ST4 - MDS

Market Dynamics

Using VAR

Hugo MURET, Hamza BEN MENA, Amélie BELLAZI, Mael DORARD & Joseph SERVIGNE

Summary

Objectives

MARKET DYNAMICS

Project Objectives

- Main objective : Using Autoregressive Models to predict Stock Market Dynamic

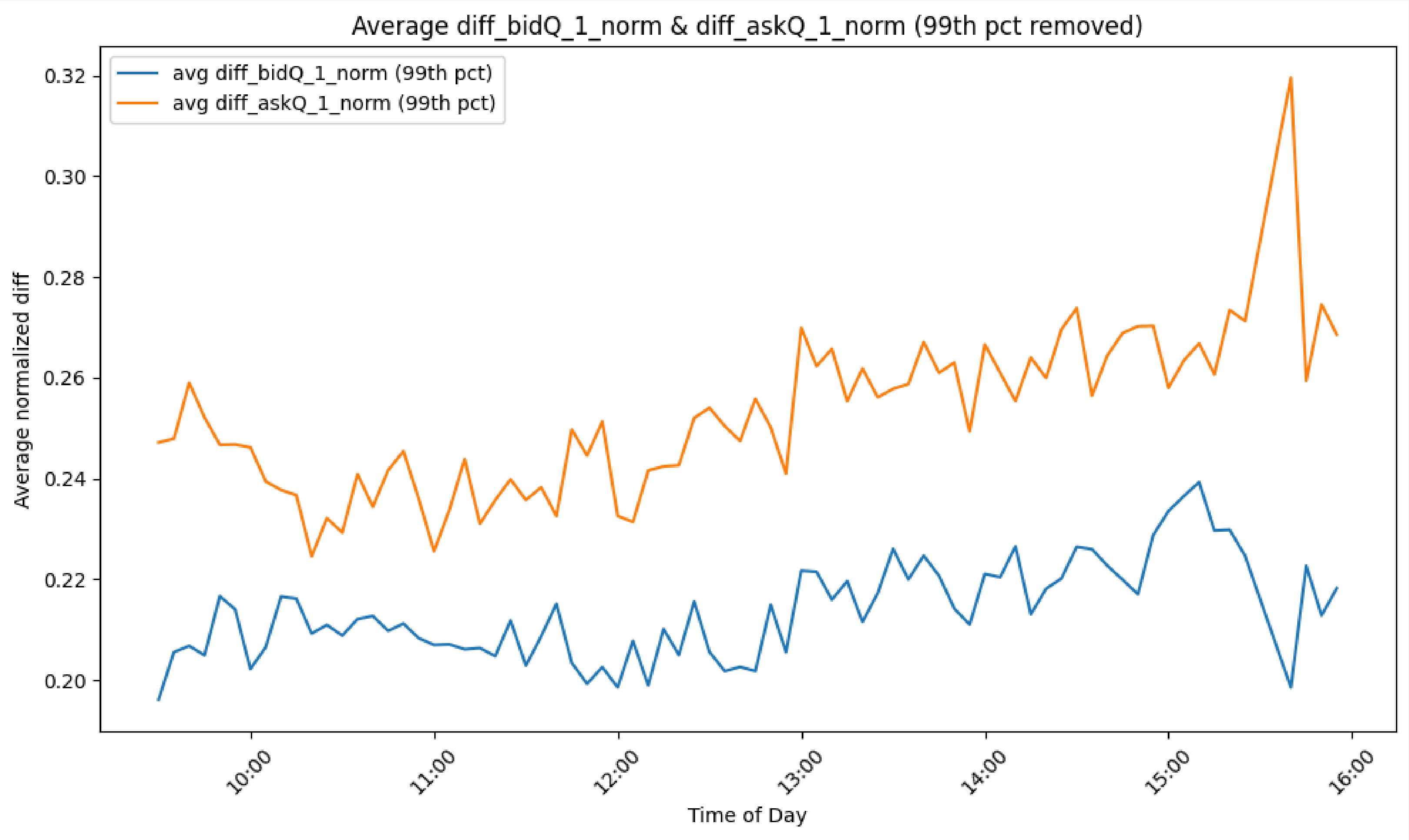
How to achieve it ?

- Filter and extract meaningful price changes
- Build aggregated order flow features
- Fit different VAR models to capture temporal dynamics and stability
- Compare them and pick the best

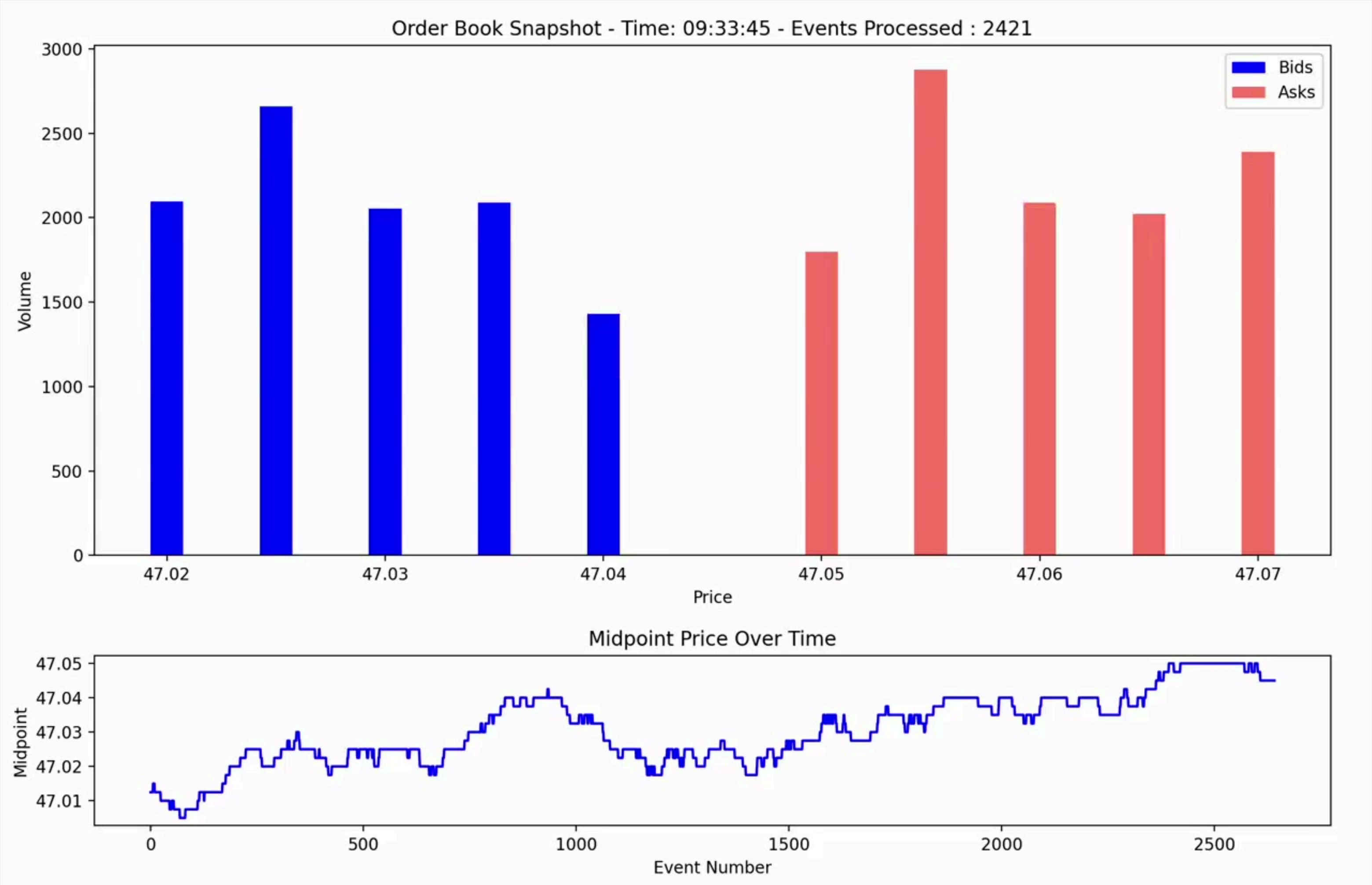
Introducing The Data

MARKET DYNAMICS

I - Data explanation and preparation



=> shows clear volume spikes at the beginning and the end of the day, caused by the increased trading activities



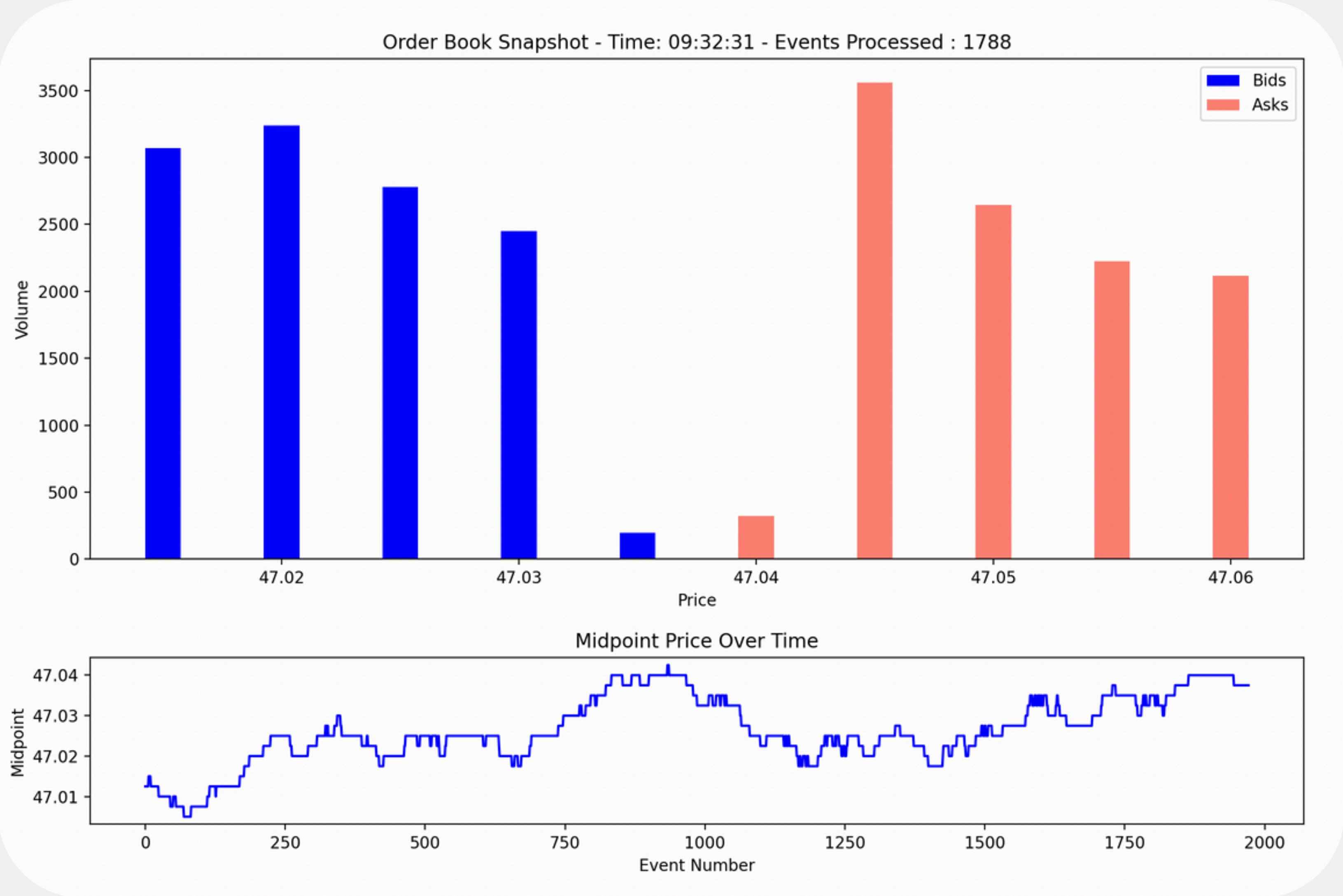
=> Evolution of 5 levels of bid / ask prices and mid price during one day

I - Data explanation and preparation

Analysis of the six variables :

Vol_lo_bid'	'Vol_c_ask'
'Vol_lo_ask'	'Vol_ex_bid'
'Vol_c_bid'	'Vol_ex_ask

Extracted from **TOTF_book_03_04_2017** and **TOTF_trade_2014_2017**



We decided to aggregate the data on 5min intervals rather than 1min to reduce noise

What's An Autoregressive Model ?

A **stochastic process** where past (**lagged**) values for a variable influences its current value (ex : stock price).

Formula AR(q) : $X_t = \varepsilon_t + \varphi_1 \cdot X_{t-1} + \varphi_2 \cdot X_{t-2} + \ldots + \varphi q \cdot X_{t-}q$

- X_t = Value of the series at time t, has to be **stationary**
- ε_t = White-noise error term (**mean 0**, **variance σ^2**)
- φ_i = Autoregressive coefficient for lag i
- q = Number of past lags used
  = The number of past observations influencing X_t.

Stationary if all the roots of the polynomial $z^p - \varphi_1 z^{p-1} - \varphi_2 z^{p-2} - \ldots - \varphi_p = 0$ lie outside the unit circle ($|z_i| > 1$ for all i)

The Models

MARKET DYNAMICS

First Model : VARI

II. Step 1 : Making the datas stationary

Purpose : Determine if a time series Xt has constant mean and variance over time.

Stationaty Tests :

- *ADF (Augmented Dickey–Fuller) :*
  - Null hypothesis (H_0): "Series has a unit root" (non-stationary)
  - If p-value < 0.05 → reject H_0 → series is stationary
- *KPSS (Kwiatkowski–Phillips–Schmidt–Shin) :*
  - Null hypothesis (H_0): "Series is stationary"
  - If p-value < 0.05 → reject H_0 → series is non-stationary

	Variable	ADF_p-value	KPSS_p-value	result
0	Vol_lo_bid	0.8138	0.01	Non-stationary
1	Vol_lo_ask	0.8138	0.01	Non-stationary
2	Vol_c_bid	0.9783	0.01	Non-stationary
3	Vol_c_ask	0.9705	0.01	Non-stationary
4	Vol_ex_bid	0.9478	0.01	Non-stationary
5	Vol_ex_ask	0.0000	0.10	Stationary

5 out of 6 variables are non-stationary => we must apply transformations to achieve stationarity.

II. Step 1 : Making the datas stationary

Purpose : Determine if a time series Xt has constant mean and variance over time.

```
if adf_p < signif and kpss_p > signif:
    decision = "Stationary"

elif adf_p >= signif and kpss_p <= signif:
    decision = "Non-stationary"

elif adf_p < signif and kpss_p <= signif:
    decision = "Trend-stationary"
else:
    decision = "Inconclusive"
```

	Variable	ADF_p-value	KPSS_p-value	result
0	Vol_lo_bid	0.8138	0.01	Non-stationary
1	Vol_lo_ask	0.8138	0.01	Non-stationary
2	Vol_c_bid	0.9783	0.01	Non-stationary
3	Vol_c_ask	0.9705	0.01	Non-stationary
4	Vol_ex_bid	0.9478	0.01	Non-stationary
5	Vol_ex_ask	0.0000	0.10	Stationary

5 out of 6 variables are non-stationary => we must apply transformations to achieve stationarity.

II. Step 1 : Making the datas stationary

Data Transformation :

- **Data clipping :**
  - Intraday Flow Profile: Volumes rise after 2:30 PM, introducing non-stationarity.
  - **Clipping off end-of-day** data removes the late-day spikes, producing a more stationary time series suitable for the VAR model.
- **Differentiation** :
  - first-order differencing on all numeric columns to achieve stationarity

Variable	ADF_p-value	KPSS_p-value	result
Vol_lo_bid	0.0	0.1	Stationary
Vol_lo_ask	0.0	0.1	Stationary
Vol_c_bid	0.0	0.1	Stationary
Vol_c_ask	0.0	0.1	Stationary
Vol_ex_bid	0.0	0.1	Stationary
Vol_ex_ask	0.0	0.1	Stationary

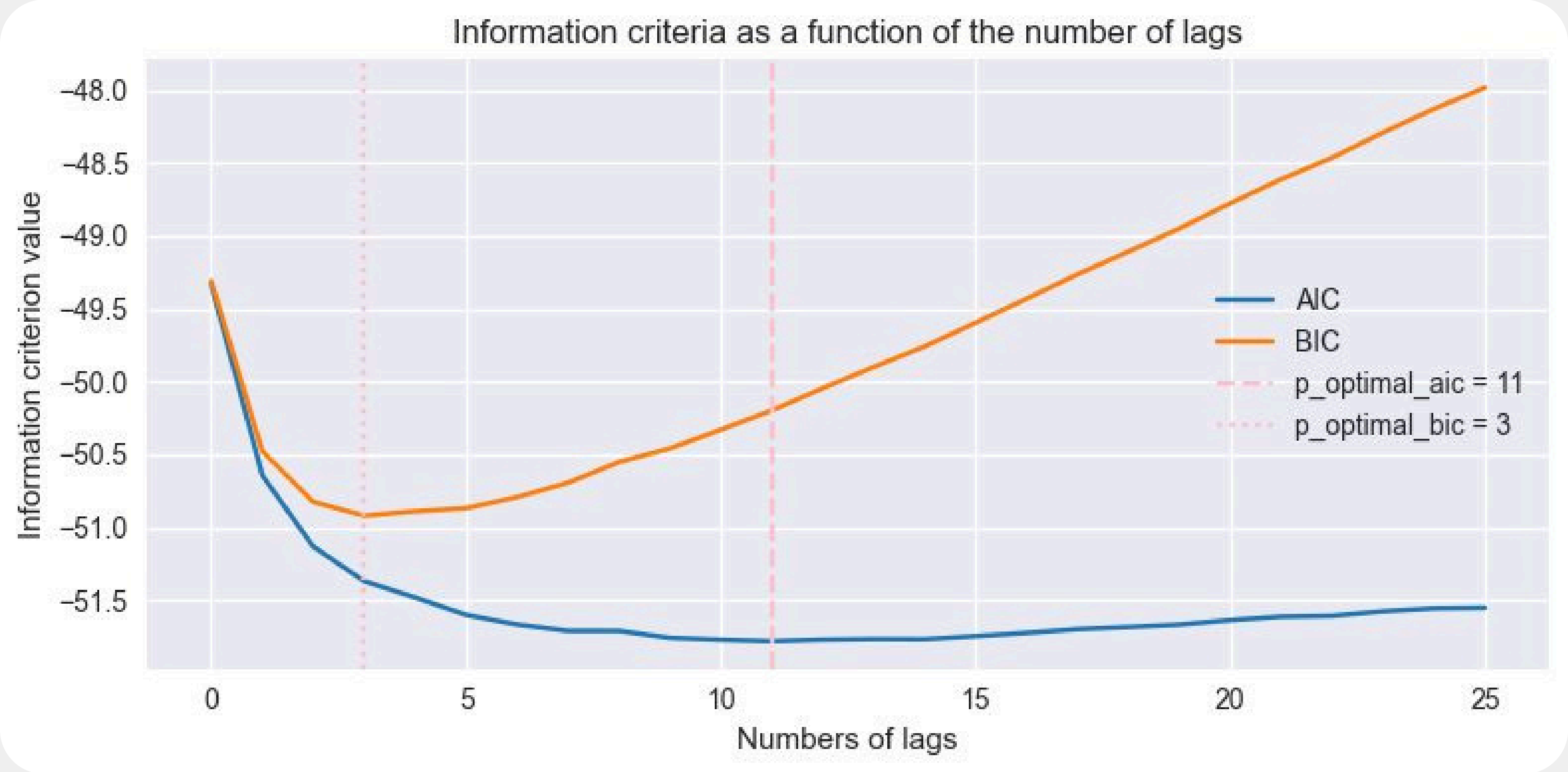
II. Step 2 : Lag Order Selection

AIC (Akaike Information Criterion) :

- AIC = -2 × ln(L^) + 2 × k
- Allows more parameters if they improve fit, favoring fit over simplicity.

BIC (Bayesian Information Criterion)

- BIC = -2 × ln(L^) + (ln n) × k
- Heavier penalty "(ln n) × k" => favors simpler models when n is large



RESULTS : AIC model : 11 lags
BIC model : 3 lags

<u>Fit vs Complexity</u> : choose lag that minimize the BIC => avoid overfitting

Second Model : Box-Cox

III. Step 1 : Making the data stationary

Same stationaty Tests :

- *ADF (Augmented Dickey–Fuller)*
- *KPSS (Kwiatkowski–Phillips–Schmidt–Shin)*

Data Transformation :

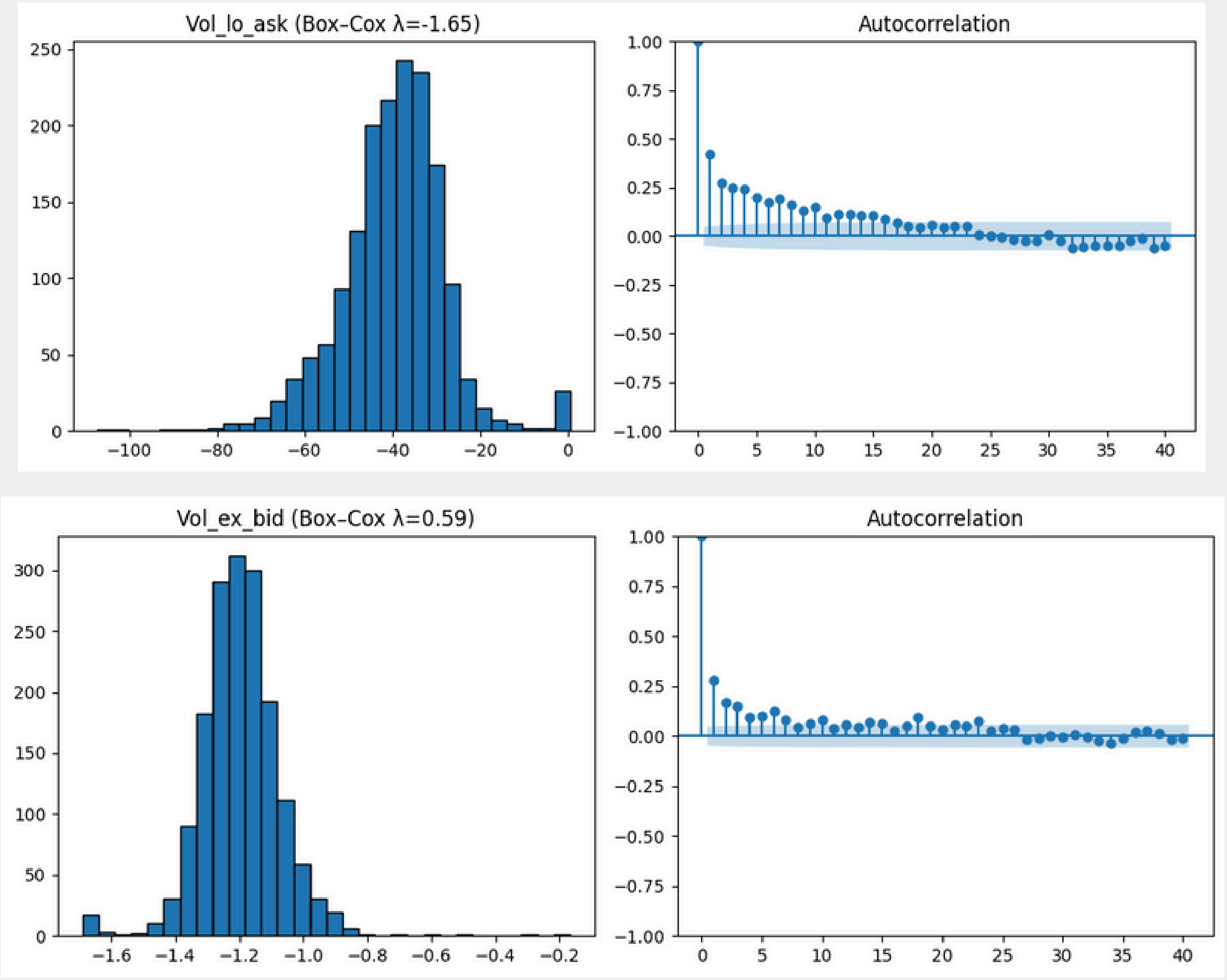
- Find optimal λ by maximizing log-likelihood
- Apply : $$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \ln(y), & \lambda = 0 \end{cases}$$

Estimated Box–Cox λ Parameters for Each Variable :

Vol_lo_bid: λ = 0.0193
Vol_lo_ask: λ = −1.6470
Vol_c_bid: λ = 0.0341
Vol_c_ask: λ = −0.8032
Vol_ex_bid: λ = 0.5916
Vol_ex_ask: λ = 0.4619

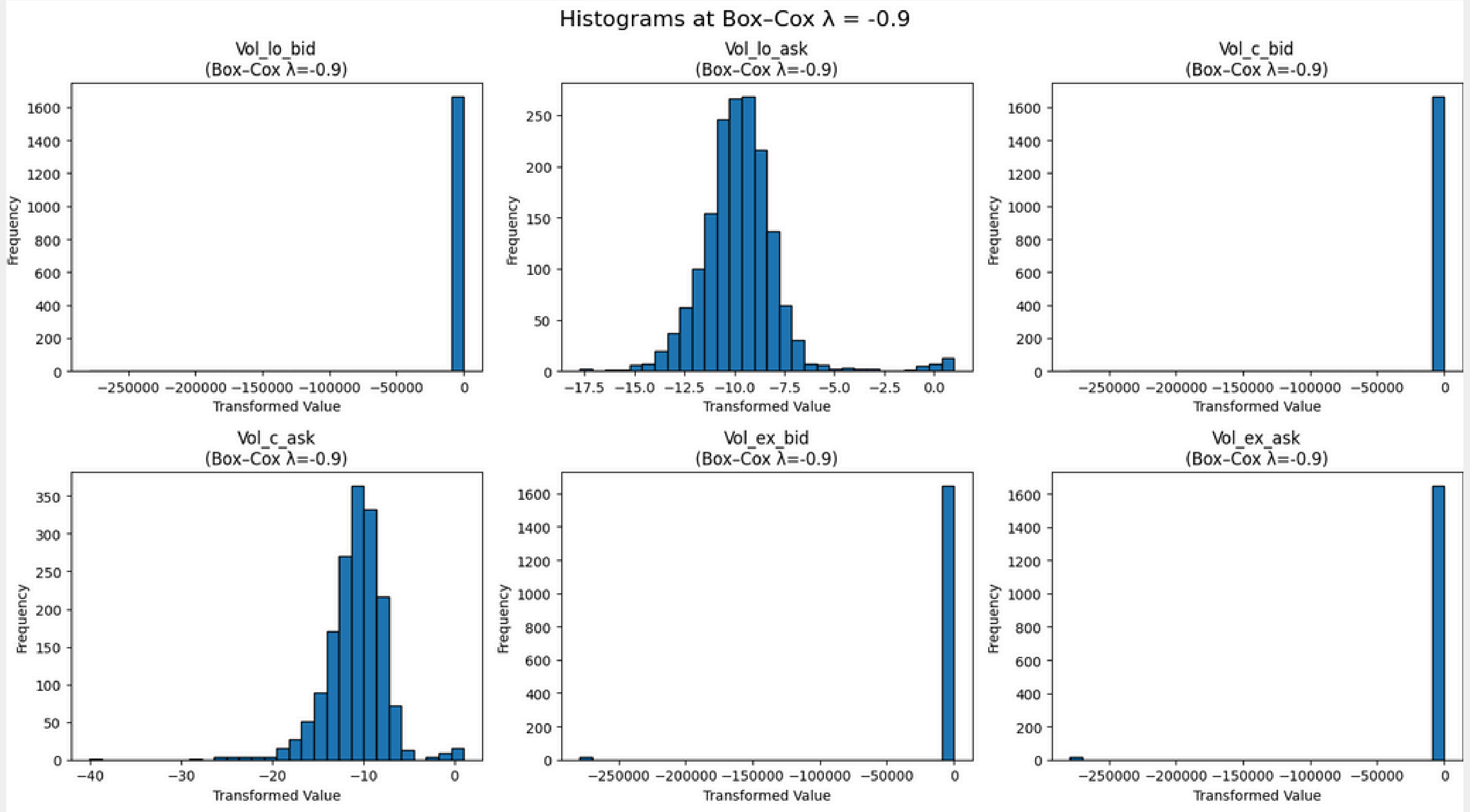
III. Step 1 : Making the datas stationary

1) Applying the Box Cox transformation to each variable



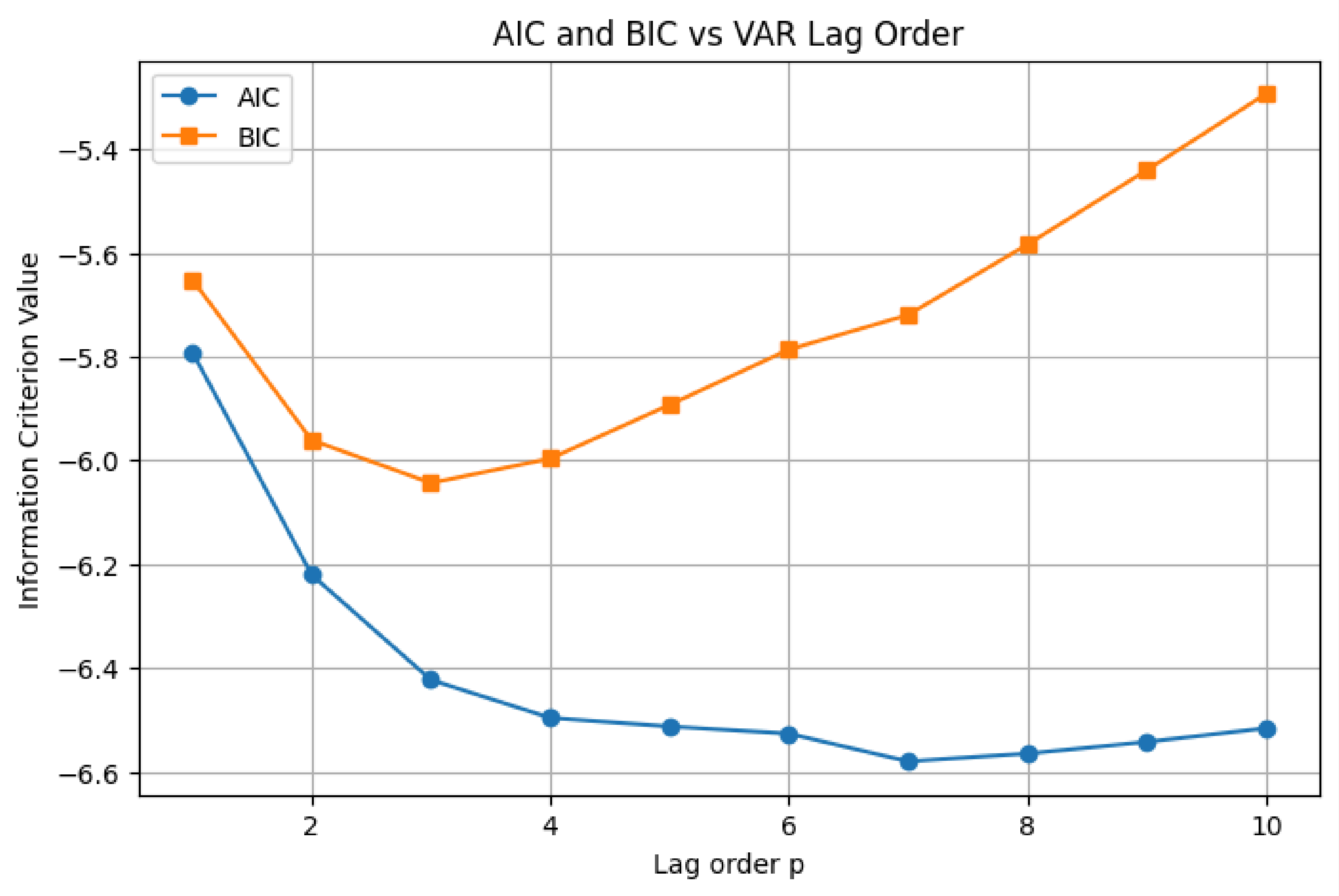
- Normal distributions
- 2/6 variables stationary

If we use the same λ for the 6 variables
=> non normal distribution



2) Differentiation of the remaining variables

III. Step 2 : Lag Order Selection



RESULTS : AIC model : 7 lags
BIC model : 3 lags

<u>Fit vs Complexity</u> : As for the fisrt model, we chose the lag that minimize the BIC => avoid overfitting

Results and Model Comparison

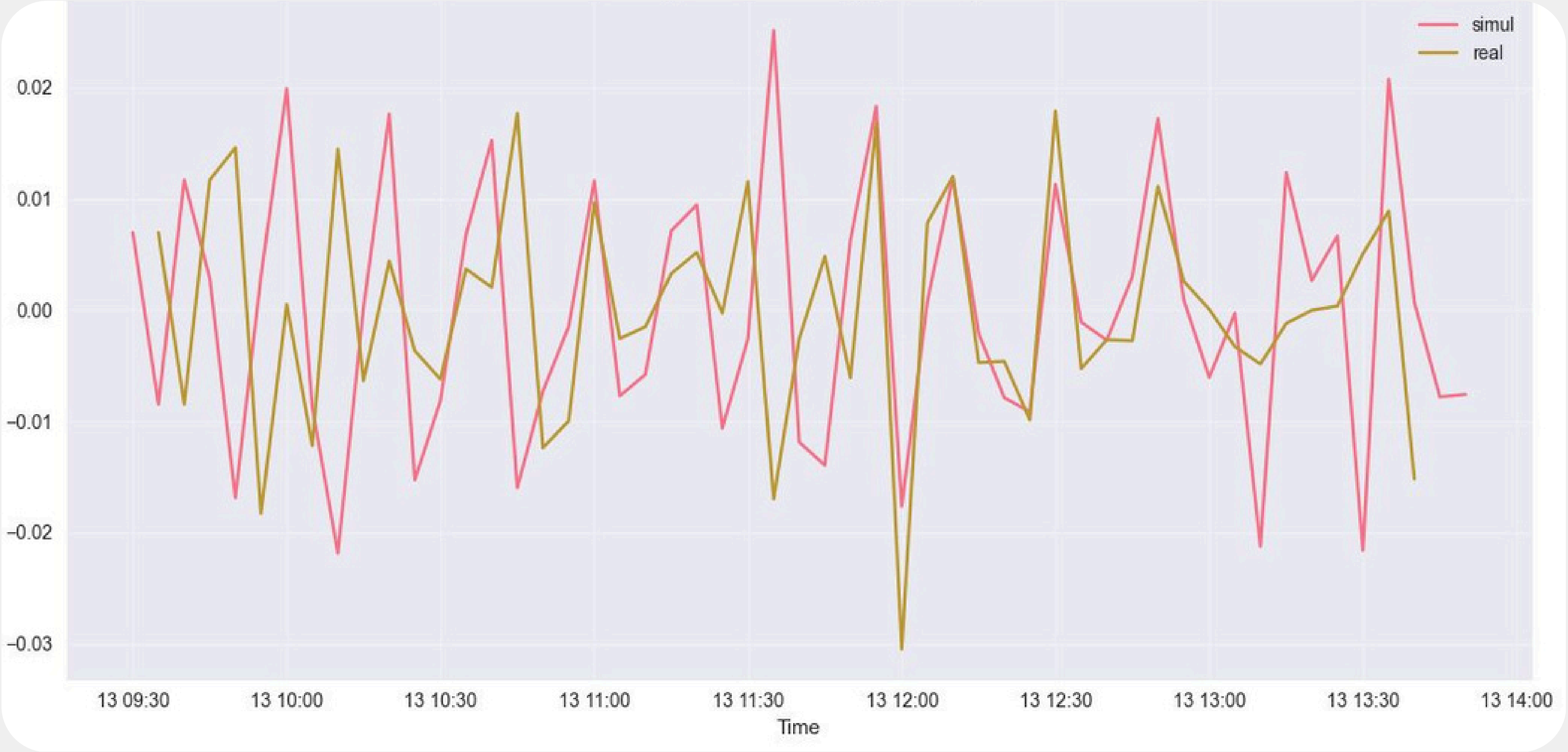
MARKET DYNAMICS

One-Day Fit - VARI Model

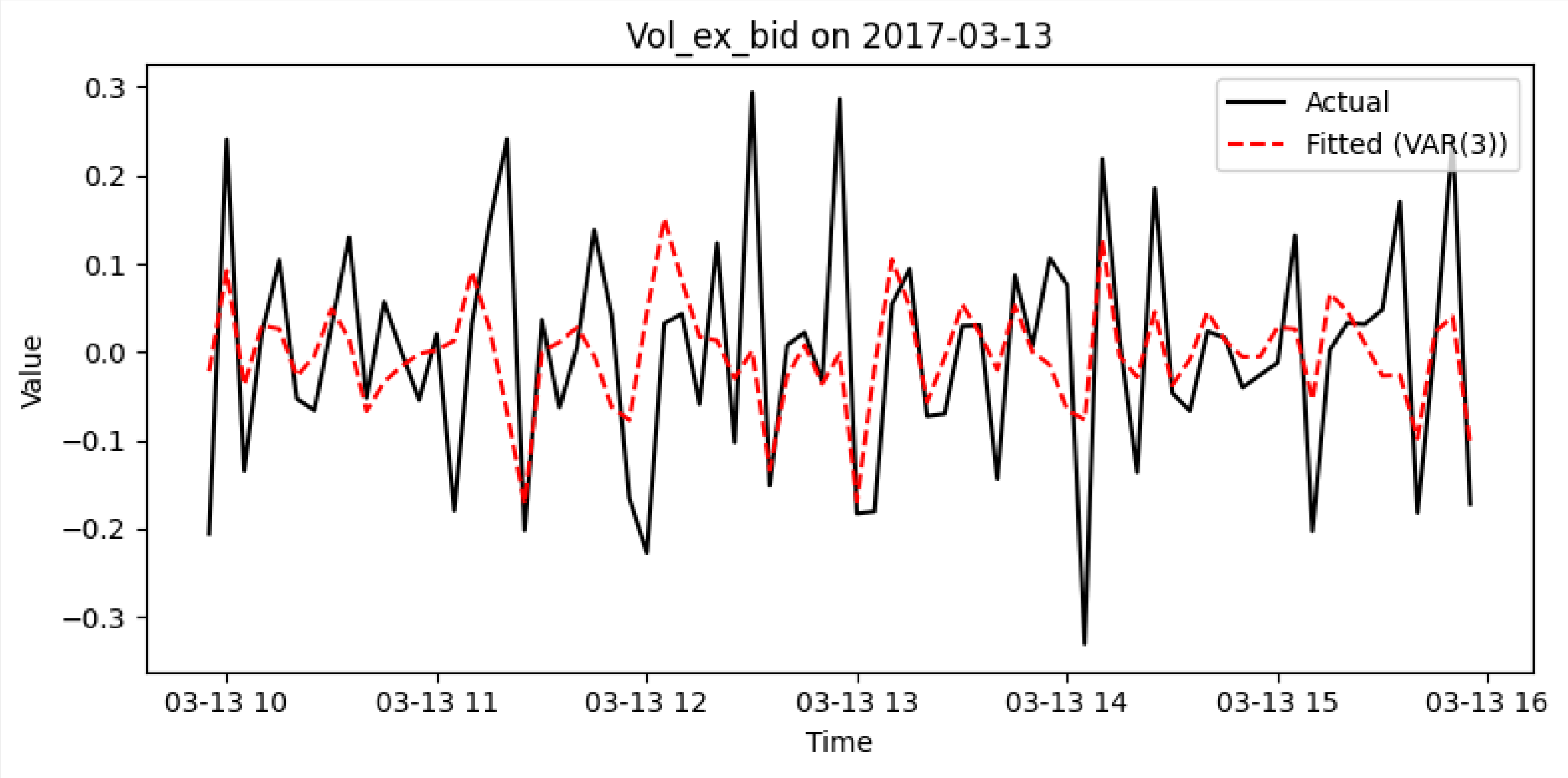
2 executions

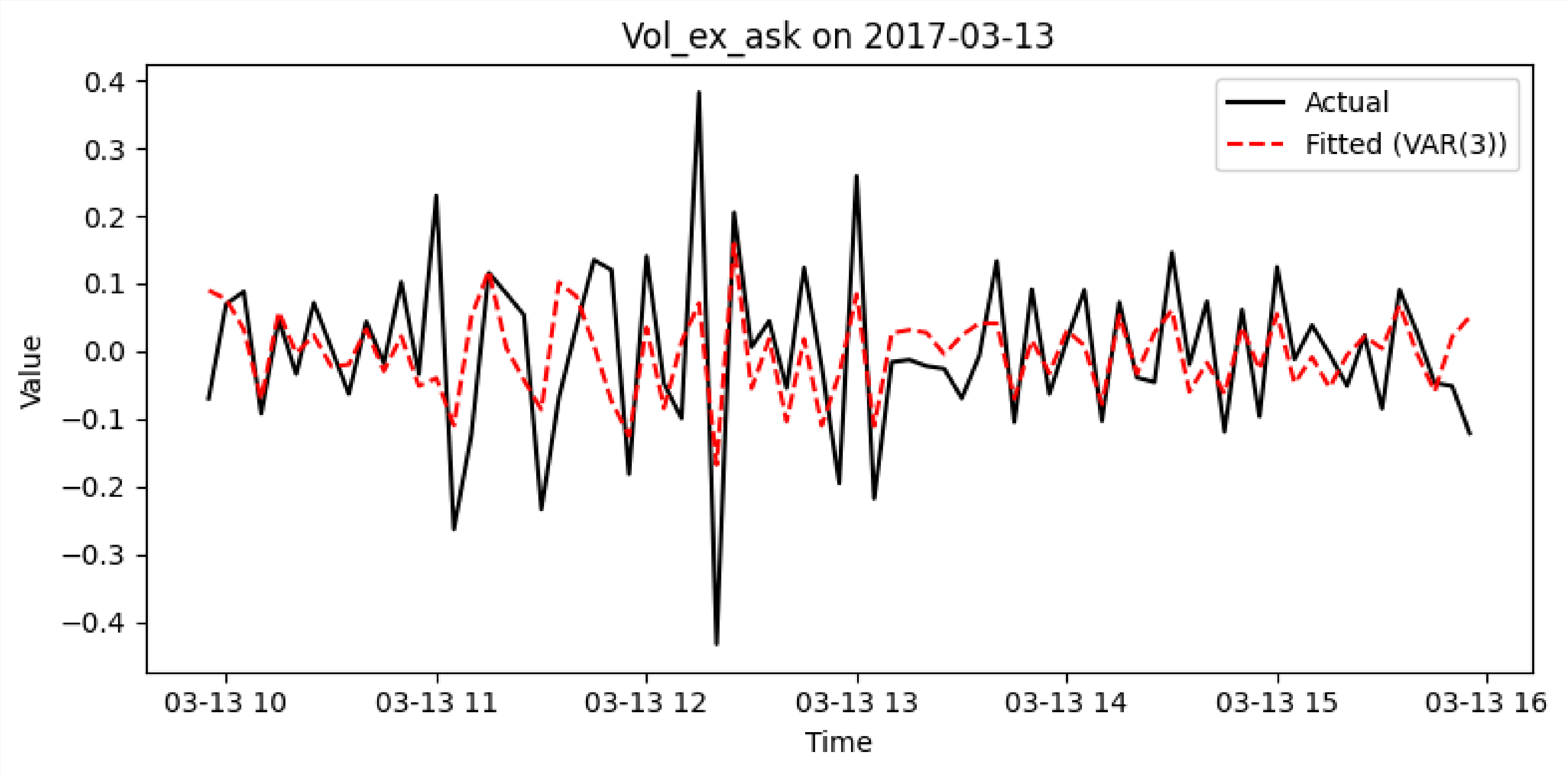
50-step VAR(3) simulation (5min per step)

50 times step VAR(3) simulation (5min per step)

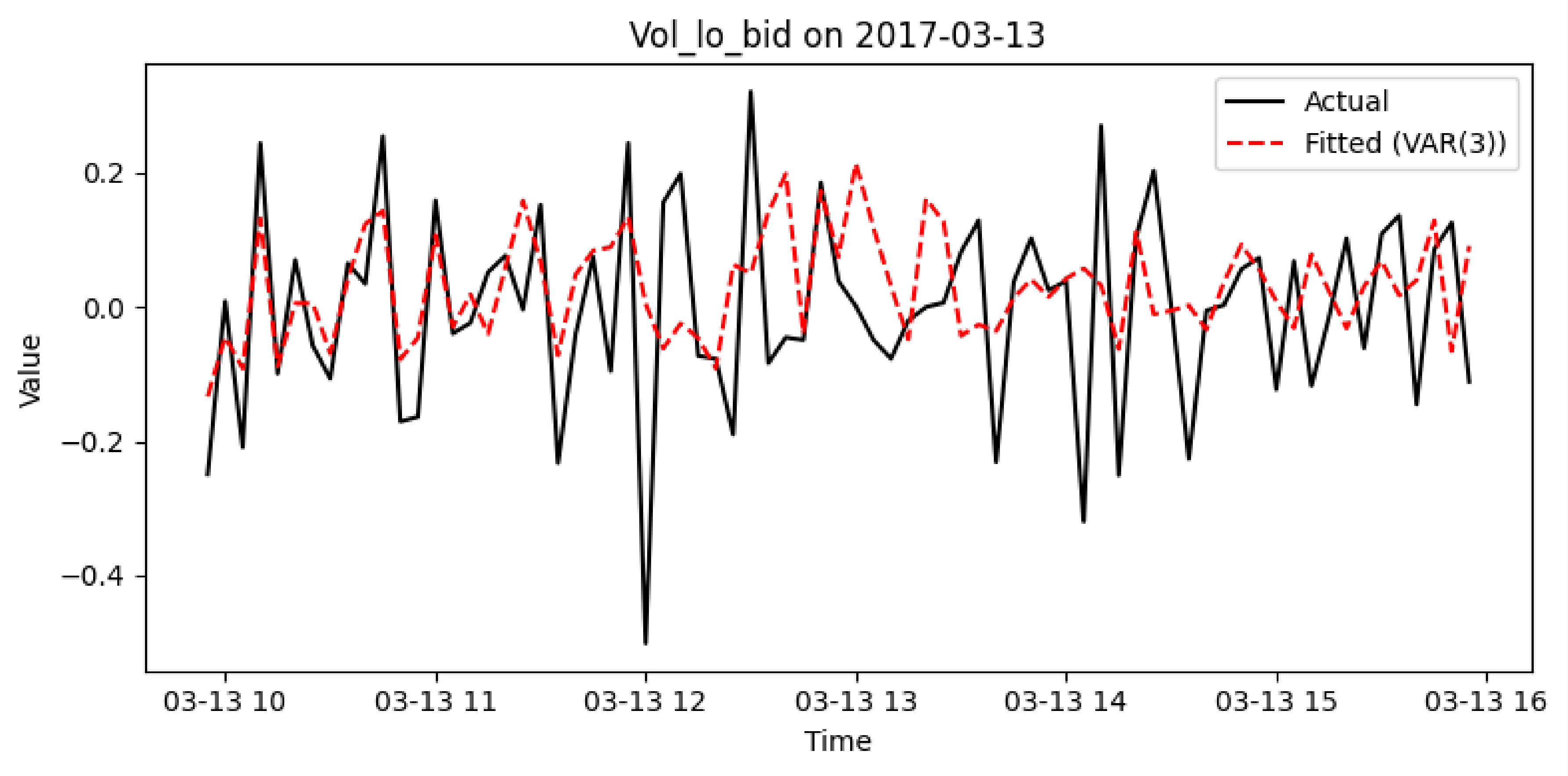


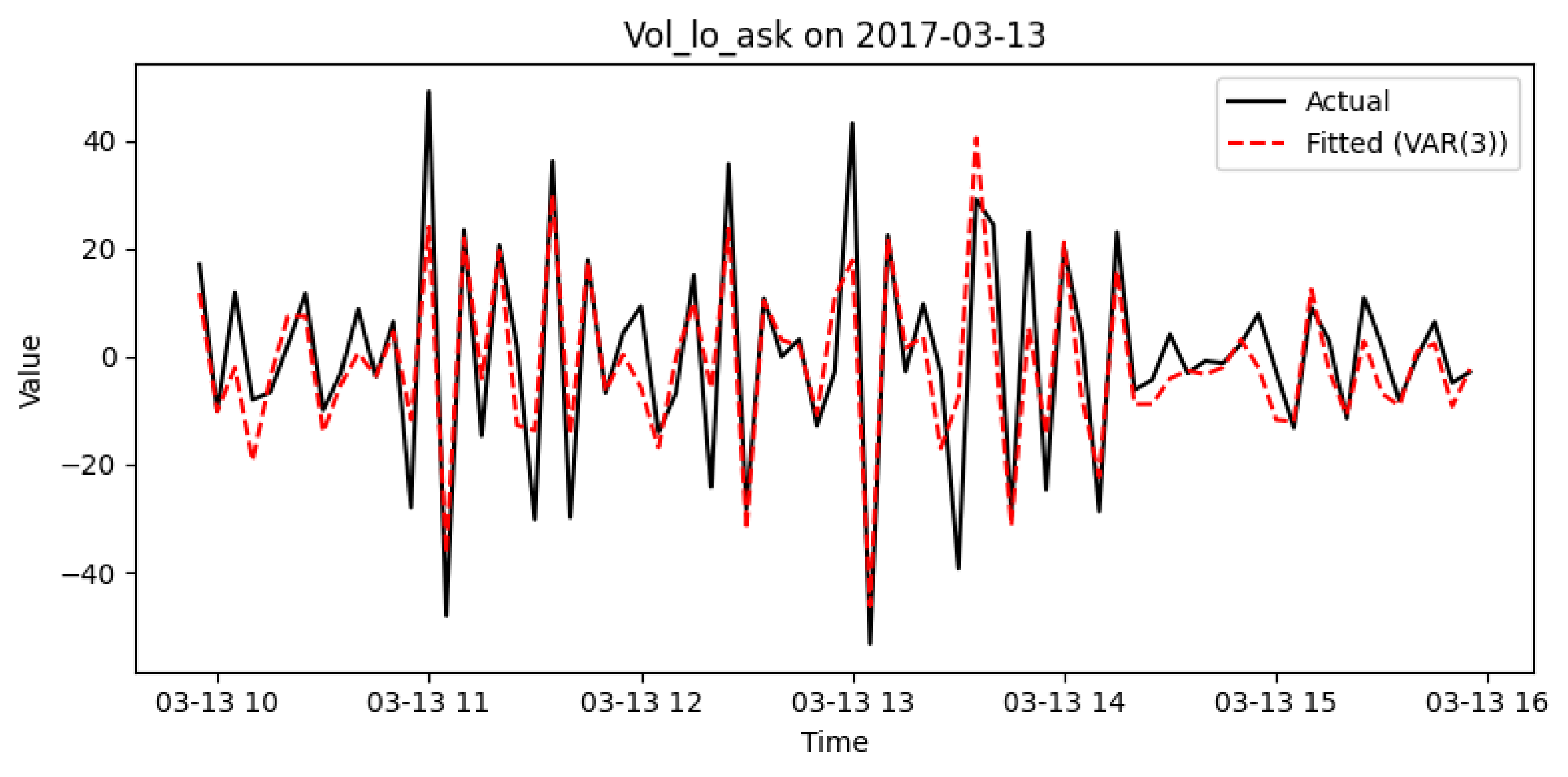
One-Day Fit - Box Cox Model



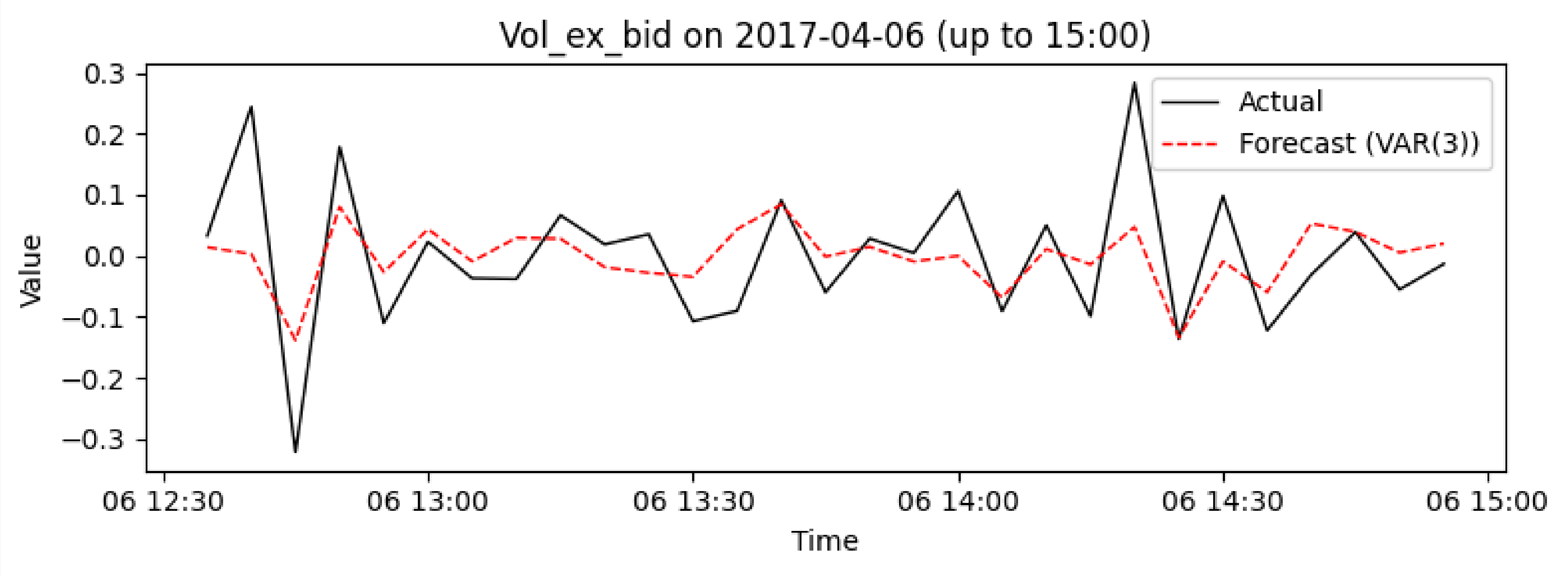


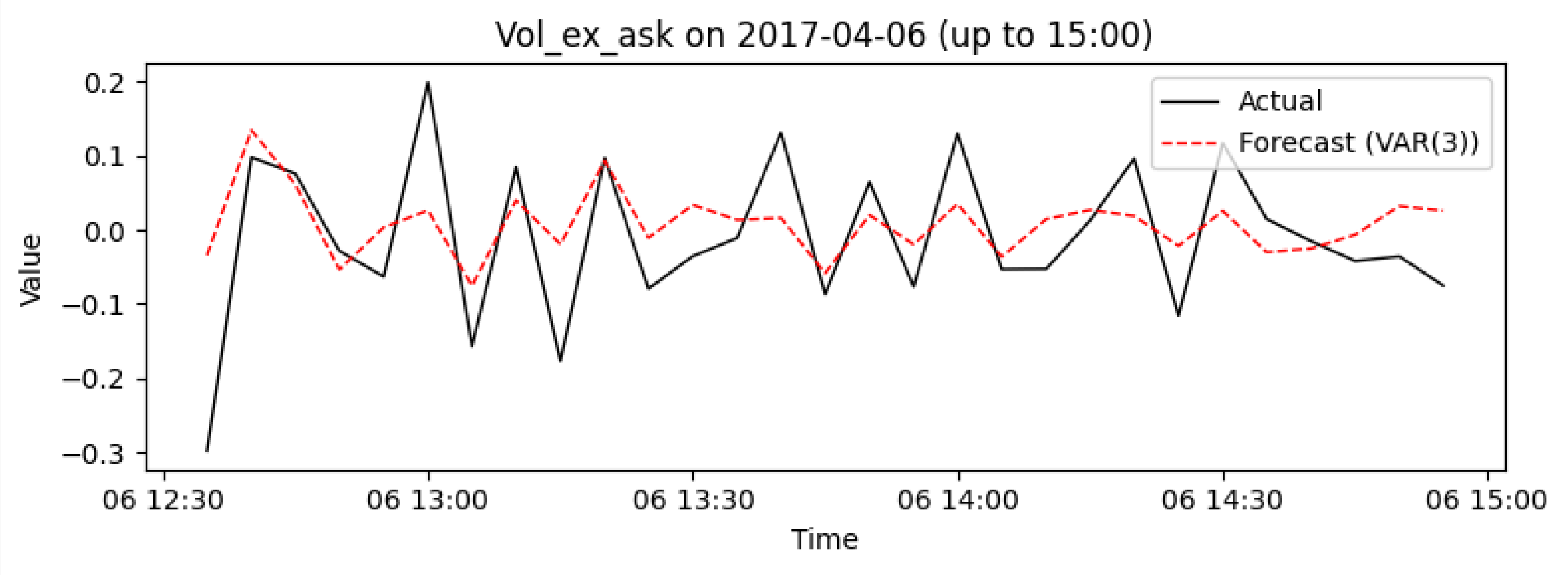
One-Day Fit - Box Cox Model

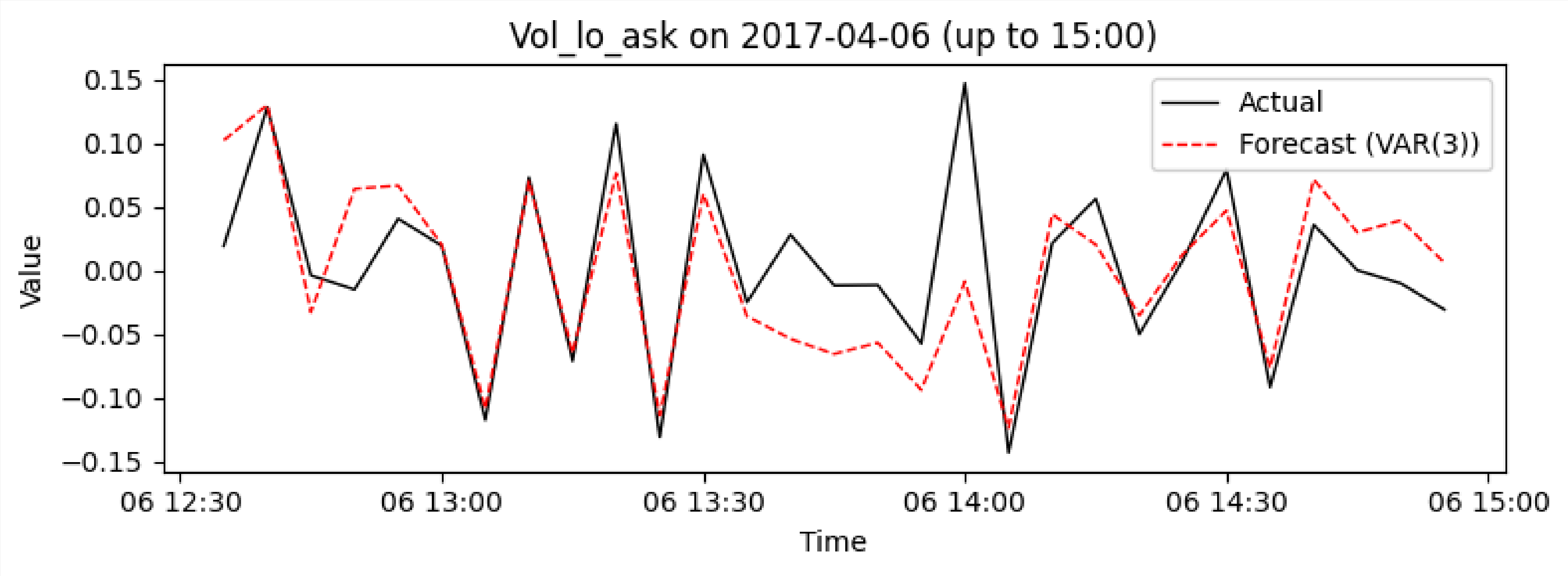




One-Day Prediction - Box Cox Model







- Captures accurately the direction of the flow
- Do not capture exceptional events
- Underestimate the magnitude of the peaks

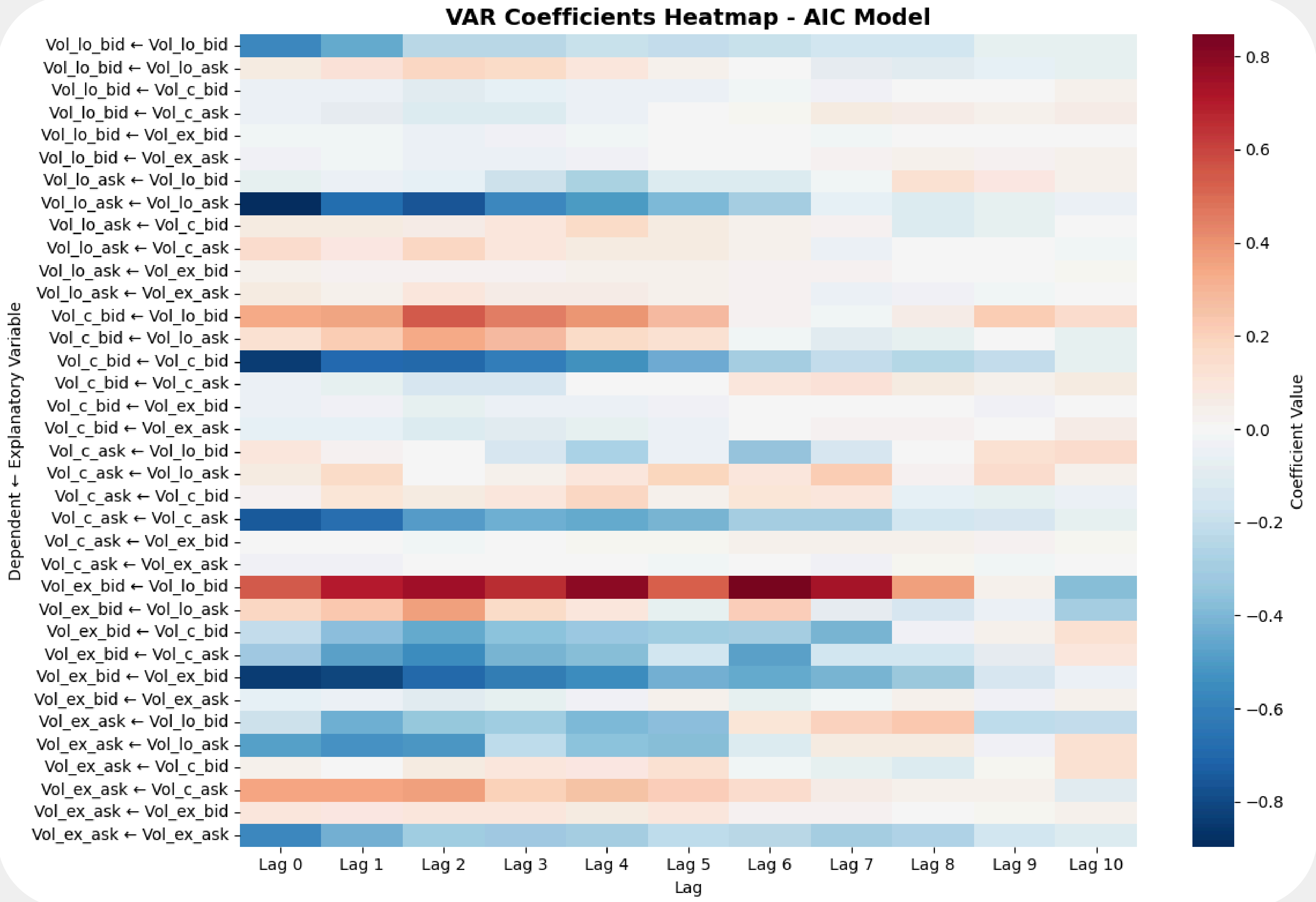
VARI Model Diagnostics

VARI coefficient matrices across lags for each explanatory variable

$$X(t) = A\,X(t-1) + \dots$$

$$\begin{pmatrix} \text{Vol\_lo\_bid} \\ \text{Vol\_lo\_ask} \\ \text{Vol\_c\_bid} \\ \text{Vol\_c\_ask} \\ \text{Vol\_ex\_bid} \\ \text{Vol\_ex\_ask} \end{pmatrix}_t = A \begin{pmatrix} \text{Vol\_lo\_bid} \\ \text{Vol\_lo\_ask} \\ \vdots \end{pmatrix}_{t-1} + \dots$$

$$\text{Vol\_lo\_bid}(t) = a_{11}\,\text{Vol\_lo\_bid}(t-1) + a_{12}\,\text{Vol\_lo\_ask}(t-1) + \dots$$

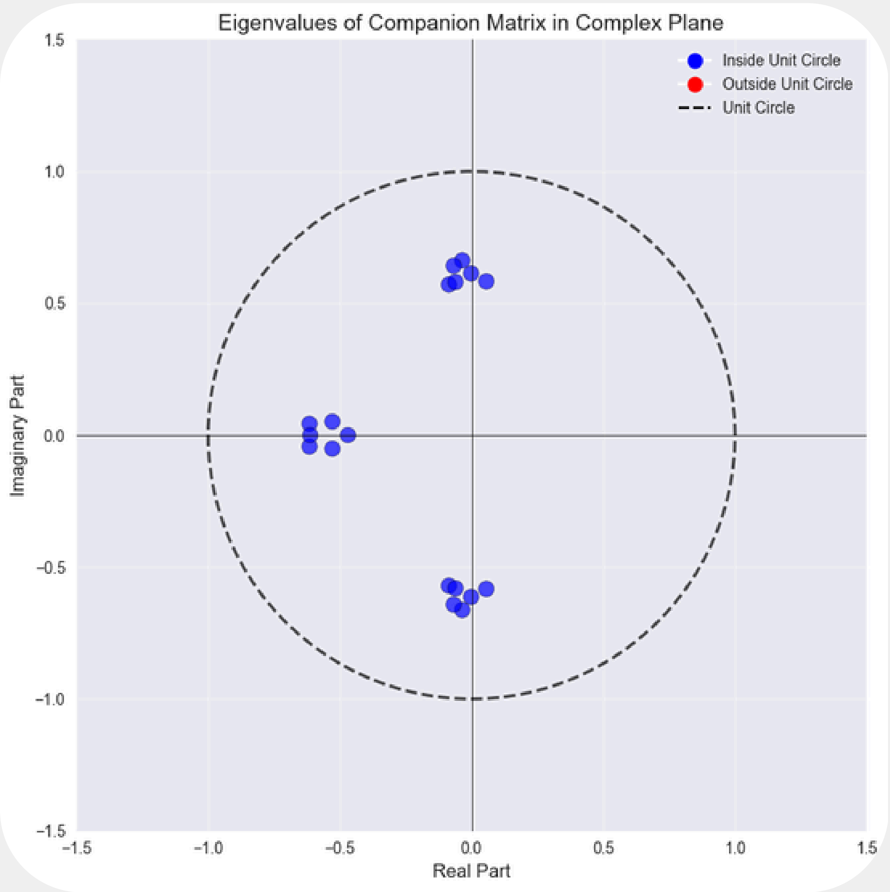


Models Stability

Stability of the VAR Model :

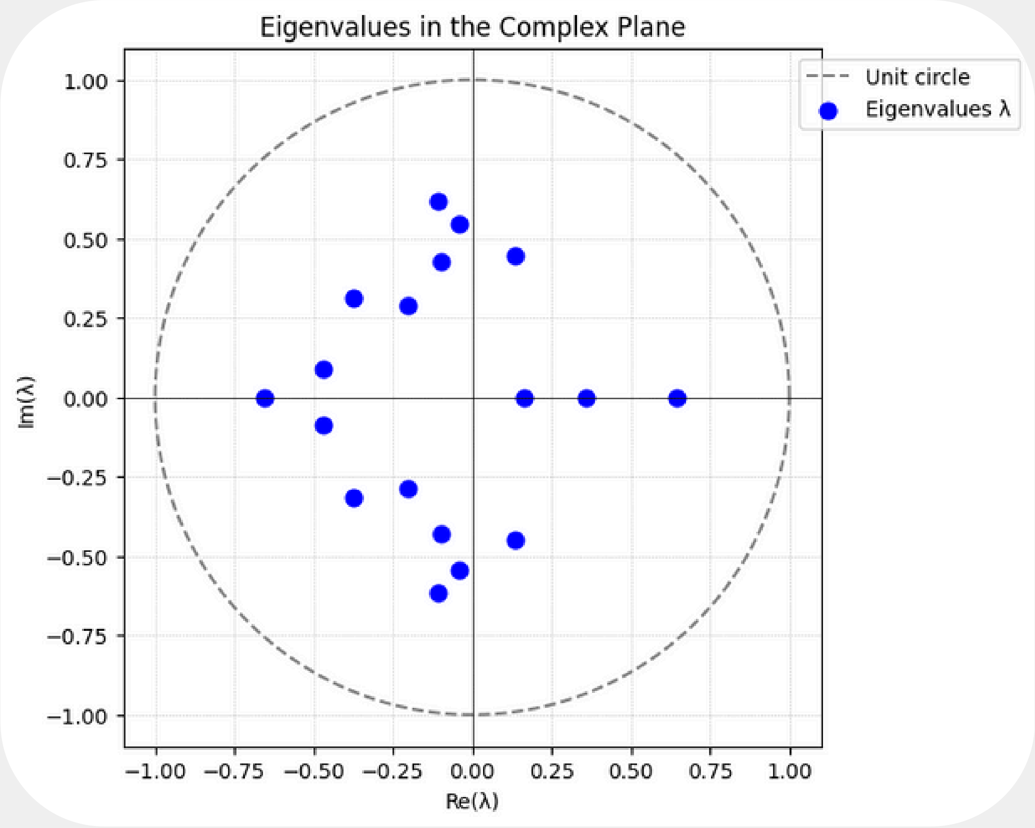
- Computing the eigenvalues of the Companion Matrix
- Checking that all eigenvalues are in the unit circle

Stability of VARI Model :



All eigenvalues are within the unit circle for both models => confirms their **stability**

Stability of Box Cox Model:



Assessing our Model's Prediction Accuracy

We use **MAE** (Mean Absolute Error) and **MASE** (Mean Absolute Scaled Error)

MAE measures the **average magnitude of the errors between predictions and actual values**.
It has the same unit as the values.
The **lower it is, the best the prediction.**

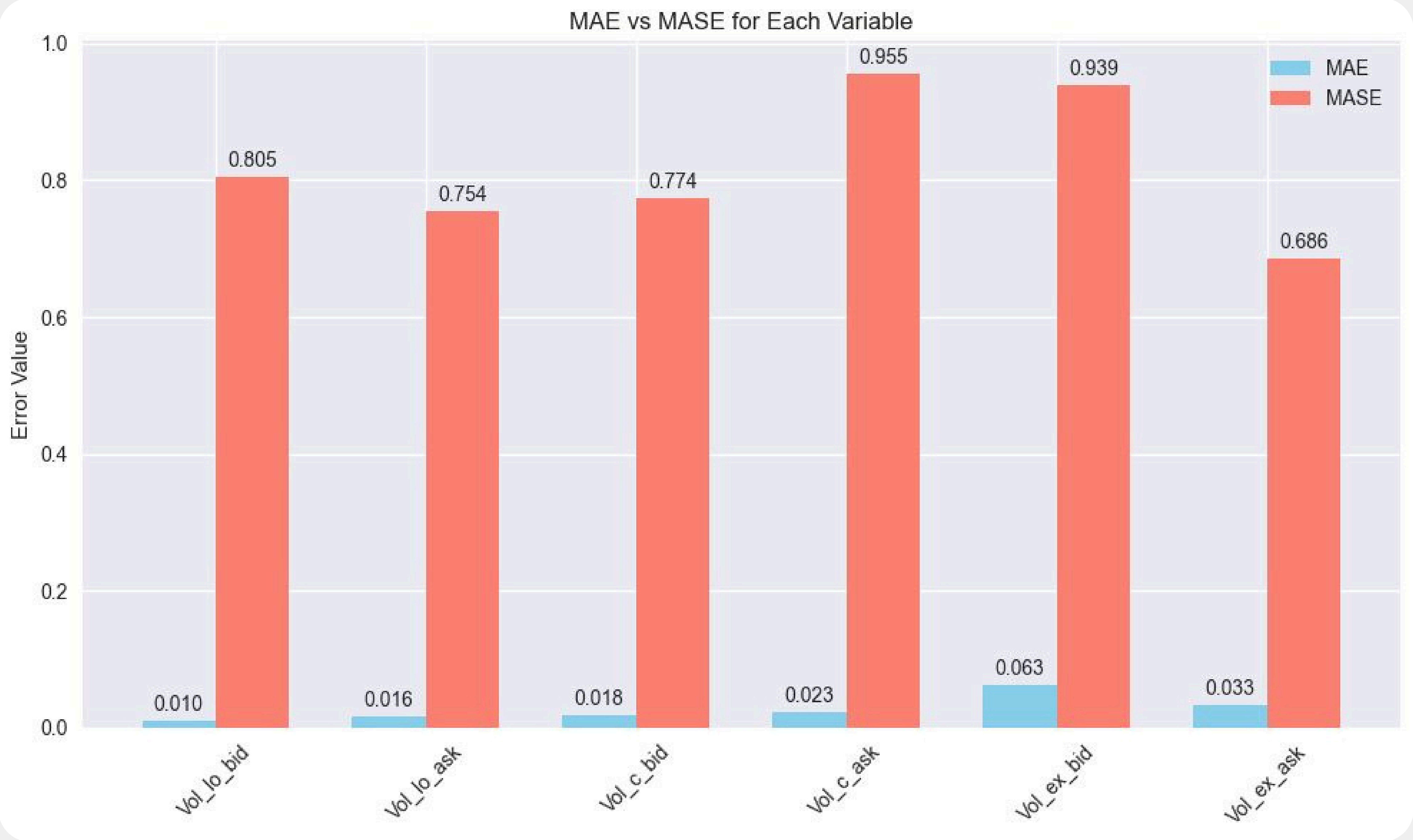
$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}|\hat{y}_t - y_t|$$

MASE scales MAE by the average error of a naïve forecast (e.g. using the previous actual value).

- If **MASE < 1, your model performs better than the naïve forecast.**
- If **MASE > 1, the naïve method outperforms your model.**

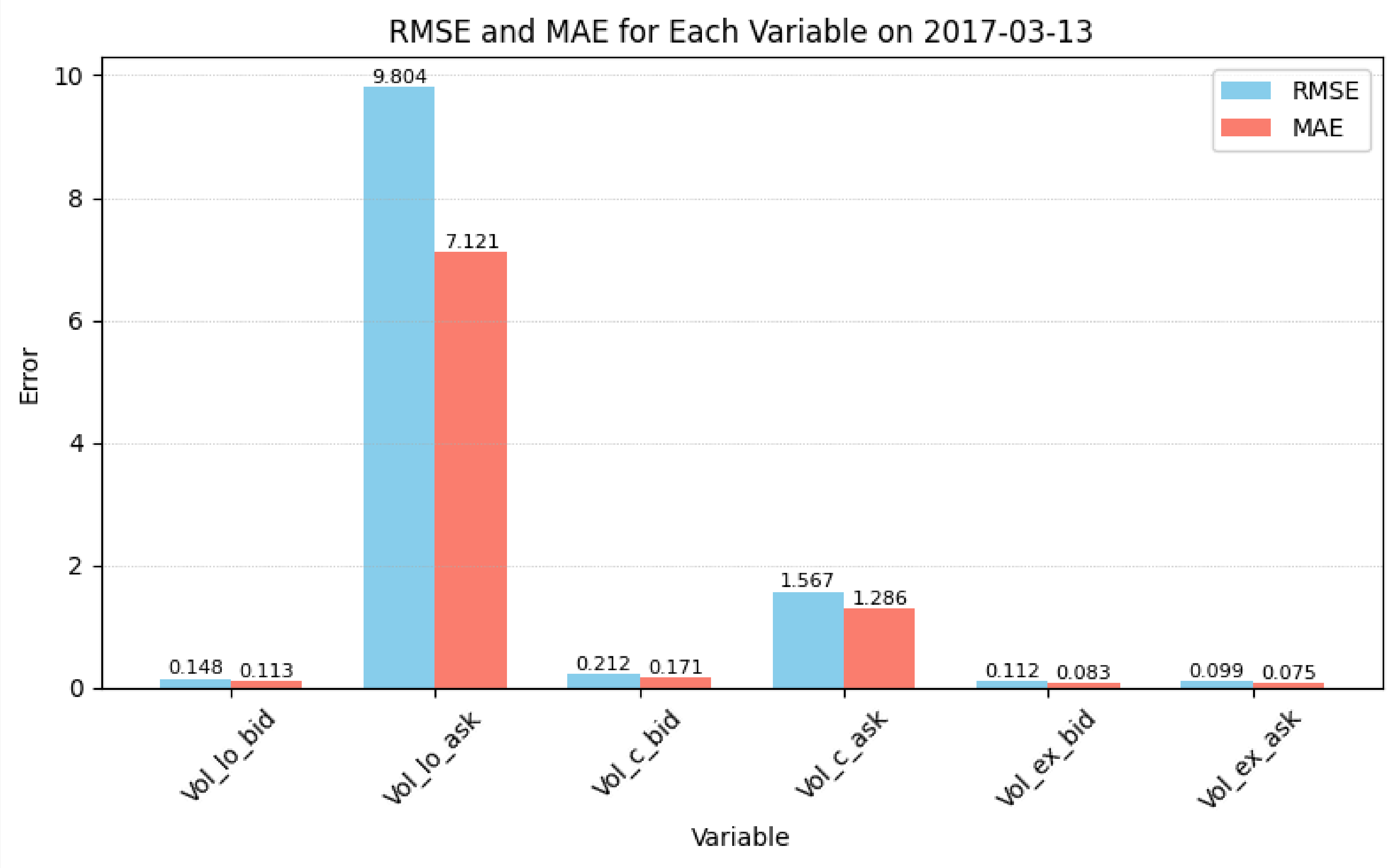
$$\text{MASE} = \frac{\text{MAE}}{\frac{1}{n-1}\sum_{t=2}^{n}|y_t - y_{t-1}|}$$

MAE and MASE criteria for VARI Model

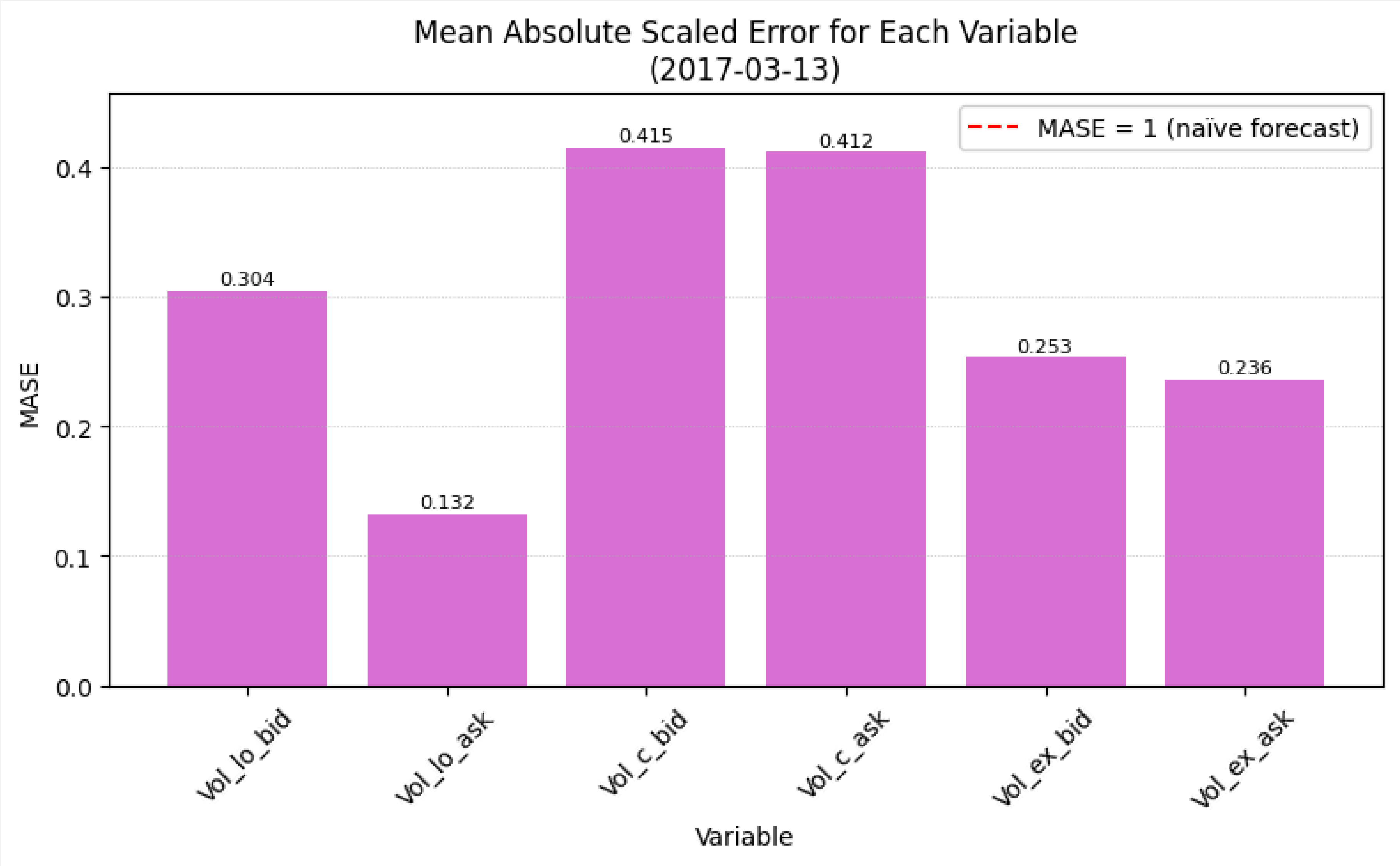


MAE and MASE criteria for Box-Cox transformation

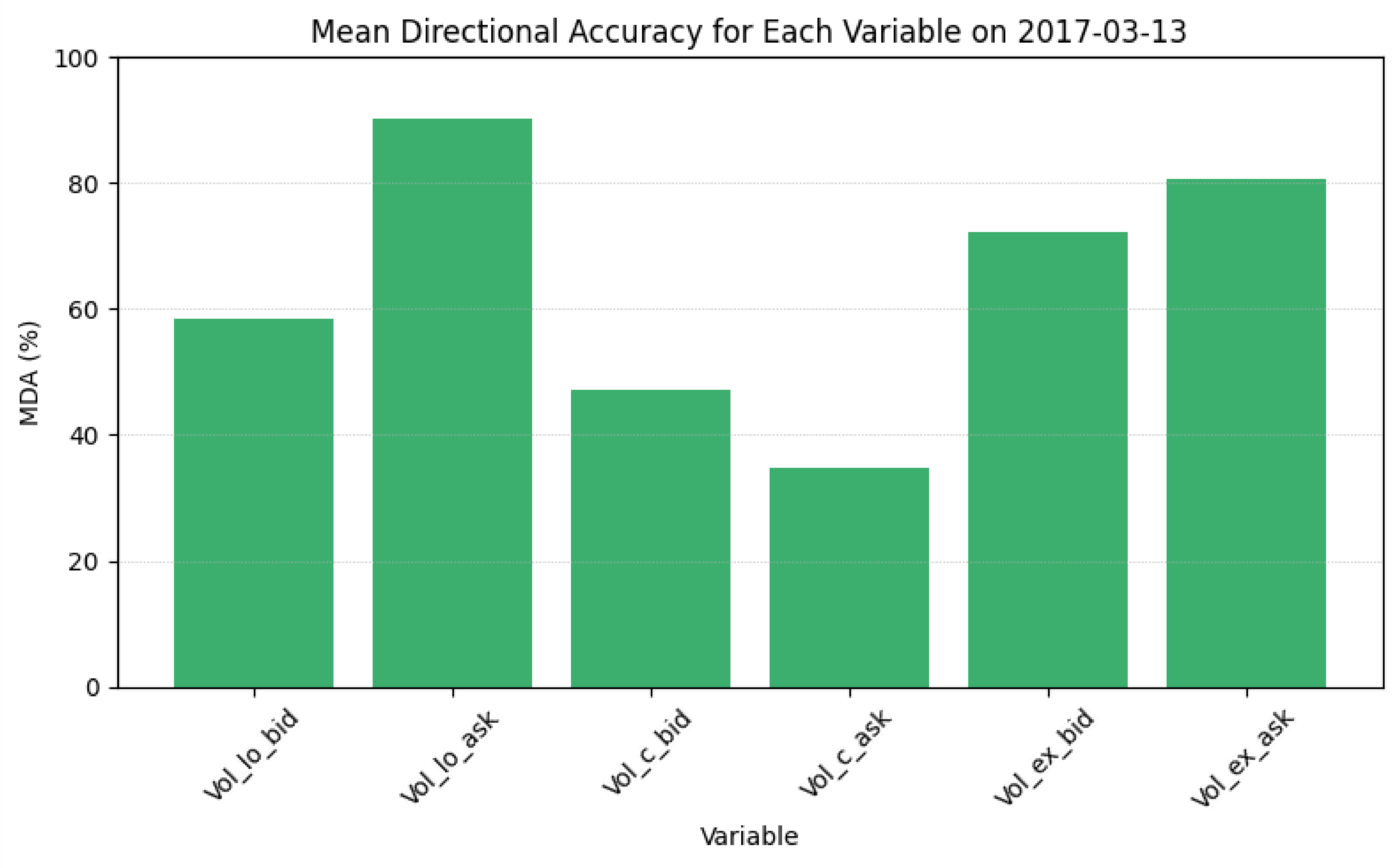
MAE :



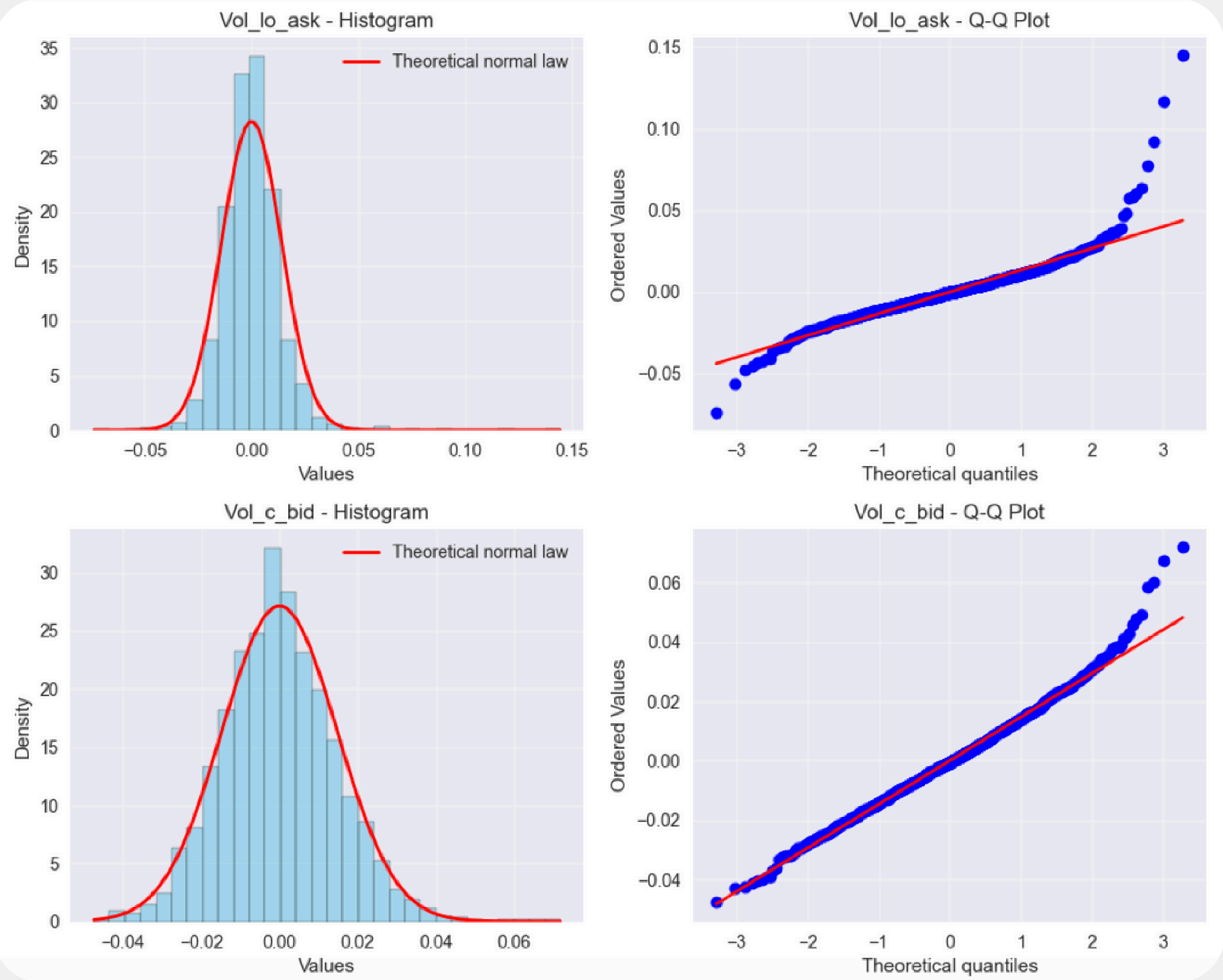
MASE :

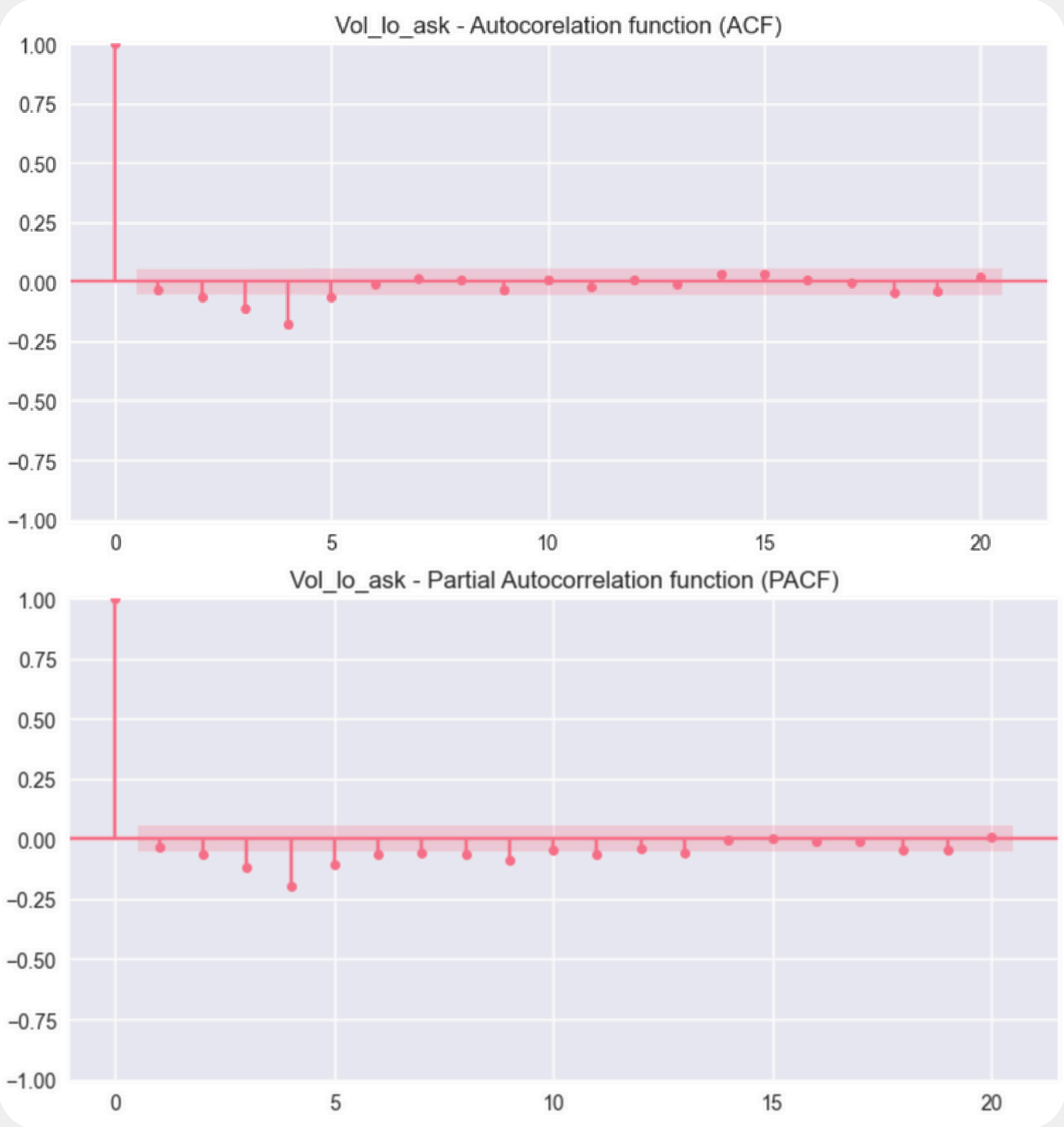


MDA criteria for Box-Cox transformation



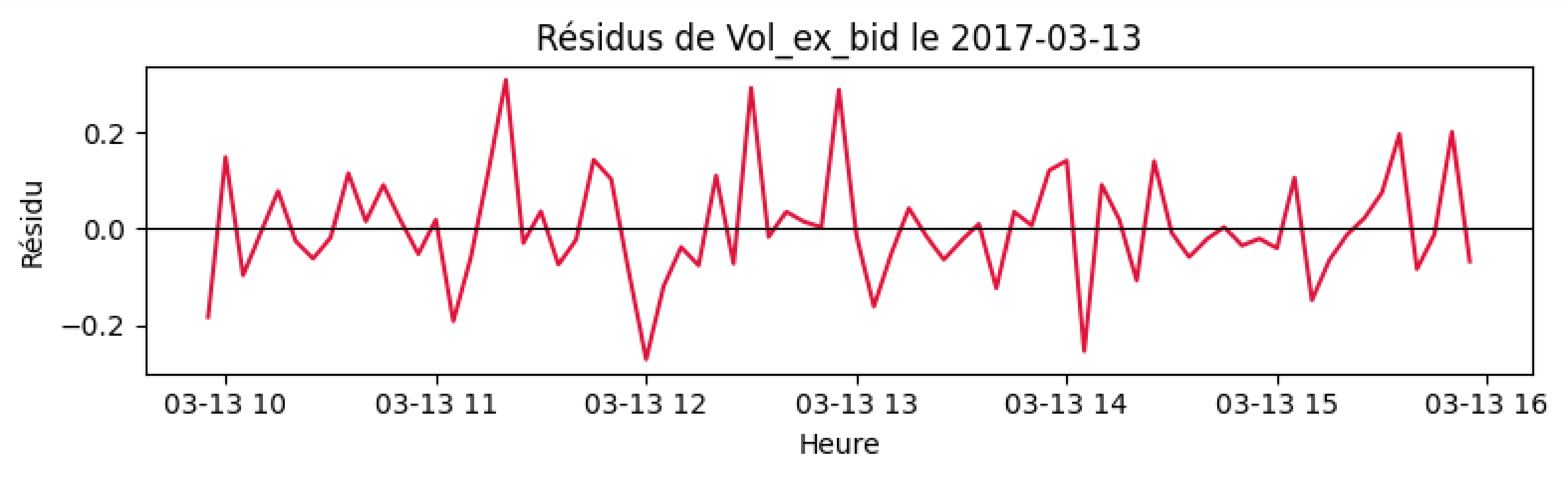
Residuals Analysis - VARI Model

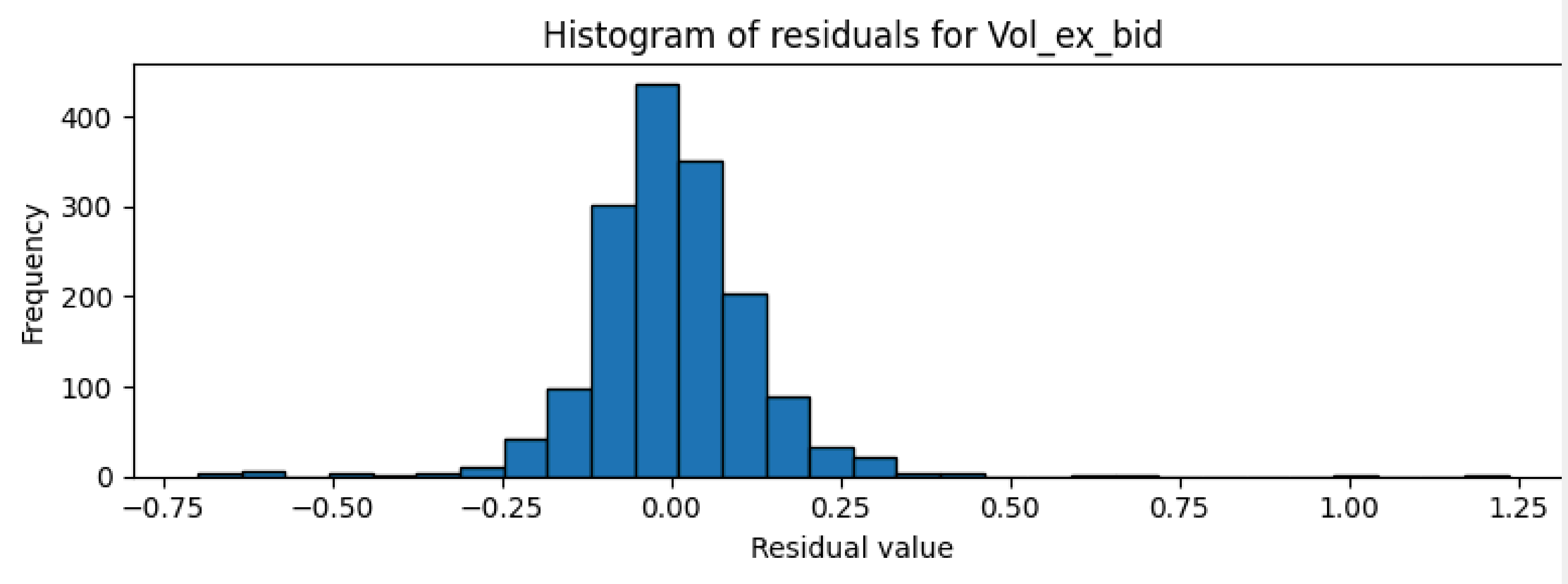




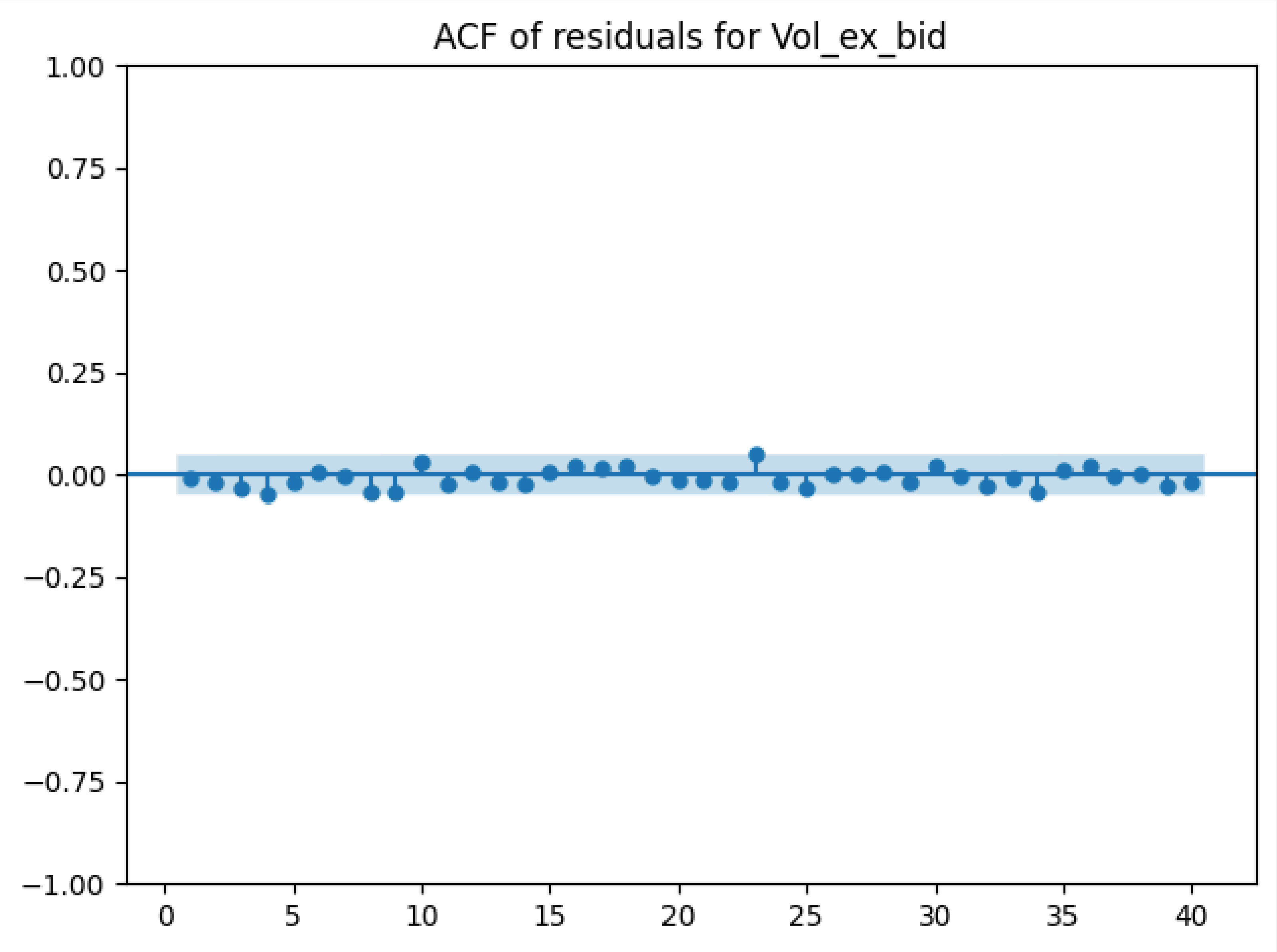
Residuals behave like a white noise in accordance with VARI model

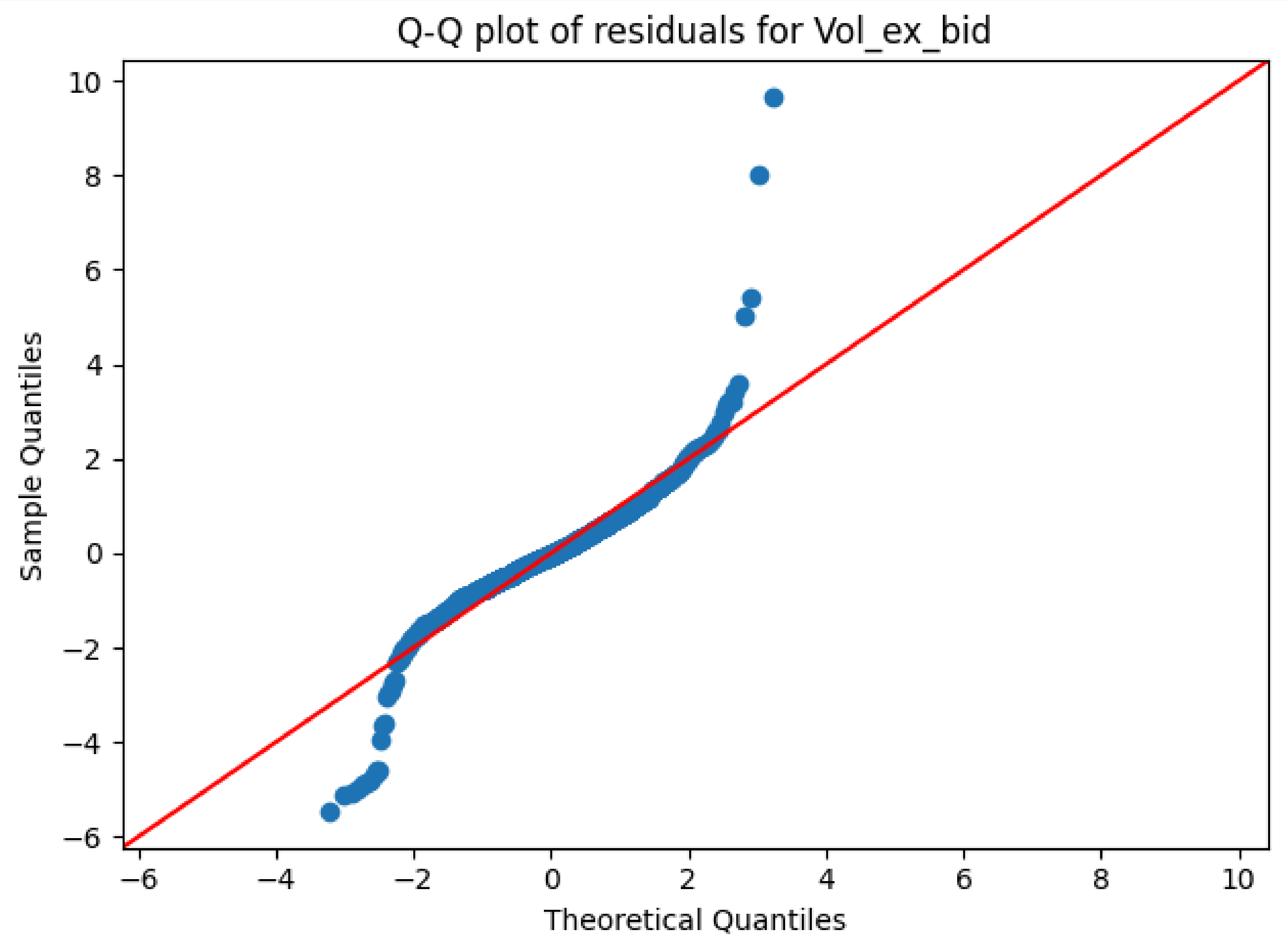
Residuals Analysis - Box Cox Model





Residuals Analysis - Box Cox Model





Conclusion

MARKET DYNAMICS

Conclusion

After our analysis and comparisons, we first concluded that our two models were better than a random prediction (MAE / MASE)

Moreover, we believe the best model we tested was Box-Cox's given its lower MAE, but it shows some aberrations for Vol Low Ask.

End

Thank you for listening