Kiểm tra giữa kỳ

Đề 3: Phân tích Sarima, Arimax

Họ và tên: Hoàng Hiểu Nhi – Lớp: 63TTNT

Mã sinh viên: 2154020987

Github: https://github.com/joyhh29/KT_ts.git

I, Phân tích mô hình Sarima, Arimax:

- Phân tích chuỗi thời gian (Time series analysis) là một phương pháp phân tích một loạt các điểm dữ liệu được thu thập trong một khoảng thời gian. Phân tích chuỗi thời gian là một lĩnh vực quan trọng trong nhiều lĩnh vực, bao gồm kinh tế, tài chính, khoa học, môi trường, …Hai mô hình được sử dụng để phân tích chuỗi thời gian ở đây là mô hình Sarima và mô hình Arimax.

Phần 1: Phân tích bằng mô hình Sarima.

- Sarima là viết tắt của Seasonal Autoregressive Integrated Moving Average. Đây là một mở rộng của mô hình ARIMA được bổ xung thêm tính mùa vụ từ đó giúp mô hình phù hợp hơn với các dữ liệu có tính biến động theo chu kỳ thời gian.
- Mô hình SARIMA được xác định bởi:
  + Bộ ba tham số: p, d, q.
    - p: Số hạng tự hồi quy (AR): Thể hiện số giá trị quá khứ ảnh hưởng đến giá trị hiện tại.
    - d là thứ tự tích hợp (I), số lần chuỗi thời gian cần được sai phân để trở nên dừng.
    - q: Số hạng trung bình trượt bậc tự hồi quy (MA): Thể hiện số lỗi dự báo quá khứ ảnh hưởng đến giá trị hiện tại.
  + Chu kỳ mùa vụ: S: Xác định số chu kỳ trong một năm.

+ Ngoài ra, mô hình còn có thể bao gồm các thành phần AR và MA mùa vụ (P, D, Q) để mô phỏng chính xác hơn các biến động theo chu kì.

+ Auto regression (AR):

- 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 trong chuỗi thời gian.
- Độ trễ dài hoặc ngắn trong quá trình AR phụ thuộc vào tham số trễ p
- Quá trình AR(p) của chuỗi xt được biểu diễn như bên dưới:

$$AR(p)=\emptyset_0+\emptyset_1.x_{t-1}+\emptyset_2.x_{t-2}+\ldots+\emptyset_p.x_{t-p}$$

+ Moving Average (MA):

- Moving Average (MA)
- Quá trình dịch chuyển hoặc thay đổi giá trị trung bình của chuỗi theo thời gian.
- Quá trình moving average sẽ tìm mối liên hệ về mặt tuyến tính giữa các phần tử ngẫu nhiên ϵt (stochastic term).
- Chuỗi này phải là một chuỗi nhiễu trắng thỏa mãn các tính chất:

$$\begin{cases} E(\epsilon_t) & = 0 & (1) \\ \sigma(\epsilon_t) & = \alpha & (2) \\ \rho(\epsilon_t, \epsilon_{t-s}) & = 0, \forall s <= t & (3) \end{cases}$$

- Quá trình trung bình trượt được biểu diễn theo nhiễu trắng như sau:

$$MA(q) = \mu + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

- Ứng dụng: SARIMA được ứng dụng rộng rãi trong nhiều lĩnh vực như:

+ Kinh tế: Dự báo tỷ giá hối đoái, lạm phát, GDP, ...
+ Tài chính: Dự báo giá cổ phiếu, lãi suất, ...
+ Marketing: Dự báo nhu cầu, doanh số bán hàng, ...
+ Khoa học môi trường: Dự báo lượng mưa, nhiệt độ, ...
+ Kỹ thuật: Dự báo nhu cầu năng lượng, rung động máy móc, ...

- Công thức Sarima:

$$(1-\phi_1 B)(1-\Phi_1 B_s)(1-B)(1-B_s)y_t = (1+\theta_1 B)(1+\Theta_1 B_s)\varepsilon_t$$

Trong đó,

- y_t là chuỗi thời gian quan sát tại thời điểm t,
- B là toán tử dịch chuyển ngược, đại diện cho toán tử độ trễ (tức là $By_t = y_{t-1}$),
- ϕ_1 là hệ số autoregressive không theo mùa,
- Φ_1 là hệ số autoregressive theo mùa,
- θ_1 là hệ số moving average không theo mùa,
- Θ_1 là hệ số moving average theo mùa,
- s là chu kỳ mùa,
- ε_t là thuật ngẫu nhiên trắng tại thời điểm t.

Phần 2: Phân tích bằng mô hình Arimax.

- Arimax là viết tắt của Autoregressive Integrated Moving Average with eXogenous variables
- Là một dạng mở rộng của model ARIMA. Mô hình cũng dựa trên giải định về mối quan hệ tuyến tính giữa giá trị và phương sai trong quá khứ với giá trị hiện tại và sử dụng phương trình hồi qui tuyến tính được suy ra từ mối quan hệ trong quá khứ nhằm dự báo tương lai. Nhờ đó mà cải thiện được khả năng dự báo
- Mô hình Arimax gồm có các thành phần:
  + AR: Thành phần hồi quy tự hồi quy, thể hiện mối quan hệ giữa giá trị hiện tại và một số giá trị trong quá khứ của nó
  + I : Thành phần tích phân, thể hiện được số lần lấy sai phân của chuỗi thời gian để làm cho nó trở nên tĩnh
  + MA : Thành phần trung bình động, thể hiện mô hình hóa mối quan hệ giữa giá trị hiện tại và lỗi trong dự báo giá trị trước đó
  + X : Các biến ngoại sinh là những biến số không phải là giá trị trễ của biến phụ thuộc nhưng có ảnh hưởng đến nó
- Cách xây dựng mô hình Arimax:
  + Khám phá dữ liệu
  + Kiểm tra tính dừng
  + Lấy sai phân
  + Lựa chọn các biến ngoại sinh
  + Xác định thứ tự của mô hình
  + Ước lượng các tham số

+ Kiểm định mô hình
+ Dự báo

- Ứng dụng:
  + Dự báo doanh số bán hàng
  + Dự báo giá cả
  + Dự báo nhu cầu dịch vụ
  + Phân tích thị trường chứng khoán
  + Phân tích dữ liệu khí tượng
  + Phân tích dữ liệu y tế

II,

1. Ảnh xây dựng mô hình

```
df = pd.read_csv('/kaggle/input/data-kiem-tra-2/data-kiem-tra-2.csv',encoding='latin-1', sep=',')
df
```

	date	truong_1	truong_2	truong_3	truong_4	truong_5
0	10.05.2013	4	58	3773	299.0	1
1	26.05.2013	4	58	3768	249.0	1
2	19.05.2013	4	58	4036	419.0	1
3	25.05.2013	4	58	12878	149.0	1
4	15.05.2013	4	58	12885	148.0	1
...	...	...	...	...	...	...
550033	07.11.2013	10	37	18474	199.0	1
550034	18.11.2013	10	37	18474	199.0	1
550035	24.11.2013	10	37	18484	199.0	1
550036	11.11.2013	10	37	19751	99.0	1
550037	26.11.2013	10	37	18498	199.0	1

- Đọc dữ liệu

- Biến đổi cột 'date' từ type là object sang datetime

```
df['date'] = pd.to_datetime(df['date'], format='%d.%m.%Y')
```

- Chuyển đổi cột date

- Cộng dồn những giá trị lặp theo ngày

```
df = df.groupby(['date']).sum().reset_index()
df
```

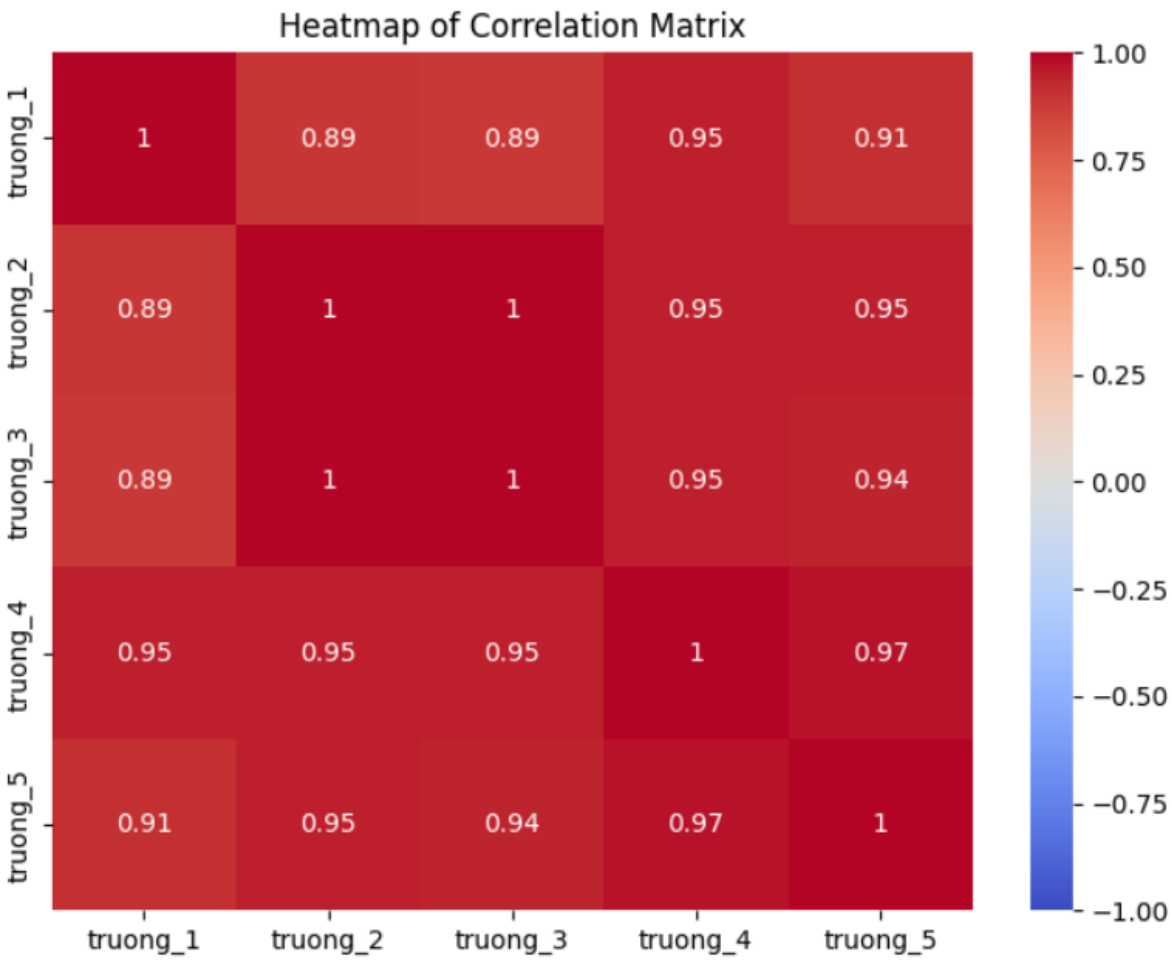
	date	truong_1	truong_2	truong_3	truong_4	truong_5
0	2013-05-01	5868	42989	15073926	8.699411e+05	1635
1	2013-05-02	5352	39975	14127032	8.025867e+05	1503
2	2013-05-03	5136	38414	13050413	8.121575e+05	1413
3	2013-05-04	4448	34386	11324760	6.420660e+05	1213
4	2013-05-05	4296	33559	11814357	5.800663e+05	1159
...	...	...	...	...	...	...
209	2013-11-26	4510	13603	5119426	3.712670e+05	548
210	2013-11-27	4050	12428	4821249	3.011002e+05	489
211	2013-11-28	4660	13724	5349662	3.666039e+05	585
212	2013-11-29	7250	21618	8082867	8.222827e+05	1499
213	2013-11-30	11350	33605	12449322	1.037198e+06	1522

- Cộng dồn những giá trị lặp theo ngày.

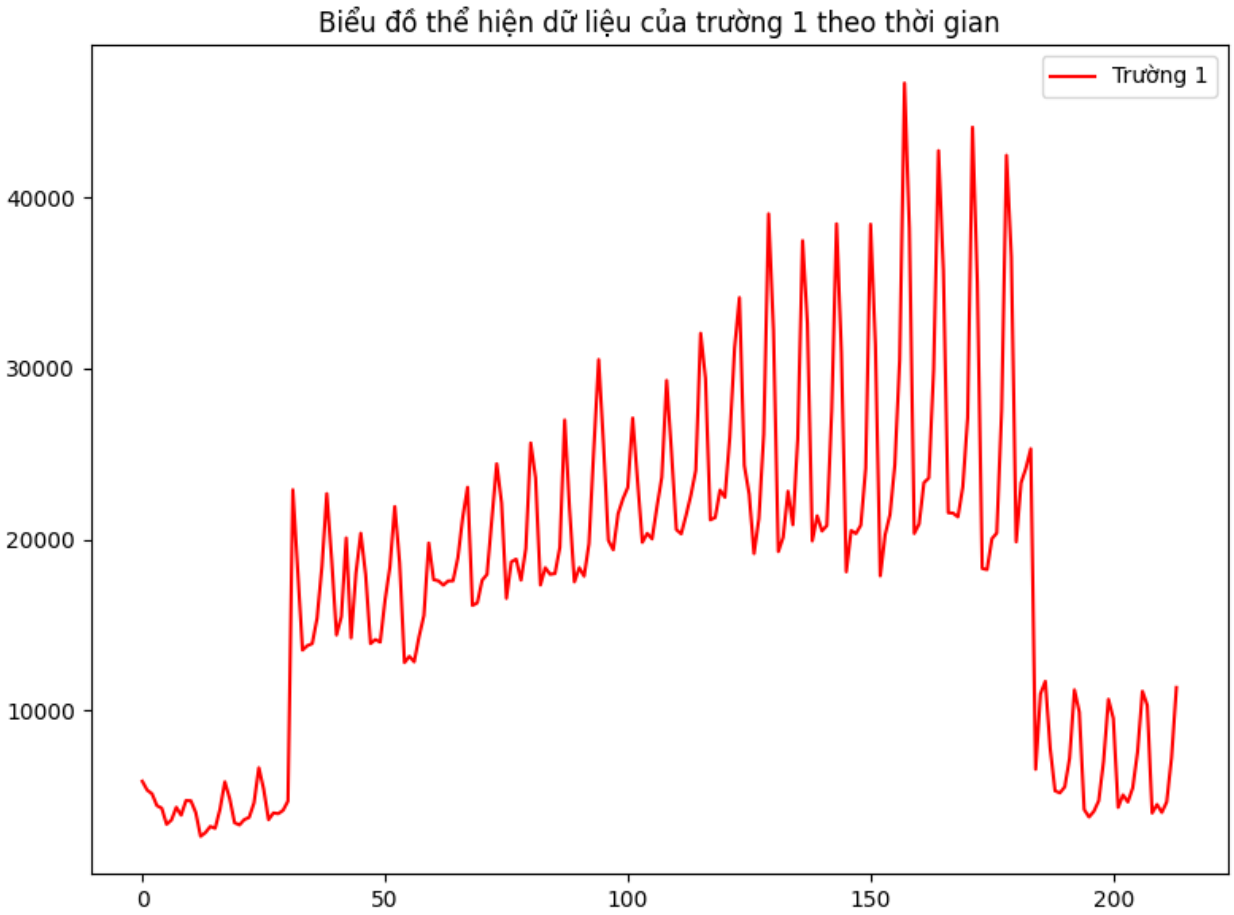
```
df = df.sort_values(by='date')
df.index = np.arange(1, len(df)+1)
df
```

	date	truong_1	truong_2	truong_3	truong_4	truong_5
1	2013-05-01	5868	42989	15073926	8.699411e+05	1635
2	2013-05-02	5352	39975	14127032	8.025867e+05	1503
3	2013-05-03	5136	38414	13050413	8.121575e+05	1413
4	2013-05-04	4448	34386	11324760	6.420660e+05	1213
5	2013-05-05	4296	33559	11814357	5.800663e+05	1159
...	...	...	...	...	...	...
210	2013-11-26	4510	13603	5119426	3.712670e+05	548
211	2013-11-27	4050	12428	4821249	3.011002e+05	489
212	2013-11-28	4660	13724	5349662	3.666039e+05	585
213	2013-11-29	7250	21618	8082867	8.222827e+05	1499
214	2013-11-30	11350	33605	12449322	1.037198e+06	1522

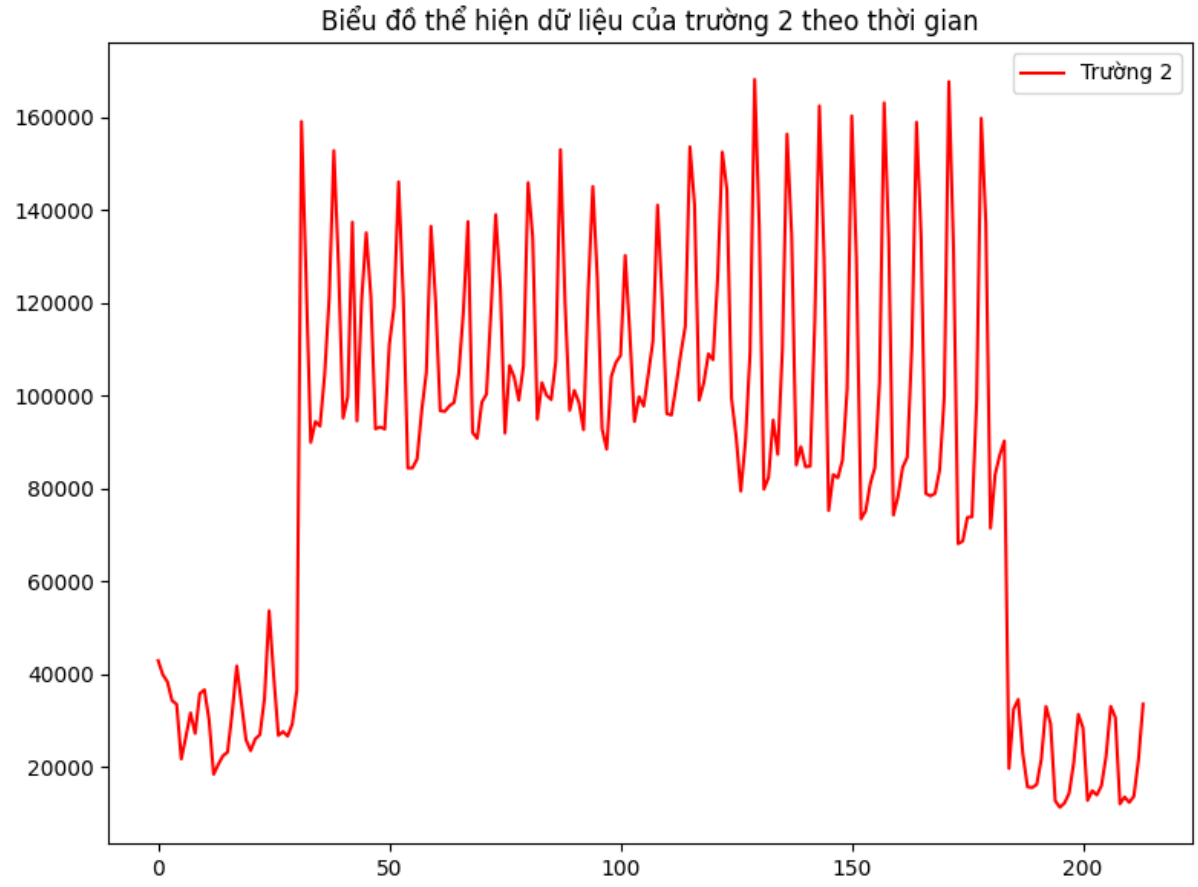
- Sắp xếp dữ liệu theo ngày.



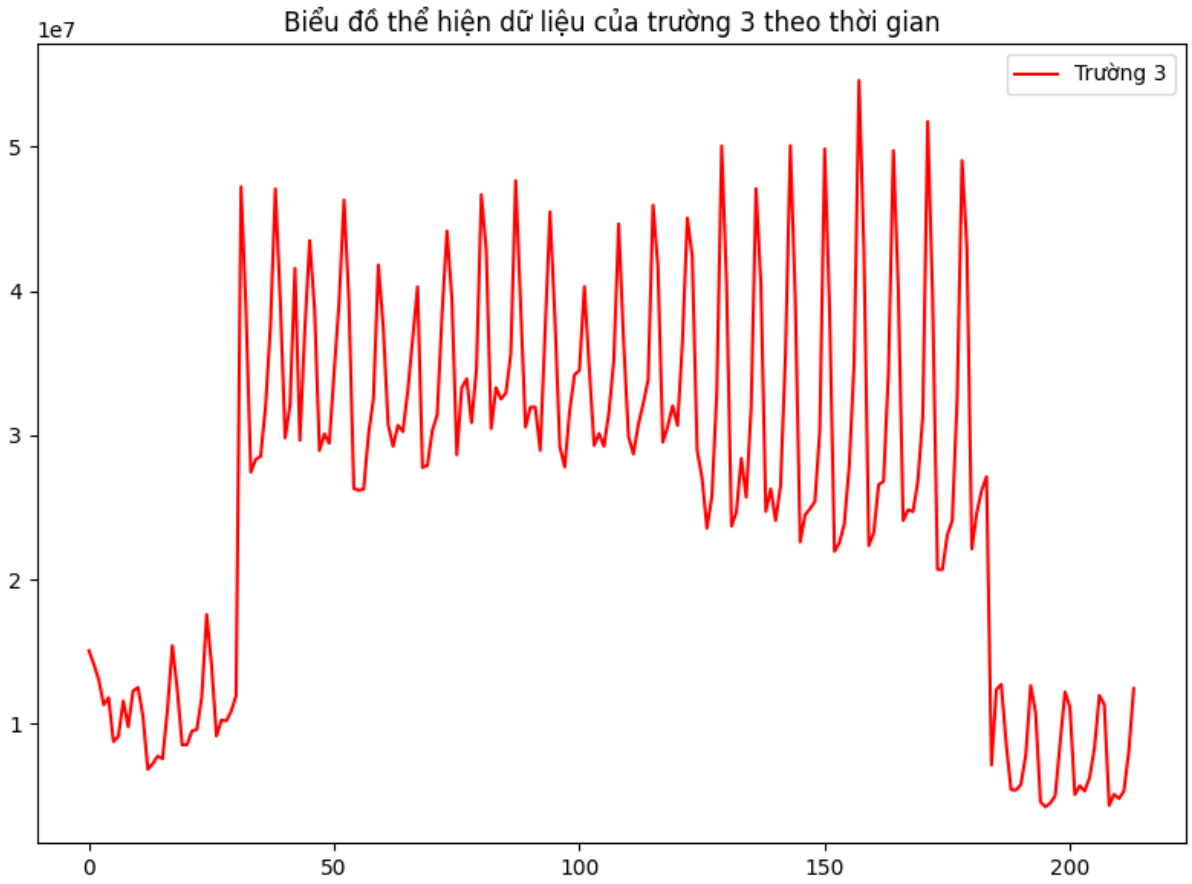
- Vẽ biểu đồ heatmap, từ đó ta thấy giữa các trường có sự tương quan với nhau.



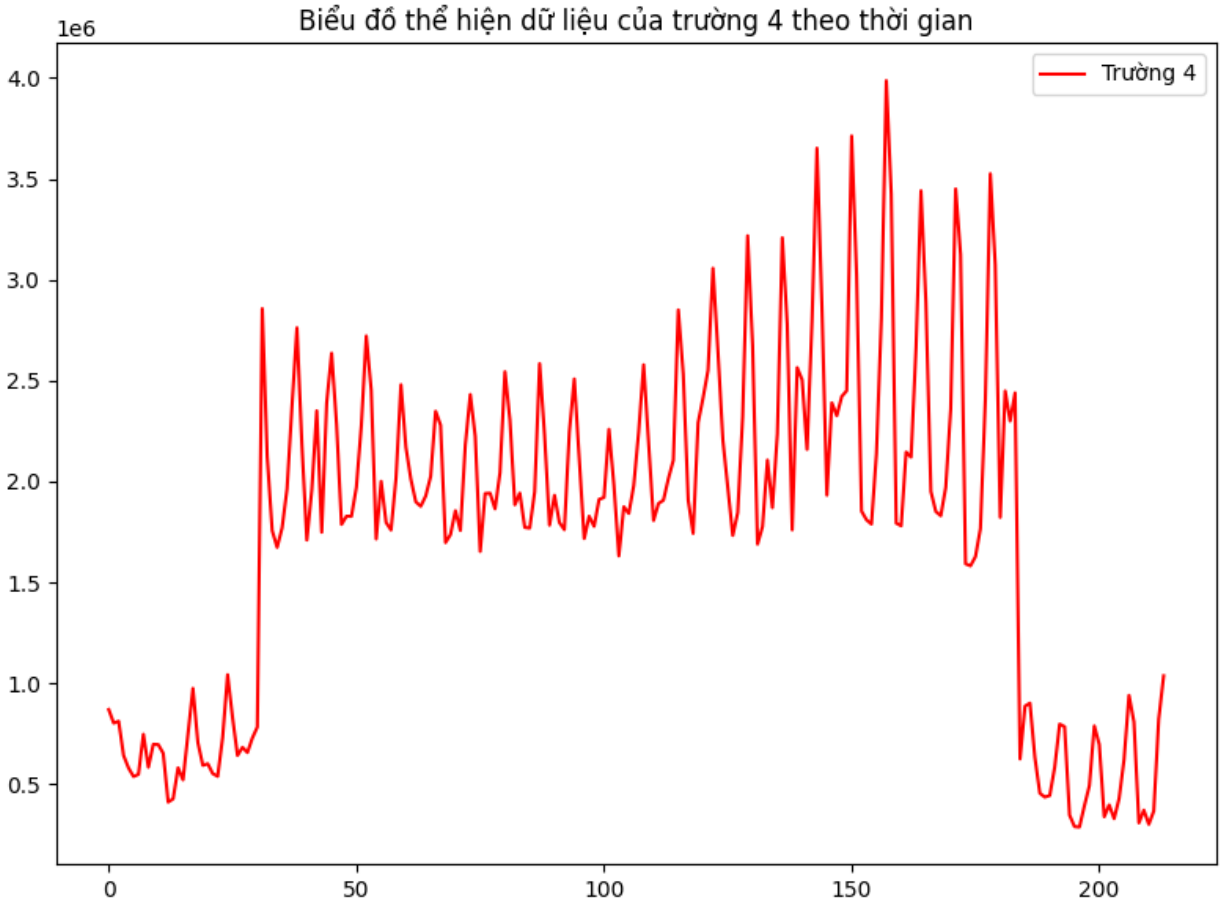
- Biểu đồ thể hiện dữ liệu của trường 1 theo thời gian.



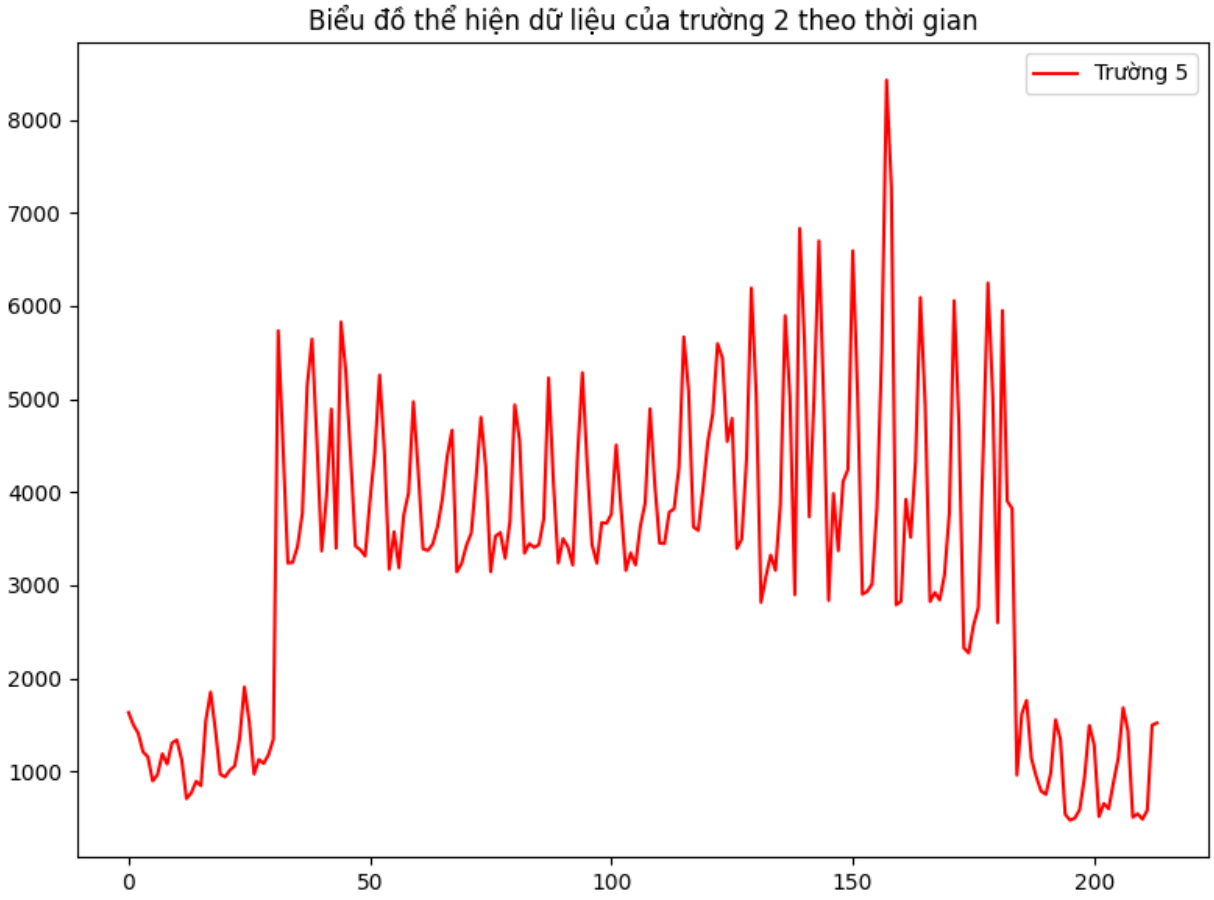
- Biển đồ thể hiện dữ liệu trường 2 theo thời gian.



- Biểu đồ thể hiện dữ liệu của trường 3 theo thời gian.



- Biểu đồ thể hiện dữ liệu của trường 4 theo thời gian.



- Biểu đồ thể hiện dữ liệu của trường 5 theo thời gian.
-

a, Mô hình Sarima:

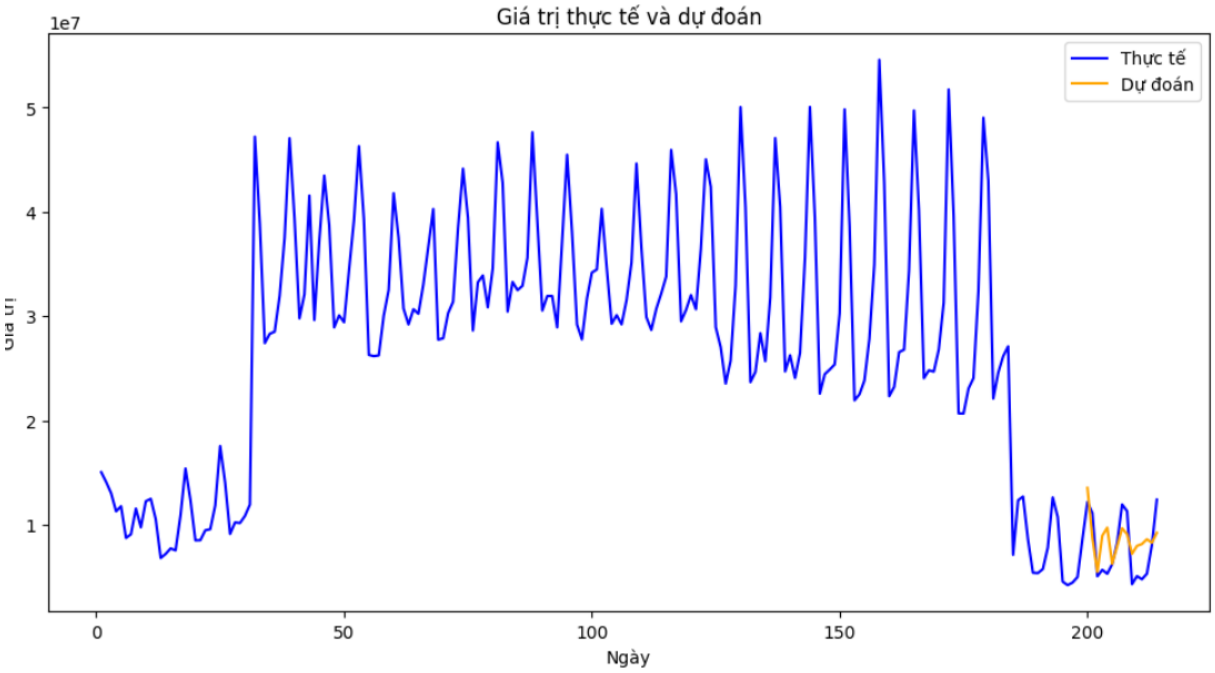
```
SARIMA_model = auto_arima(train_df["truong_3"].values.reshape(-1,1), start_p=1, start_q=1,
                          test='adf',
                          max_p=3, max_q=3,
                          m=6, #12 is the frequncy of the cycle
                          start_P=0,
                          seasonal=True, #set to seasonal
                          d=None,
                          D=1, #order of the seasonal differencing
                          trace=False,
                          error_action='ignore',
                          suppress_warnings=True,
                          stepwise=True)
print(SARIMA_model.summary())
```

```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                  199
Model:             SARIMAX(3, 0, 3)x(0, 1, [1, 2], 6)   Log Likelihood               -3293.894
Date:                                Tue, 04 Jun 2024   AIC                           6607.788
Time:                                        05:28:24   BIC                           6640.415
Sample:                                             0   HQIC                          6621.001
                                                  - 199
Covariance Type:                                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept  -7.598e+04    1.6e+05     -0.474      0.635    -3.9e+05    2.38e+05
ar.L1          0.5858      0.113      5.203      0.000       0.365       0.807
ar.L2         -0.3239      0.097     -3.344      0.001      -0.514      -0.134
```

```
sarima = SARIMAX(train_df['truong_3'].values.reshape(-1,1),
                 order=(3,0,3),
                 seasonal_order=(0,1,1,6)).fit()
predictions = sarima.forecast(steps=15)
```

```
print(predictions)
```

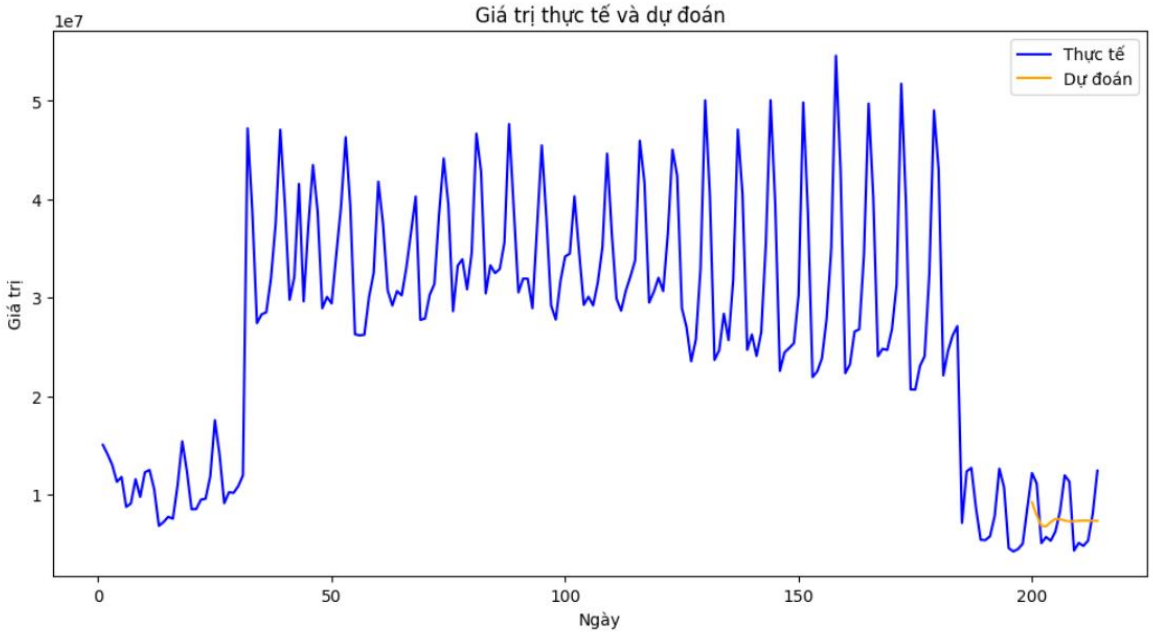
```
[13567909.15659588  8766537.83955109  5522884.94330143  8993327.13638207
  9763032.83491153  6295083.1010228   7981218.70710666  9705960.71277191
  9134636.56251001  7257368.3315203   8014958.7314067   8191567.0957039
  8653025.18627475  8309666.22728938  9270770.13810547]
```



B, Mô hình Arimax

```
ARIMAX_model = auto_arima(train_df['truong_3'].values.reshape(-1,1),
                  exogenous=train_df['truong_1'],
                  trace=True,
                  error_action="ignore",
                  suppress_warnings=True)
print(ARIMAX_model.summary())
```

```
arimax = SARIMAX(train_df['truong_3'].values.reshape(-1,1),
                 order=(2,1,1),
                 exogenous=train_df['truong_1']).fit()
predictions = arimax.forecast(steps=15)
```


Giá trị thực tế và dự đoán
1e7
Thực tế
Dự đoán
Giá trị
Ngày
0
50
100
150
200
1
2
3
4
5