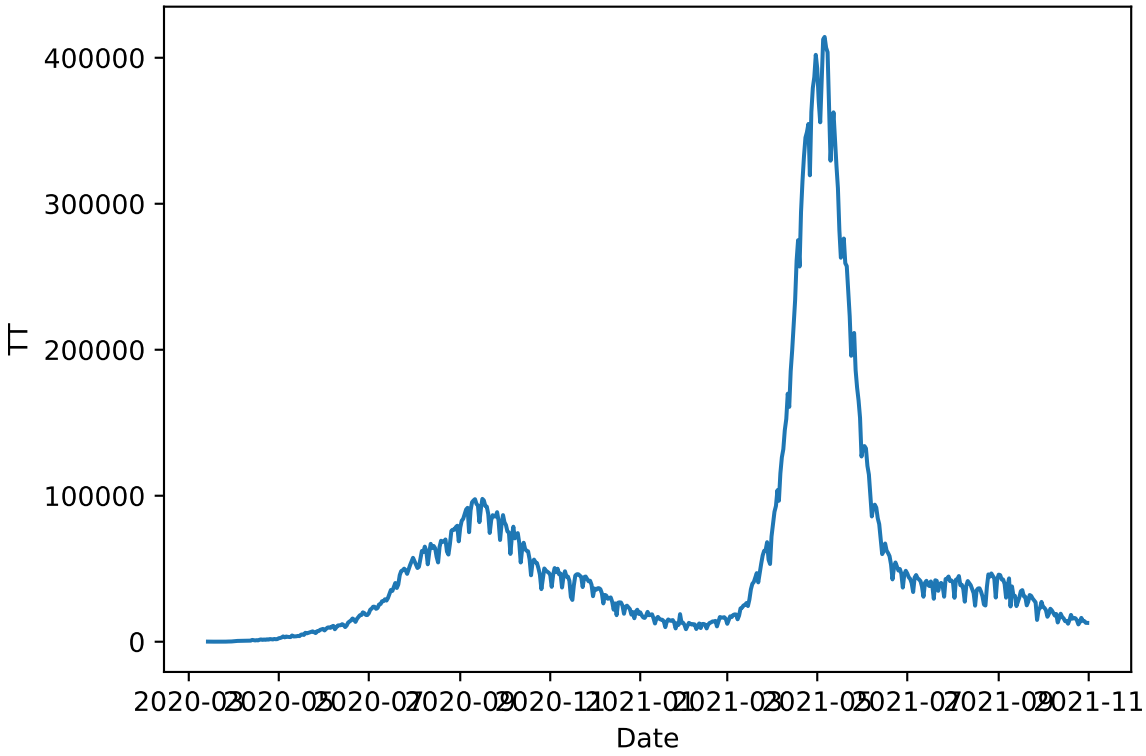
Thêm vào đó, bước tiếp theo cần chuyển Date mà đã chuyển định dạng về index và tổng số ca nhiễm (TT) về dạng float nhằm phục vụ mục đích tính toán chính xác cao hơn.

```
train_df = train_df.set_index('Date')
train_df['TT'] = train_df['TT'].astype(float)
train_df.head()
```

```
##                Date_YMD    TT
## Date
## 2020-03-14  2020-03-14  81.0
## 2020-03-15  2020-03-15  27.0
## 2020-03-16  2020-03-16  15.0
## 2020-03-17  2020-03-17  11.0
## 2020-03-18  2020-03-18  37.0
```

Trực quan dữ liệu hiện có, ta có biểu đò

```
sns.lineplot(x="Date", y="TT",legend = 'full' , data=train_df)
```



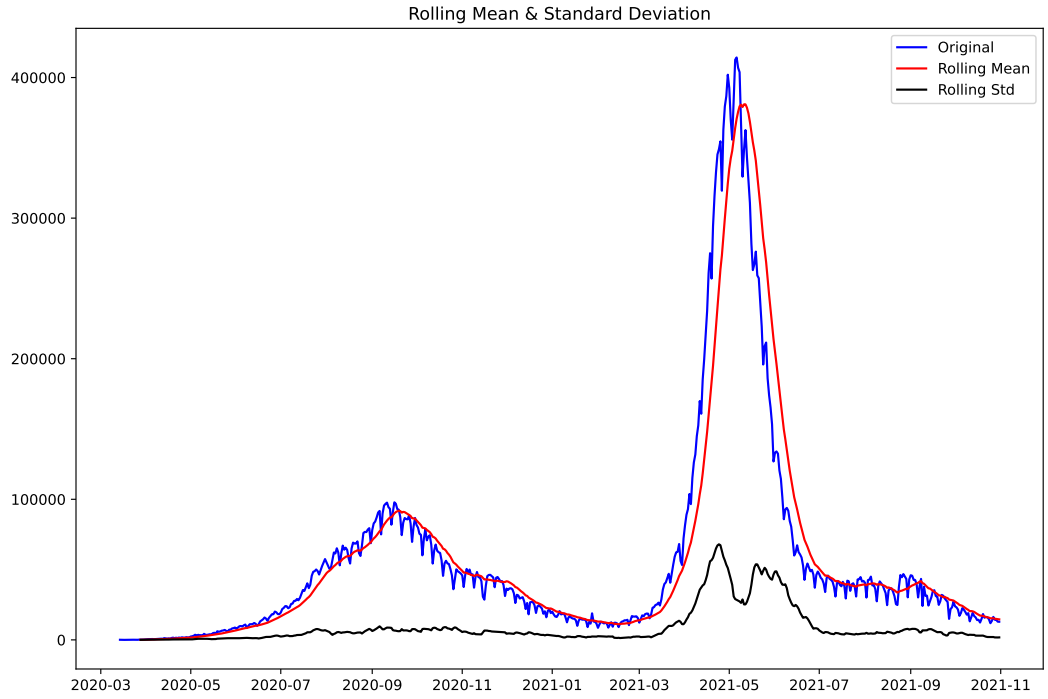
3.2 Xây dựng mô hình ARIMA trên tập dữ liệu

Xây dựng các trung bình trượt và sai số trượt cho tập dữ liệu bằng các hàm test_stationarity, và thu được kết quả như sau:

```
from statsmodels.tsa.stattools import adfuller #adfuller stands for Augmented Dickey-Fuller unit root t

#The function find mean and standard deviation of the series and and performs augmented dickey fuller t
#returns pvale .. The samaller the pvalue more stationary is the series.

def test_stationarity(timeseries, window = 15, cutoff = 0.01):
  rolmean = timeseries.rolling(window).mean()
  rolstd = timeseries.rolling(window).std()
  fig = plt.figure(figsize=(12, 8))
  orig = plt.plot(timeseries, color='blue',label='Original')
  mean = plt.plot(rolmean, color='red', label='Rolling Mean')
  std = plt.plot(rolstd, color='black', label = 'Rolling Std')
  plt.legend(loc='best')
  plt.title('Rolling Mean & Standard Deviation')
  plt.show()

  print('Results of Dickey-Fuller Test:')
  dftest = adfuller(timeseries, autolag='AIC',)
  dfoutput = pd.Series(dftest[0:4], index=['Test Statistic','p-value','#Lags Used','Number of Observati
  for key,value in dftest[4].items():
      dfoutput['Critical Value (%s)'%key] = value
```

```
  pvalue = dftest[1]
  if pvalue < cutoff:
      print('p-value = %.4f. The series is likely stationary.' % pvalue)
  else:
      print('p-value = %.4f. The series is likely non-stationary.' % pvalue)

  print(dfoutput)

test_stationarity(train_df['TT'])
```

```
## Results of Dickey-Fuller Test:
## p-value = 0.0201. The series is likely non-stationary.
## Test Statistic                  -3.198105
## p-value                          0.020095
## #Lags Used                      19.000000
## Number of Observations Used    577.000000
## Critical Value (1%)             -3.441734
## Critical Value (5%)             -2.866562
## Critical Value (10%)            -2.569445
## dtype: float64
```



Vậy dựa vào kết quả trên, ta thấy rằng p-value rất thấp (p-value = 0.02 < 0.05) nên có thể chứng minh đây là chuỗi dừng.

Nhóm đưa ra thông tin chi tiết của mô hình như sau:

```
sarimax_mod = sm.tsa.statespace.SARIMAX(train_df.TT, trend='n', order=(14,1,0)).fit()
```

```
## E:\anaconda3\envs\NLP\lib\site-packages\statsmodels\tsa\base\tsa_model.py:471: ValueWarning: No frequ
##   self._init_dates(dates, freq)
## E:\anaconda3\envs\NLP\lib\site-packages\statsmodels\tsa\base\tsa_model.py:471: ValueWarning: No frequ
##   self._init_dates(dates, freq)
```

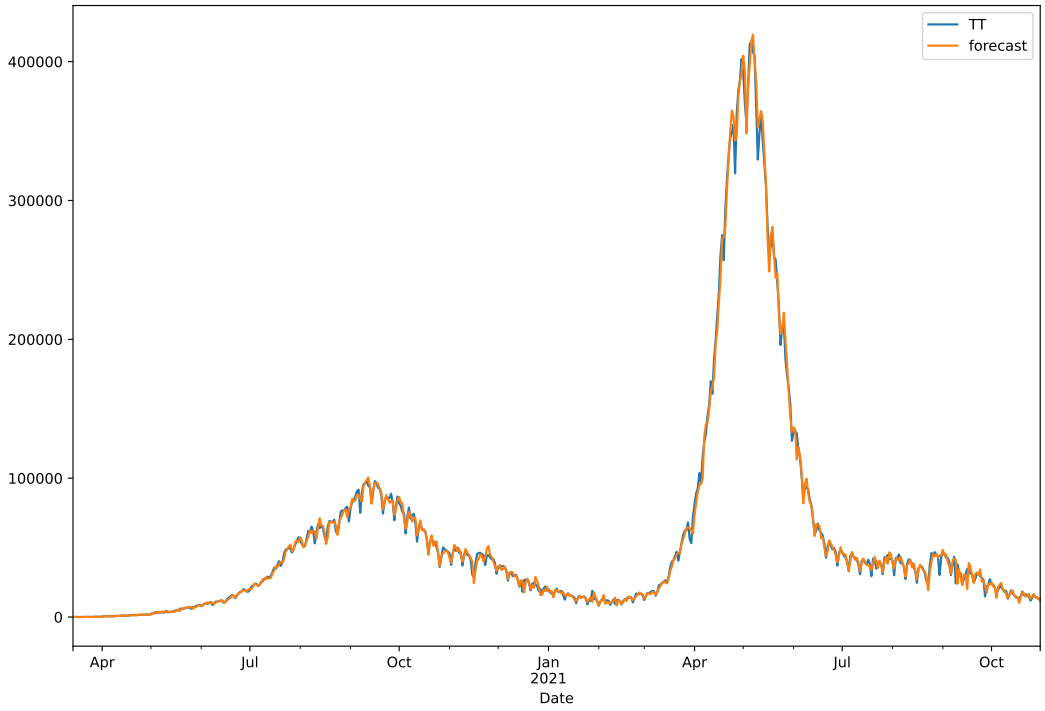
```
print(sarimax_mod.summary())
```

```
##                                SARIMAX Results
## ==============================================================================
## Dep. Variable:                      TT   No. Observations:                  597
## Model:               SARIMAX(14, 1, 0)   Log Likelihood               -5840.113
## Date:                Tue, 20 Dec 2022   AIC                          11710.226
## Time:                        03:58:32   BIC                          11776.079
## Sample:                  03-14-2020   HQIC                         11735.869
##                        - 10-31-2021
## Covariance Type:                  opg
## ==============================================================================
##                  coef    std err          z      P>|z|      [0.025      0.975]
## ------------------------------------------------------------------------------
## ar.L1         -0.0979      0.027     -3.625      0.000      -0.151      -0.045
## ar.L2          0.0975      0.027      3.675      0.000       0.045       0.149
## ar.L3          0.1132      0.024      4.741      0.000       0.066       0.160
## ar.L4          0.0839      0.028      2.959      0.003       0.028       0.139
## ar.L5          0.1381      0.026      5.284      0.000       0.087       0.189
## ar.L6          0.2109      0.021      9.903      0.000       0.169       0.253
## ar.L7          0.6451      0.022     28.725      0.000       0.601       0.689
## ar.L8          0.0888      0.028      3.190      0.001       0.034       0.143
## ar.L9         -0.1538      0.022     -6.987      0.000      -0.197      -0.111
## ar.L10        -0.1248      0.025     -4.941      0.000      -0.174      -0.075
## ar.L11        -0.0698      0.032     -2.207      0.027      -0.132      -0.008
## ar.L12        -0.1642      0.028     -5.901      0.000      -0.219      -0.110
## ar.L13        -0.1322      0.024     -5.417      0.000      -0.180      -0.084
## ar.L14         0.1628      0.025      6.436      0.000       0.113       0.212
## sigma2      1.923e+07   5.72e-10   3.36e+16      0.000    1.92e+07    1.92e+07
## ==============================================================================
## Ljung-Box (L1) (Q):                   0.07   Jarque-Bera (JB):              1502.95
## Prob(Q):                              0.79   Prob(JB):                         0.00
## Heteroskedasticity (H):               8.96   Skew:                            -0.63
## Prob(H) (two-sided):                  0.00   Kurtosis:                        10.68
## ==============================================================================
##
## Warnings:
## [1] Covariance matrix calculated using the outer product of gradients (complex-step).
## [2] Covariance matrix is singular or near-singular, with condition number 3.23e+31. Standard errors m
```

Thực hiện công việc dự đoán dựa trên thời gian đã có trong dữ liệu:

```
#Now lets predict using out model.
today = datetime.date.today() -timedelta(days=1)

start_index = '14-Mar-20'
end_index = today.strftime("%Y-%m-%d")

#adding forecasted values and plotting
train_df['forecast'] = sarimax_mod.predict(start= start_index,end = end_index,dynamic= False,)
```

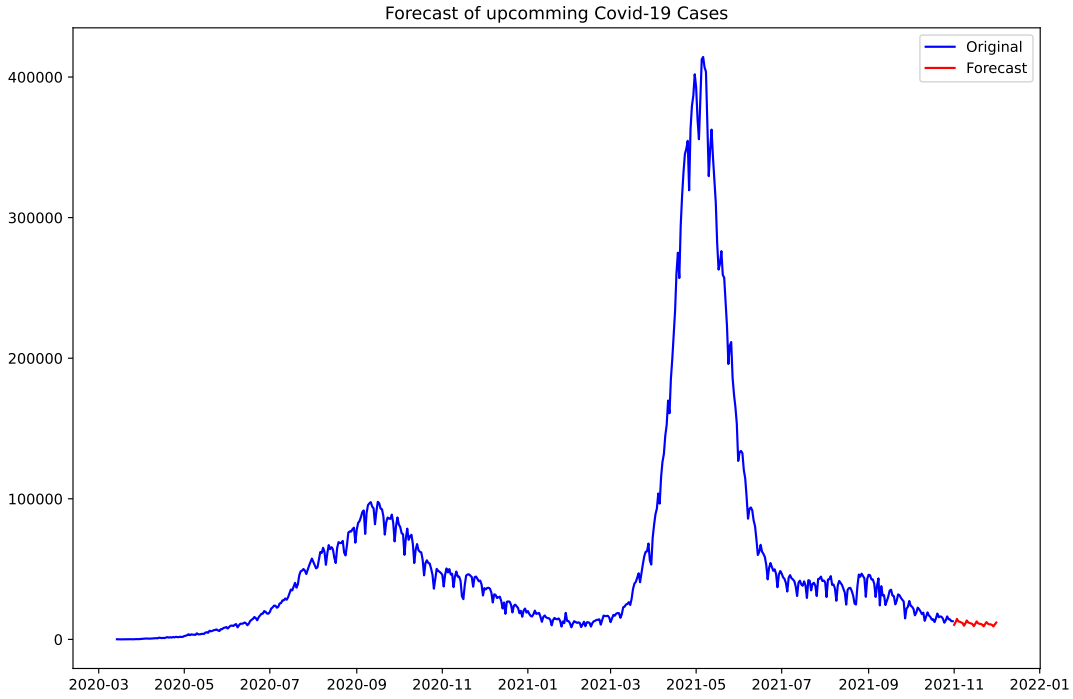
```
train_df[start_index:][['TT', 'forecast']].plot(figsize=(12, 8))
```



Có thể thấy rằng dữ liệu khá là khớp so với dữ liệu hiện hiện có

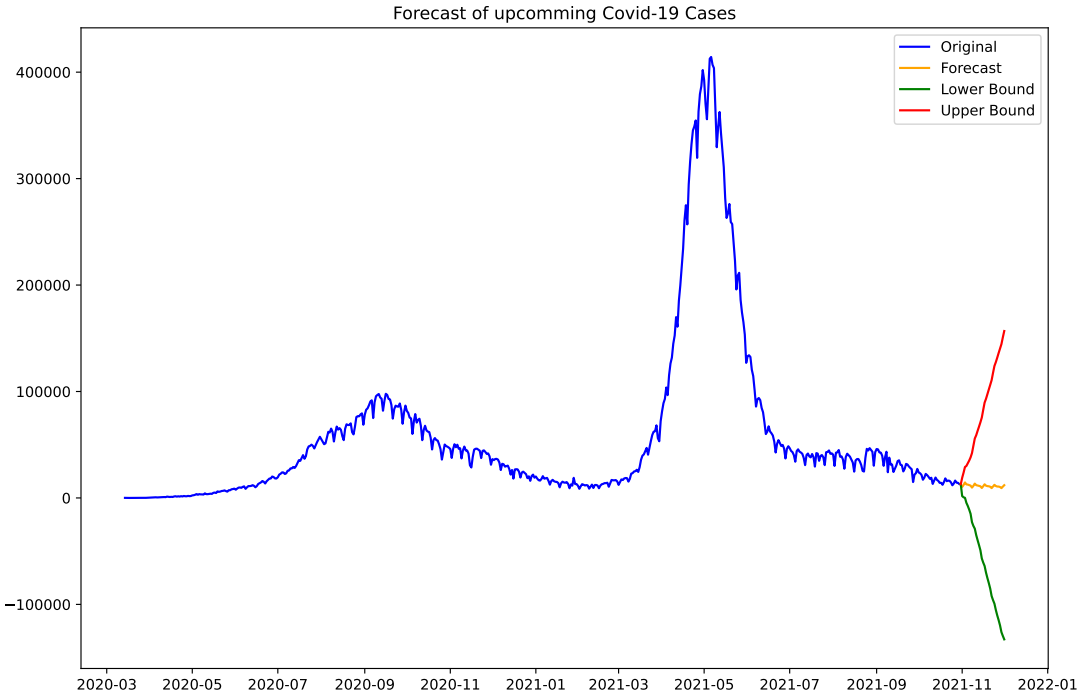
Tiếp theo, để thử nhiệm mô hình ARIMA trong dự đoán số ca COVID-19 trong tương lai, nhóm đã thử nghiệm với thời gian 30 ngày kể từ lần cuối dữ liệu được ghi nhận là 30-10-2021, ta thu được kết quả như hình sau

```
future_predict = sarimax_mod.predict(start= datetime.date(2021,11,1) ,end = datetime.date(2021,11,1)+ti
figg = plt.figure(figsize=(12, 8))
orig = plt.plot(train_df['TT'], color='blue',label='Original')
fore = plt.plot(future_predict, color='red', label='Forecast')
plt.legend(loc='best')
plt.title('Forecast of upcomming Covid-19 Cases')
plt.show()
```



Cuối cùng, nhóm đưa ra số liệu về độ tin cậy của dự đoán như sau. Đối với khoảng tin cậy trái và phải như sau:

```
f_temp = pd.DataFrame()
f_temp['date'] = future_predict.index
f_temp['values'] = future_predict.values
f_temp.loc[-1] = [train_df.index[-1],train_df['TT'][-1]]
f_temp.index = f_temp.index + 1
f_temp = f_temp.sort_index()
f_temp['date'] = pd.to_datetime(f_temp['date'], format="%d-%b-%y")
f_temp = f_temp.set_index('date')
fcast = sarimax_mod.get_forecast(datetime.date(2021,11,1)+timedelta(days=30))
fcast = fcast.conf_int()
fcast = fcast.reset_index()
fcast.loc[-1] = [train_df.index[-1],train_df['TT'][-1],train_df['TT'][-1]]
fcast.index = fcast.index + 1
fcast = fcast.sort_index()
fcast['index'] = pd.to_datetime(fcast['index'], format="%d-%b-%y")
fcast = fcast.set_index('index')
figg = plt.figure(figsize=(12, 8))
orig = plt.plot(train_df['TT'], color='blue',label='Original')
fore = plt.plot(f_temp['values'], color='orange', label='Forecast')
lower = plt.plot(fcast['lower TT'], color='green', label='Lower Bound')
upper = plt.plot(fcast['upper TT'], color='red', label='Upper Bound')
plt.legend(loc='best')
plt.title('Forecast of upcomming Covid-19 Cases')
plt.show()
```



[1] Đ. N. H. My, *Ứng dụng mô hình ARIMA trong dự báo chỉ số VN-INDEX*. Đại học kinh tế Huế, 2015.

[2] nathan, "Giải thuật time-series forecasting." magestore, 2020. Available: https://insights.magestore.com/posts/giai-thuat-time-series-forecasting

[3] P. thanh Bình, "Chuỗi dừng và chuỗi không dừng," *Econometrics by example*, vol. 13, 2018.