Mô hình VARMAX và LSTM trong dự báo chất lượng không khí

Nhóm 16
Phạm Thị Hoa - 20195874
Lê Thanh Thảo - 20195919
Phạm Thu Trang - 20195931
Giảng viên hướng dẫn: TS. Nguyễn Thị Ngọc Anh

Hanoi, 3/2023

Outline

1. Giới thiệu bài toán
2. Mô hình VARMAX
3. Mô hình LSTM
4. Kết quả thực nghiệm
5. Kết luận

Outline

1. **Giới thiệu bài toán**
2. Mô hình VARMAX
3. Mô hình LSTM
4. Kết quả thực nghiệm
5. Kết luận

Giới thiệu bài toán

- Sử dụng hai mô hình VARMAX và LSTM dự đoán chất lượng không khí.
- Đối tượng nghiên cứu: các chỉ số ảnh hưởng đến chất lượng không khí trong một khu vực: nhiệt độ, độ ẩm, lượng bụi mịn PM10, PM 2.5, nồng độ CO, SO2, NO... trong không khí
- Phạm vi nghiên cứu: từ ngày 14/4/2015 tới ngày 5/9/2019.

Outline

1. Giới thiệu bài toán
2. **Mô hình VARMAX**
3. Mô hình LSTM
4. Kết quả thực nghiệm
5. Kết luận

Mô hình VARIMAX

Mô hình VARMAX gồm 3 thành phần chính:

- Vector Auto regression (VAR)
- Moving average (MA)
- Exogenous variable (X)

Mô hình VARIMAX

Mô hình toán học của VARMAX được định nghĩa như sau:

$$y_t = v + A_1 y_{t-1} + \cdots + A_p y_{t-p} + B x_t + \epsilon_t + M_1 \epsilon_{t-1} + \cdots + M_q \epsilon_{t-q} \tag{2.1}$$

trong đó:

- y_t là một vector của các giá trị hiện tại và các giá trị y khác là các giá trị trễ.
- A_s là các hệ số tự tương quan: đối với mỗi độ trễ, chúng là một vector có cùng độ dài với số lượng chuỗi thời gian.
- M_s là vector của các hệ số của các sai số mô hình trễ. Chúng đại diện cho phần trung bình cộng của mô hình.

Outline

1. Giới thiệu bài toán
2. Mô hình VARMAX
3. **Mô hình LSTM**
4. Kết quả thực nghiệm
5. Kết luận

Mô hình Long Short-term Memory

- Mạng bộ nhớ dài-ngắn (Long Short Term Memory networks), thường được gọi là LSTM - là một phiên bản mở rộng của mạng thần kinh hồi quy (RNN) nhân tạo được sử dụng trong lĩnh vực học sâu.
- LSTM được giới thiệu bởi **Hochreiter & Schmidhuber** vào năm 1997.
- LSTM được thiết kế để giải quyết các bài toán về phụ thuộc xa (long-term dependency) trong mạng RNN do bị ảnh hưởng bởi vấn đề gradient biến mất.

Mô hình Long Short-term Memory

Một đơn vị LSTM thông thường bao gồm:

- **Tế bào (cell)**
- **Tầng cổng quên (forget gate)**
- **Tầng cổng vào (input gate)**
- **Tầng ổng ra (output gate)**

Trạng thái tế bào (Cell state)

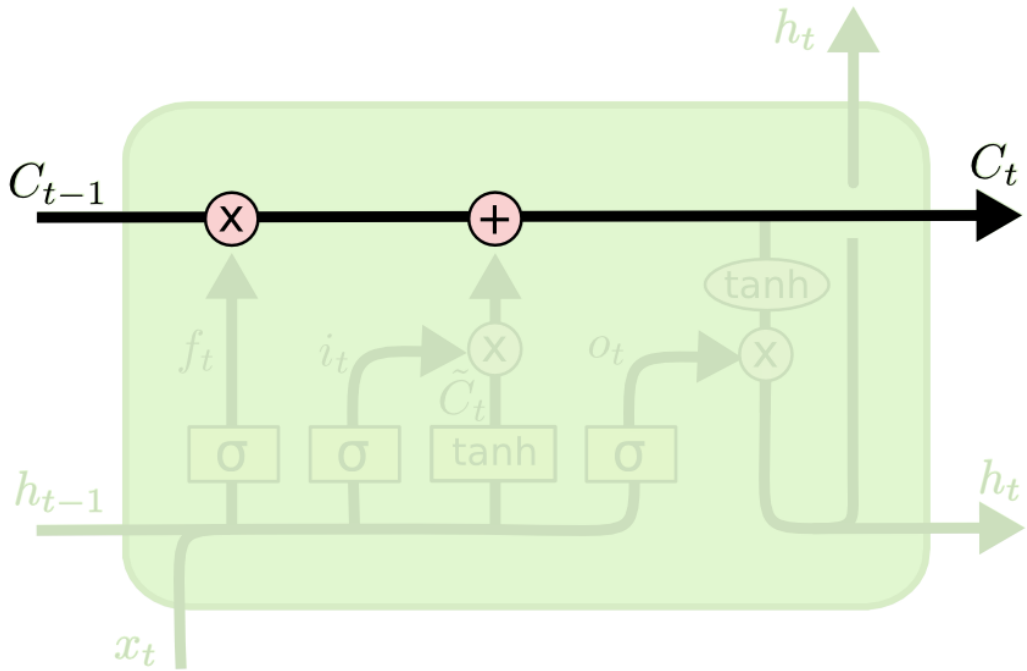


Figure: Cell state in LSTM.

Kiến trúc bên trong LSTM

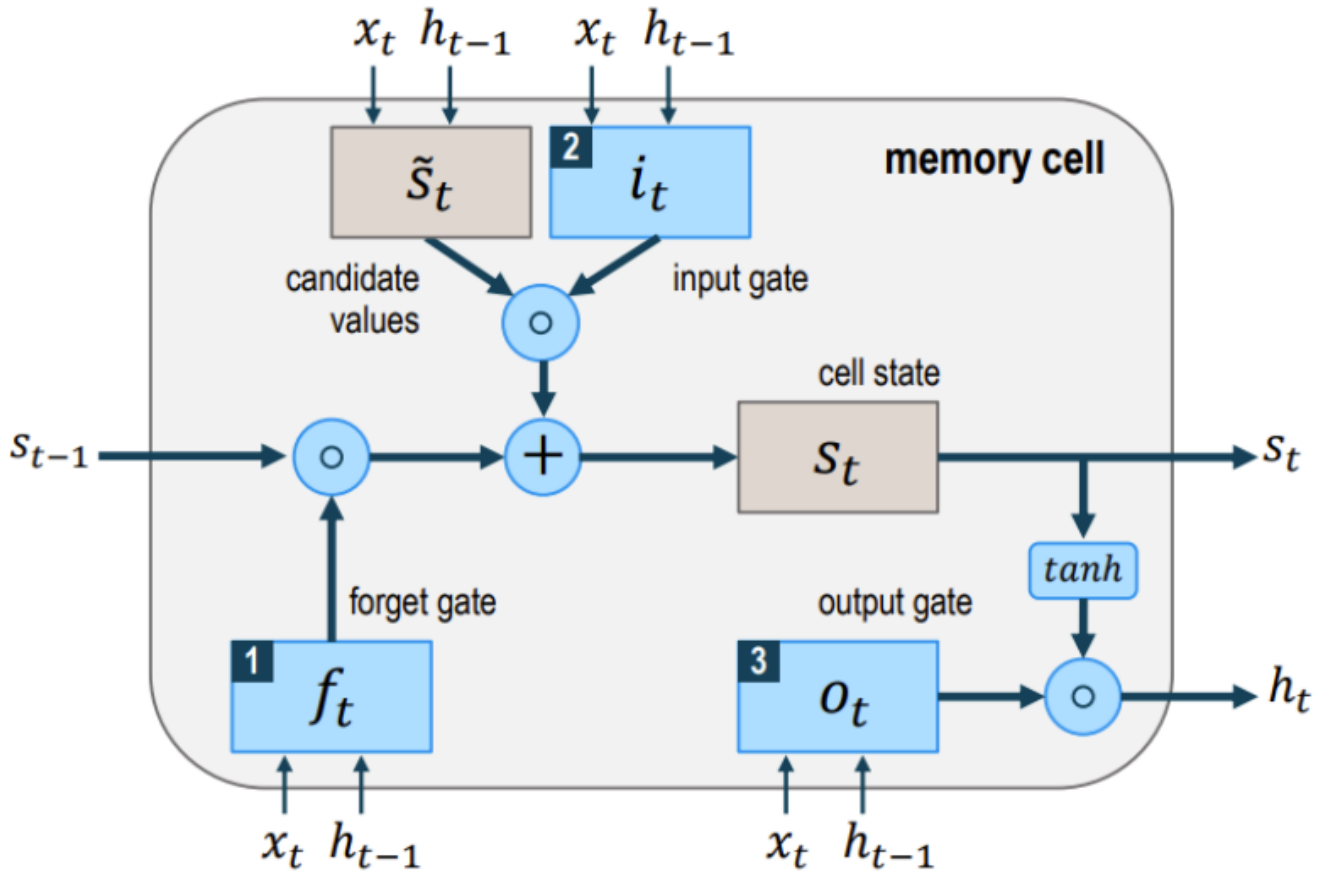
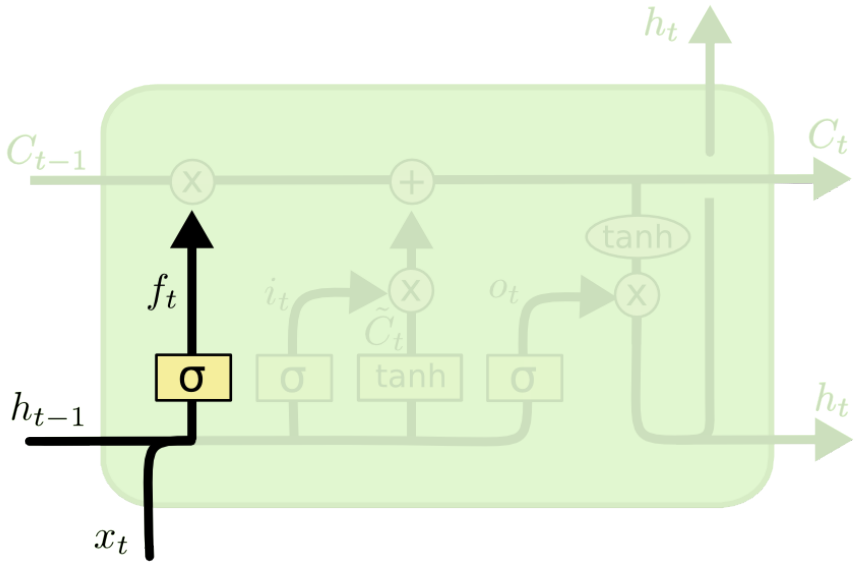


Figure: Sơ đồ biểu diễn kiến trúc bên trong của một tế bào LSTM.

Tầng cổng quên

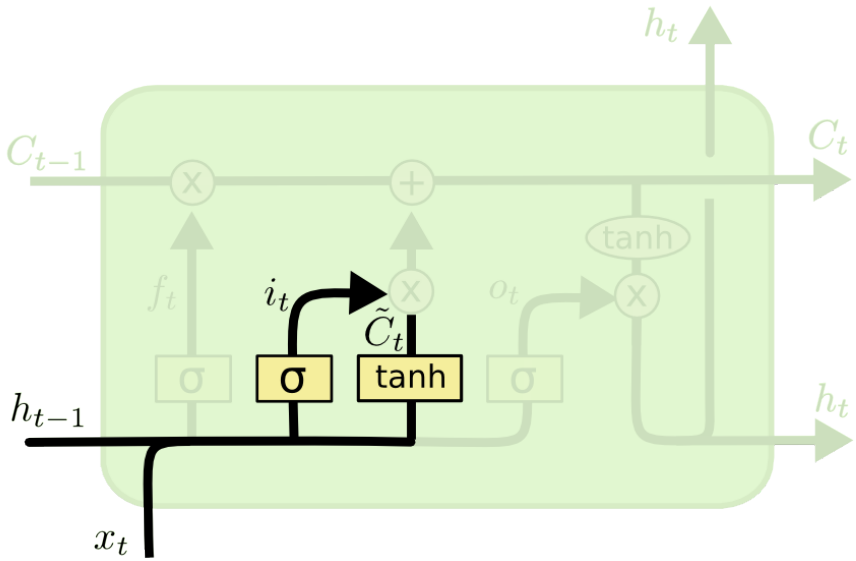


$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

Figure: Tầng cổng quên.

Tầng cổng quên có nhiệm vụ loại bỏ những thông tin không cần thiết nhằn được khỏi cell internal state.

Tầng cổng vào



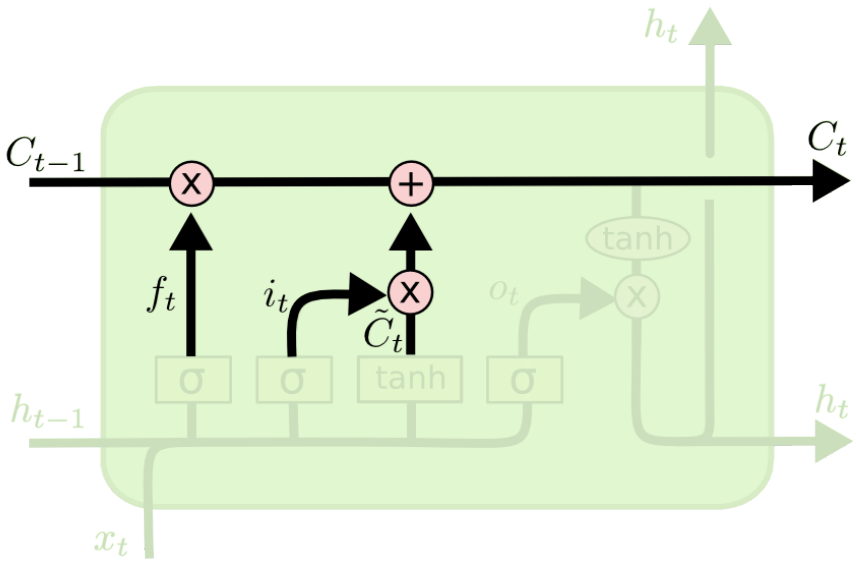
$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

Figure: Tầng cổng vào.

Tầng cổng vào có nhiệm vụ chọn lọc những thông tin cần thiết nào được thêm vào cell internal state.

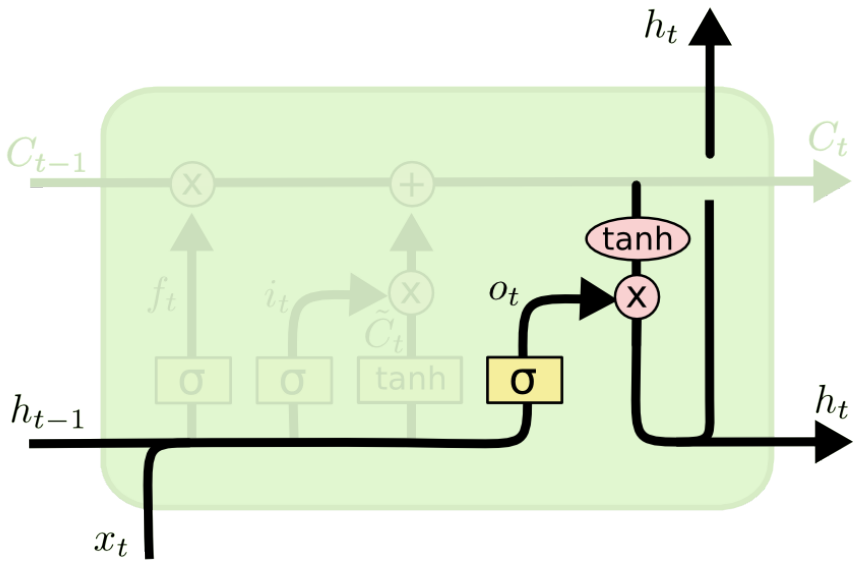
Cập nhật trạng thái tế bào



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Figure: Cập nhật trạng thái tế bào.

Cổng ra



$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

Figure: Cổng ra.

Tầng cổng ra có nhiệm vụ xác định những thông tin vào từ cell internal state được sử dụng như đầu ra.

Outline

1. Giới thiệu bài toán
2. Mô hình VARMAX
3. Mô hình LSTM
4. **Kết quả thực nghiệm**
5. Kết luận

Mô tả bộ dữ liệu

Để thực hiện dự đoán với hai mô hình LSTM và VARMAX, chúng em sẽ thực nghiệm trên dữ liệu về chỉ số môi trường không khí.

- Nguồn thu thập: Được cung cấp bởi giảng viên bộ môn
- Thời gian của dữ liệu: 142 ngày từ 7:00 14/4/2019 đến 6:00 5/9/2019 tính theo từng giờ.
- Kích thước dữ liệu: 556KB
- Giá trị cần dự đoán: mô hình dự đoán tất cả các giá trị trong bộ dữ liệu

Mô tả bộ dữ liệu

	Barometer	Temp	NO	PM-10	RH	Radiation	WindDir	SO2	NOx	NO2	...	CO	PM-1	O3	Wind Spd	TSP	time	WinDir1	Temp1	ASKQ	Wind Spd (sai)
0	1010.280000	25.550000	18.57	18.12	98.55	0.01	188.880000	19.17	68.232	78.45	...	461.82	3.43	-17.17	1.480000	31.02	14/4/2019 07:00:00	NaN	NaN	NaN	NaN
1	1011.210000	25.170000	11.61	17.25	99.62	32.83	194.240000	13.49	48.492	58.17	...	595.40	3.43	-3.70	1.550000	31.13	14/4/2019 08:00:00	NaN	NaN	NaN	NaN
2	1012.360000	25.860000	8.79	12.04	97.46	311.73	108.600000	6.68	30.564	34.38	...	553.42	2.43	3.09	1.240000	21.29	14/4/2019 09:00:00	NaN	NaN	NaN	NaN
3	1012.660000	27.670000	11.33	10.91	88.45	532.91	119.740000	5.77	30.996	31.16	...	446.55	2.14	8.60	1.780000	19.24	14/4/2019 10:00:00	NaN	NaN	NaN	NaN
4	1012.780000	29.140000	5.50	8.91	77.99	610.51	116.880000	5.72	23.676	28.68	...	435.10	1.87	33.08	1.560000	18.48	14/4/2019 11:00:00	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3409	1006.197202	24.618810	2.31	12.86	62.53	NaN	177.290079	70.69	2.448	0.29	...	122.13	2.98	91.97	0.994953	23.62	5/9/2019 2:00	103.96	30.69	1002.97	2.46
3410	1004.847083	27.431410	2.30	10.30	62.61	NaN	131.708012	72.00	2.400	0.23	...	114.50	3.01	90.53	1.789547	18.27	5/9/2019 3:00	84.25	30.64	1002.61	2.57
3411	1004.401372	24.187061	2.47	10.48	62.80	NaN	148.633002	73.61	2.520	0.17	...	110.68	2.98	88.60	1.083361	17.82	5/9/2019 4:00	74.20	30.63	1002.26	3.08
3412	1009.252225	30.123856	2.48	11.45	63.54	NaN	211.588451	73.22	2.388	NaN	...	122.13	3.07	88.27	1.401751	18.58	5/9/2019 5:00	91.78	30.46	1002.10	2.75
3413	1004.019334	27.841429	2.55	14.54	70.41	NaN	109.182176	77.79	4.584	3.27	...	232.82	3.70	76.07	1.305240	24.03	5/9/2019 6:00	249.09	29.01	1002.14	1.58

3414 rows × 22 columns

Figure: Dữ liệu môi trường không khí từ 7:00 14/4/2019 đến 6:00 5/9/2019

Tiền xử lý dữ liệu

Nhận thấy tỷ lệ dữ liệu khuyết thiếu của từng cột cũng không cao lắm nên có thể giữ lại và sẽ sử dụng Linear-Regression để dự đoán các giá trị còn thiếu ta thu được kết quả:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3414 entries, 0 to 3413
Data columns (total 22 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   Barometer       3414 non-null   float64
 1   Temp            3414 non-null   float64
 2   NO              3414 non-null   float64
 3   PM-10           3414 non-null   float64
 4   RH              3414 non-null   float64
 5   Radiation       3414 non-null   float64
 6   WindDir         3414 non-null   float64
 7   SO2             3414 non-null   float64
 8   NOx             3414 non-null   float64
 9   NO2             3414 non-null   float64
 10  Compass         3414 non-null   float64
 11  PM-2-5          3414 non-null   float64
 12  CO              3414 non-null   float64
 13  PM-1            3414 non-null   float64
 14  O3              3414 non-null   float64
 15  Wind Spd        3414 non-null   float64
 16  TSP             3414 non-null   float64
 17  time            3414 non-null   object 
 18  WinDir1         3414 non-null   float64
 19  Temp1           3414 non-null   float64
 20  ASKQ            3414 non-null   float64
 21  Wind Spd (sai)  3414 non-null   float64
dtypes: float64(21), object(1)
memory usage: 586.9+ KB
```

Figure: Thông tin dữ liệu sau khi xử lý

Các metrics đánh giá

Root Mean Squared Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(y_i - \hat{y}_i)^2} \tag{4.1}$$

với y_i là giá trị thực sự cần dự đoán, và $\hat{y}_i$ là giá trị mô hình dự đoán, n là kích thước của dữ liệu cần dự đoán.

Các metrics đánh giá

Mean Squared Error (MSE):

MSE được hiểu là giá trị sai số bình phương trung bình hoặc là lỗi bình phương trung bình. Nó đề cập đến giá trị trung bình của chênh lệch bình phương giữa tham số dự đoán và tham số quan sát được và có công thức như sau:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4.2}$$

với y_i là giá trị thực sự cần dự đoán, và $\hat{y}_i$ là giá trị mô hình dự đoán, n là kích thước của dữ liệu cần dự đoán..

Kết quả -LSTM

LOSS: được tính bằng MSE, LOSS của mô hình sau khi huấn luyện qua 200 epoch

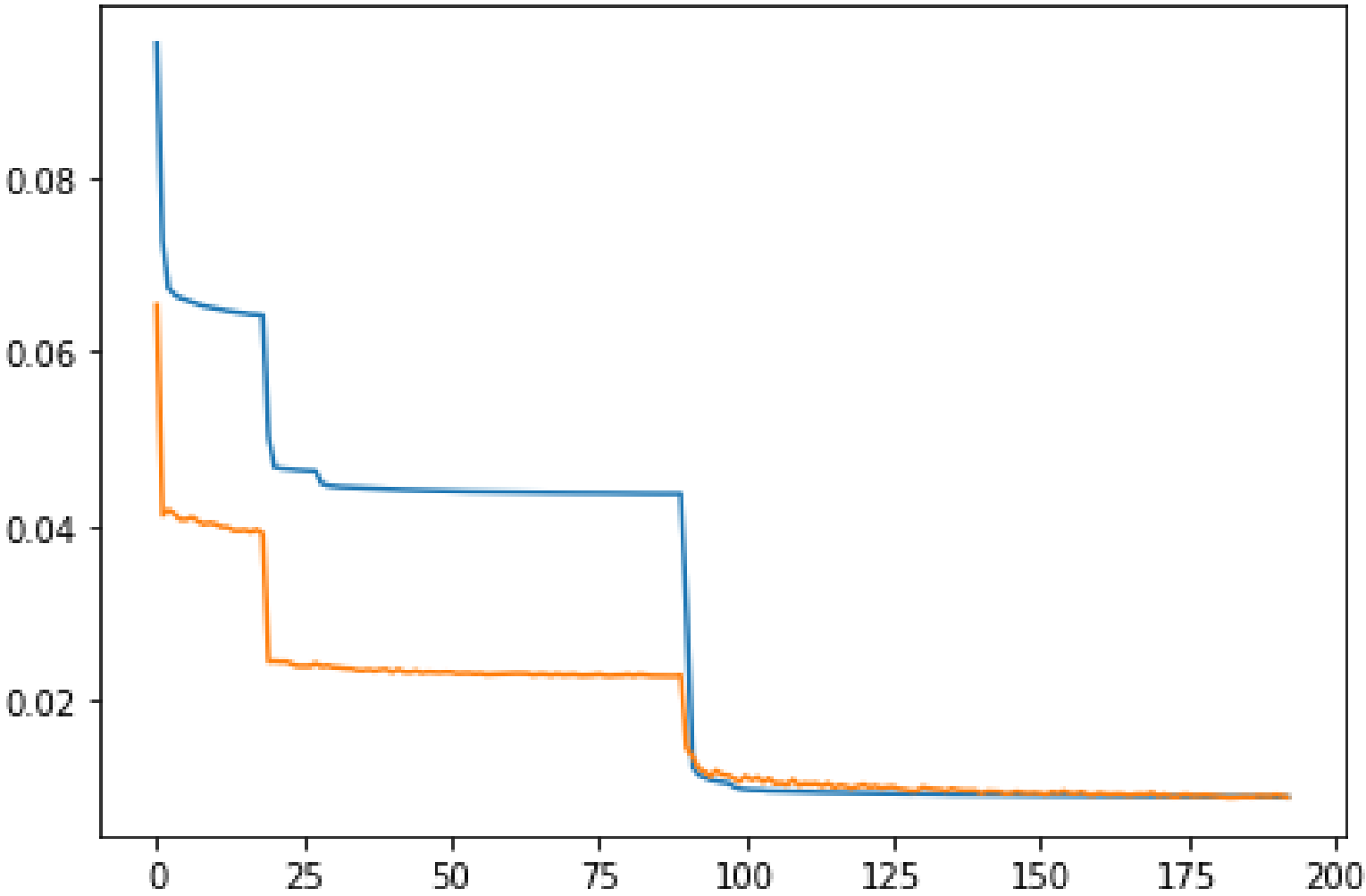


Figure: LOSS của mô hình LSTM

Kết quả

RMSE của mô hình sau khi huấn luyện qua 200 epoch:

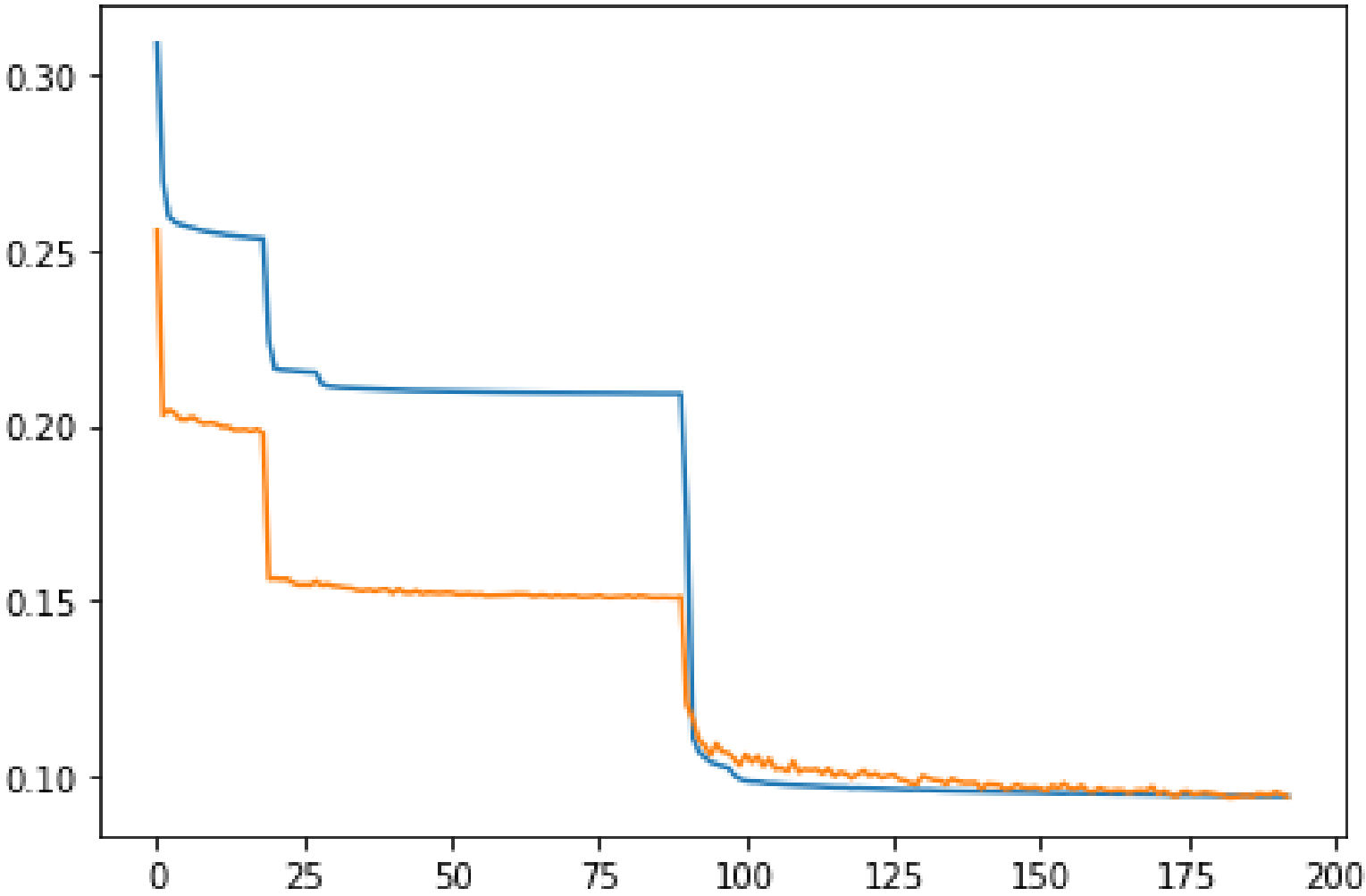


Figure: RMSE của mô hình LSTM

Kết quả

Kết quả dự đoán của mô hình với 48h cuối của bộ dữ liệu so với giá trị thực trên tập dữ liệu, trên một số cột:

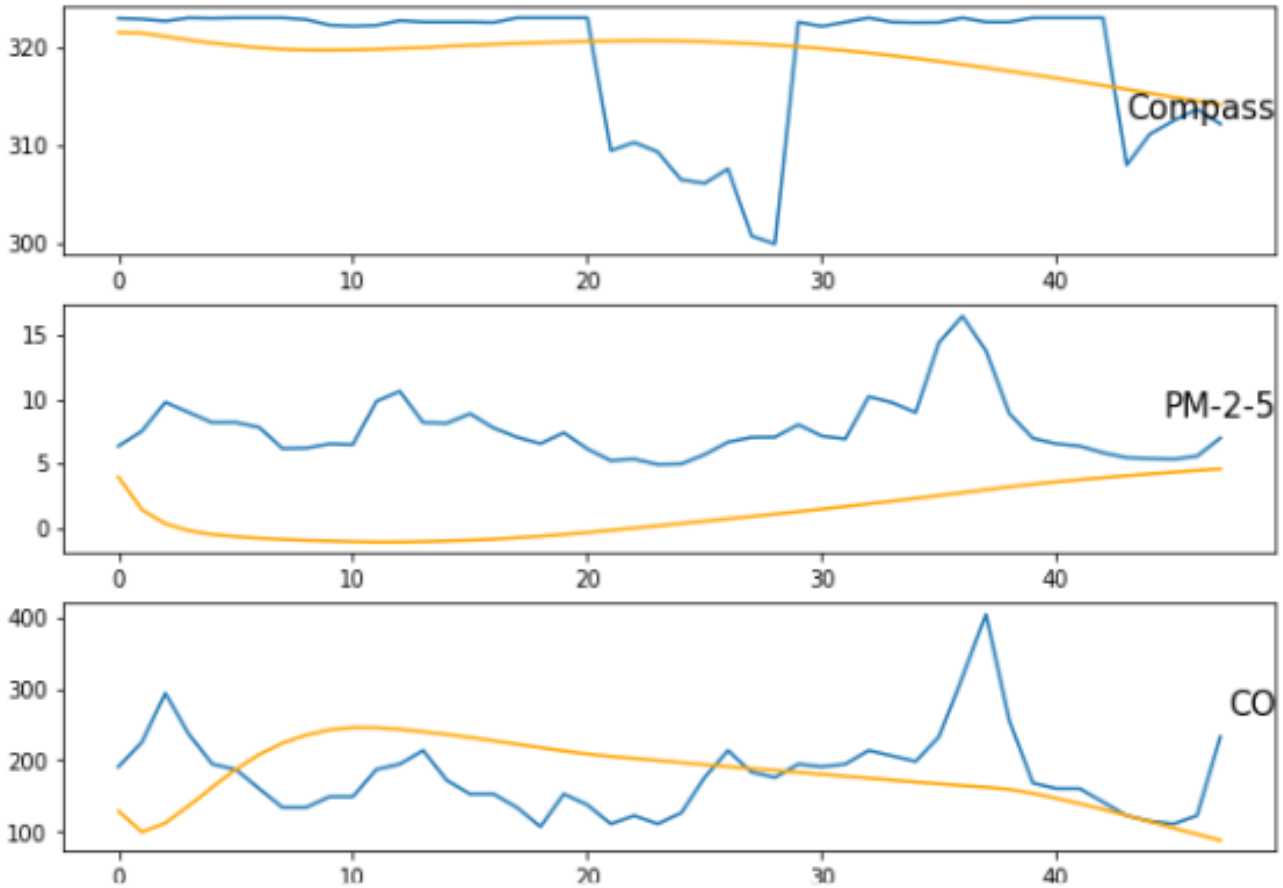


Figure: Compass, PM-2-5 và CO

Kết quả -VARMAX

- Giá trị MSE của mô hình sau khi huấn luyện là **1031.6635015800691**
- Kết quả đánh giá theo một số cột:

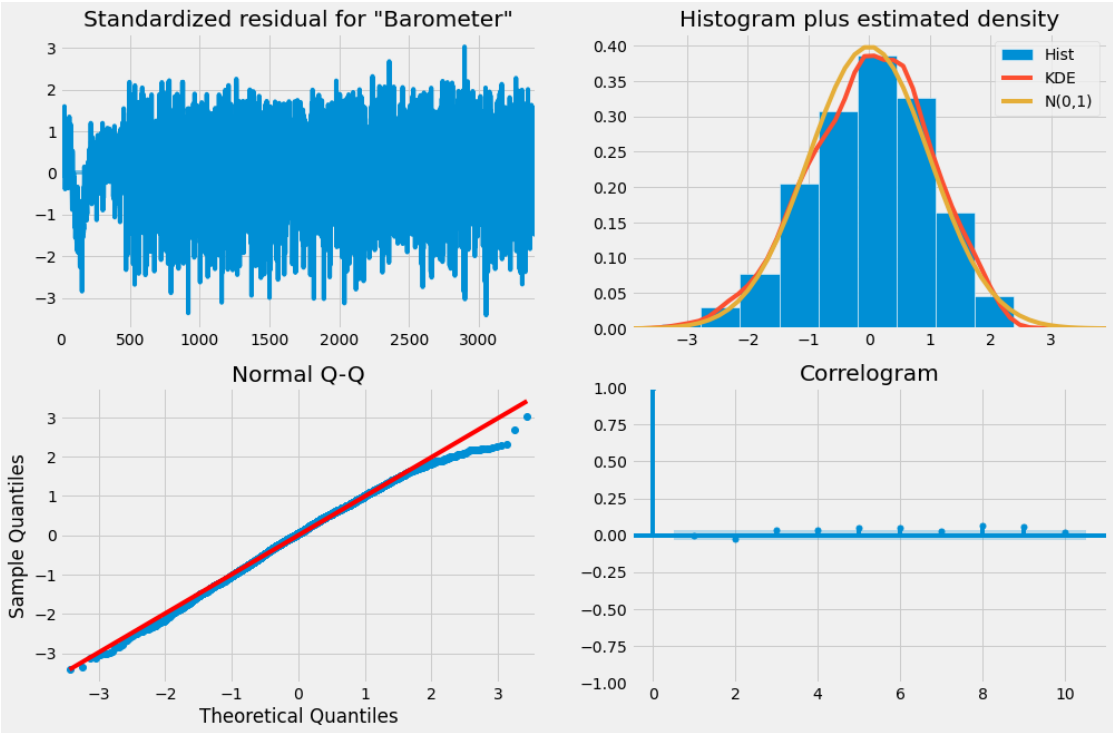


Figure: Barometer

Kết quả

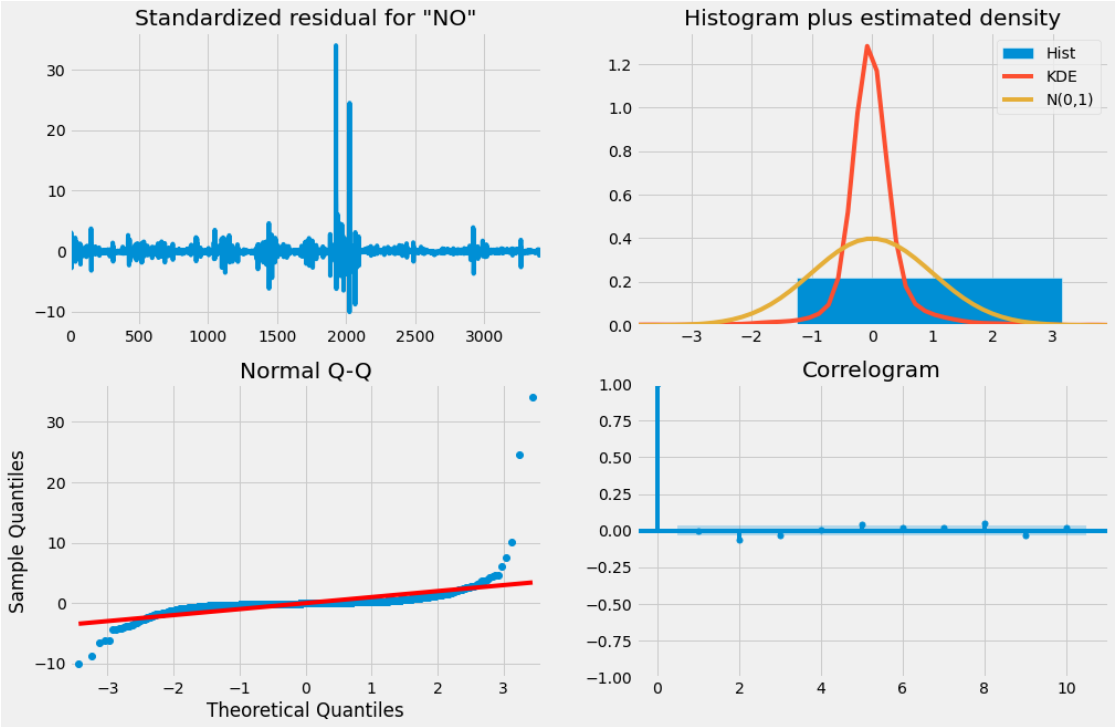


Figure: NO

Kết quả

Kết quả dự đoán so với tập test

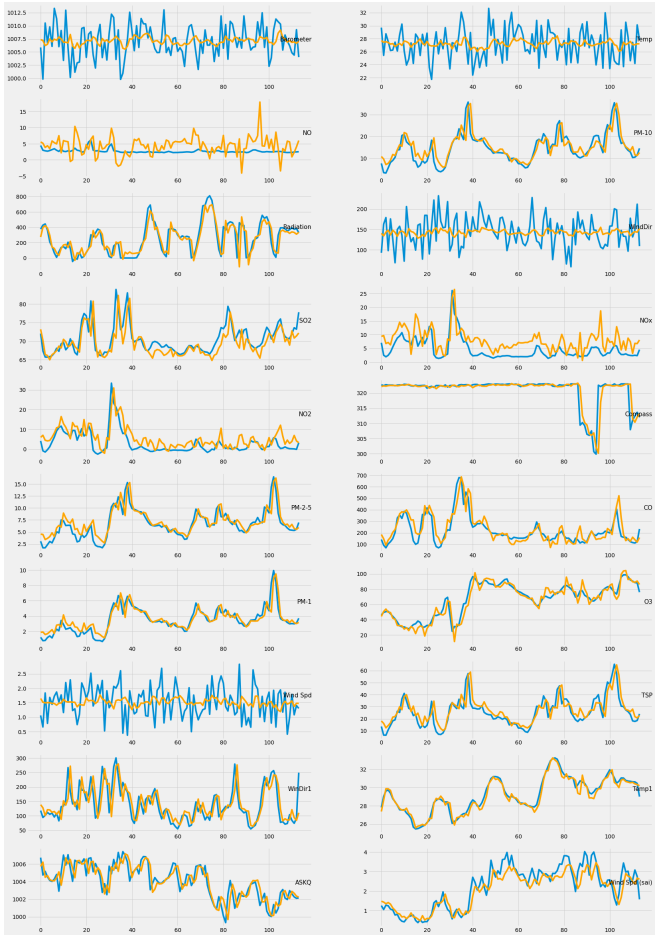


Figure: Kết quả dự đoán so với tập test

Outline

1. Giới thiệu bài toán
2. Mô hình VARMAX
3. Mô hình LSTM
4. Kết quả thực nghiệm
5. **Kết luận**

Những điều đã làm được

Trong phạm vi nội dung của bài tập lớn, một số nội dung mà nhóm chúng em đã đạt được:

- Thành công trong việc giới thiệu 2 mô hình dự đoán chuỗi thời gian VARMAX và LSTM.
- Ứng dụng được vào bài toán dự đoán chất lượng không khí.

Thanks for listening!