Time Series Data Processing and Modeling - Based on ARIMA

July 30, 2024

Abstract

A tap on Time Series Modelling.

1 Introduction

Time series data, which is a collection of data points arranged in chronological order, is a crucial form of data in many fields. This type of data originates from various applications, such as stock prices in financial markets, daily observations of economic indicators, meteorological data collection, and physiological signals in the biomedical field. The purpose of processing and modeling time series data is to uncover its inherent patterns, trends, and potential relationships, thereby providing insights into future developments. In this era of information explosion, the ability to effectively utilize time series data has become increasingly important.

There are various methods for modeling time series data, with foundational models including TensorFlow, Statsmodels, and Scikit-Learn. This series will focus on the ARIMA model, aiming to introduce the fundamental concepts, methods, and practices of time series data processing and modeling. It will explore the entire process from data cleaning and preprocessing to building predictive models, covering common statistical methods and the underlying statistical logic.

This series aims to clarify the following questions regarding time series:

1. What is a time series? What are its components? How can time series data be decomposed?
2. What is ”stationary” data? What are the characteristics of stationary data? Why is stationary data required for time series modeling? How can non-stationary data be transformed into stationary data?
3. What are ACF, PACF, and ARIMA?
4. What are the steps involved in modeling?
5. Practical implementation of models.

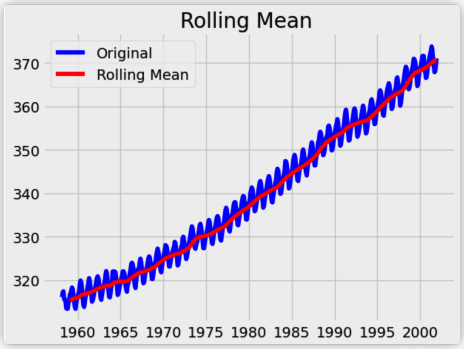


Figure 1: Why rolling mean is a good way to extract the trend?

2 Components and Decomposition of Time Series

Time series data consists of the following components.

$$Y = T + S + C + I,$$

where Y represents the time series data, T represents the trend, which refers to a tendency or state of a phenomenon that continues to develop and change over a long period, S represents seasonal fluctuations, which are regular changes in the level of a phenomenon due to seasonal variations, C represents changes in behavior (broadly speaking, C can also be viewed as part of the trend), I represents outlying observations.

Here, since there is no treatment group and no significant event impacts, we consider C=0.

In fact, from another perspective, we can consider time series data to be composed of: adding seasonal fluctuations S to a stationary random residual term, then adding trend T, and finally adding event C.

2.1 Decomposing the Trend

For the decomposition of a time series, the first step is to decompose the trend T (de-trend), which can be done using various methods, including linear regression, polynomial regression, and spline regression. The most commonly used methods are differencing and moving averages.

Moving average method. This is because the residuals are stationary, fluctuating around a mean value. Averaging can offset randomness to some extent and can also cancel out the seasonal factors of cyclical fluctuations. Therefore, after applying the moving average, we can consider the result to contain only the T and C components, that is, the total trend component.

Differencing.In fact, among T, S, C, and I, only T is related to time t (S is cyclical and is not

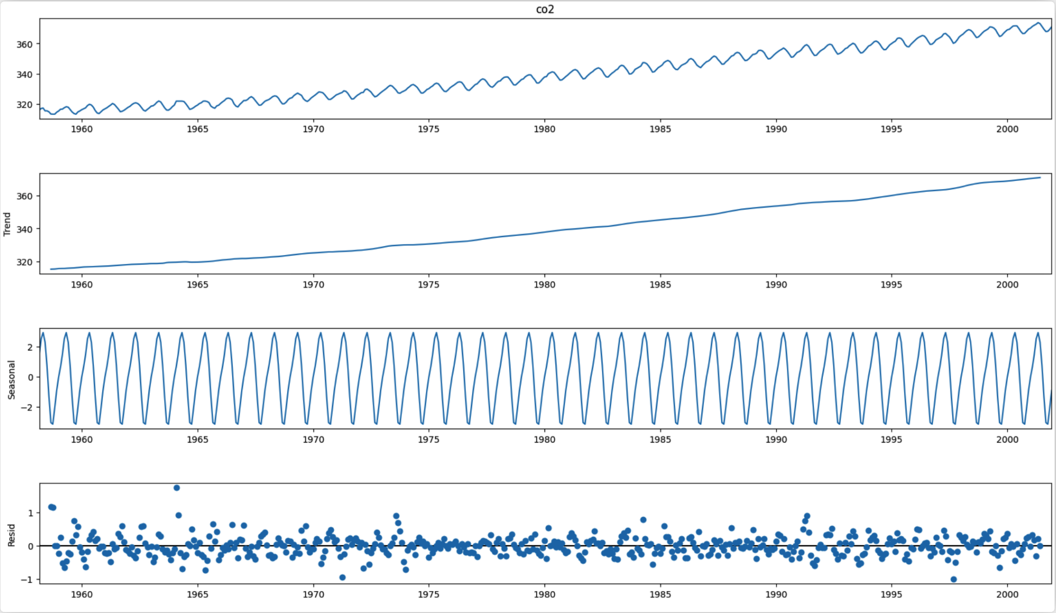


Figure 2: Decomposed time series

concerned with the specific year or month). Decomposing the trend means isolating the terms that are related to time t. Under the assumption that the trend and time t are linearly related, we have:

$$EX_t = C_t + a \times t, EX_{t-1} = C_{t-1} + a \times (t-1),$$

$$\therefore EX_t - EX_{t-1} = C_t - C_{t-1} + a$$

After decomposition, there are no longer terms related to time t; the trend has been extracted.

2.2 Decomposing Seasonality

After decomposing the trend, we are left with the S and I terms, and then we proceed to decompose the seasonal component (de-seasonality). For the seasonal component, we have $S_t = S_{t-d}$, where d is a cycle. The most commonly used decomposition method is differencing, which performs first-order differencing every d time intervals.

The final decomposition result is shown in the figure.

After gaining a general understanding of the data table (including the columns, their data types, the number of records, and the number of unique values), we can begin building the model.

3 Stationary Data and Stationarized Data

3.1 Definition of Stationary Data

Stationary data refers to data whose statistical properties remain constant over time. If a time series is stationary, its basic statistical characteristics, such as mean and variance, should remain stable across different time periods. Stationarity is an important assumption in time series analysis because many time series models are based on the premise that the data is stationary. Stationarity can be divided into strong stationarity and weak stationarity.

3.1.1 Weak Stationarity

$$Let \quad \mu_x(t) = E[X_t], Y_x = Cov(X_r, X_s),$$

$$\begin{cases} \mu_x(t) \quad is \quad IND \quad of \quad t, \\ Y_x(t+h, t) \quad is \quad IND \quad of \quad t \quad for \quad each \quad h \end{cases}$$

For weak stationarity, the mean and variance are constants, and the covariance at the same time point is also a constant. That is, although the entire probability distribution does not need to be identical, the basic statistical properties remain unchanged. At this point, the residuals do not contain information about time t, but they do contain information about the time interval h.

3.1.2 Strong Stationarity

$$Let \quad \mu_x(t) = E(X_t), Y_x = Cov(X_r, X_s), F(\cdot) be \quad the \quad cdf \quad of \quad X,$$

$$F_{X_{t1}, X_{t2}, \cdots, X_{tn}}(X_1, X_2, \cdots, X_n) = F_{X_{t1+h}, X_{t2+h}, \cdots, X_{tn+h}}(X_1, X_2, \cdots, X_n), \quad \forall t, h$$

The entire probability distribution remains unchanged over time. For any t and s(where $t \neq s$), the joint distribution of the data is the same. If the residuals are strongly stationary, then the residuals contain no information related to time (neither related to time t nor to the time interval h).

3.2 Conditions for Stationarity

The conditions for determining weak stationarity are:

$$\begin{cases} \mu_x(t) \quad is \quad IND \quad of \quad t, \\ Y_x(t+h, t) \quad is \quad IND \quad of \quad t \quad for \quad each \quad h \end{cases}$$

The conditions for determining strong stationarity are:

$$F_{X_{t1},X_{t2},\cdots,X_{tn}}(X_1, X_2, \cdots, X_n) = F_{X_{t1+h},X_{t2+h},\cdots,X_{tn+h}}(X_1, X_2, \cdots, X_n), \quad \forall t, h$$

In time series data, we need stationary residuals because we must assume that the residuals (the remaining items after decomposing the trend, seasonal fluctuations, and behavior changes) do not vary over time; otherwise, we cannot make appropriate predictions about the future. Again, it is important to note that stationary residuals are the foundation of time series data modeling. We must first ensure the stationarity of the residuals before beginning the modeling process.

For non-stationary data, there may be trends (the overall direction that changes over time), seasonality (cyclical changes, such as seasonal factors at certain times of the year), and cyclicality (non-fixed length cyclical factors), among others. When conducting time series modeling or analysis, it is usually necessary to first test and process the data for stationarity. This can be achieved through differencing methods.

4 What are ACF, PACF, ARMA, and ARIMA?

ACF (AutoCorrelation Function) is the autocorrelation function, which is used to measure the correlation of a time series with itself at different time points. It can help identify repeating patterns and seasonality within the time series.

$$ACF: \quad \rho_x(h) = \frac{Y_x(h)}{Y_x(o)}$$

$$SampleACF: \quad Y_x\hat{(h)} = \frac{1}{n}\sum(x_{t+h} - \bar{x})(x_t - \bar{x})$$

PACF (Partial AutoCorrelation Function) is the partial autocorrelation function. PACF measures the autocorrelation between two different time points after removing the influence of the intermediate time points. It can help determine the direct relationship between the lagged terms of a time series. PACF is typically used to identify the order of autoregressive models.

$$AR(p): \quad X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + z_t$$

$$PACF(q) = \phi_q \quad q \leq p$$

4.1 Two Types of Autoregressive Methods: AR and MA.

$$AR(p): \quad X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + z_t \quad \phi_t \neq 0 \quad (1)$$

$$MA(q): \quad X_t = \mu + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots - \theta_q\varepsilon_{t-q} \quad \theta q \neq 0 \quad (2)$$

The AR model (Autoregressive Model) describes the relationship between the current observation and several past observations, meaning that the current value is a linear combination of past values. It is typically denoted as AR(p), where p is the order of the lag. The coefficients ϕ in the model represent the weights of each lagged term, and c is a constant, while the error term z_t is noise.

The MA model (Moving Average Model) describes the relationship between the current observation and several past white noise error terms, meaning that the current value is a linear combination of past error terms. It is typically denoted as MA(q), where q is the order of the moving average. The coefficients θ_j in the model represent the weights of each error term, μ is the mean, ϵ_j is the white noise error, which refers to the residuals after decomposing the time series.

These two models can be combined to form the ARMA model, while differencing can be introduced to form the ARIMA model. These models are commonly used in time series analysis to describe and predict the dynamics and trends of data.

When ACF shows a tail while PACF does not, an AR model should be used for residual modeling. When PACF shows a tail while ACF does not, an MA model should be used for residual modeling. If both show tails, then the ARMA and ARIMA models described below should be used for modeling. This is because, when the ACF shows a tailing off pattern, it means that the current value of the time series is correlated with its past values, and this correlation gradually weakens over time. This pattern often suggests that the data has long-term autocorrelation, which is suitable for modeling with an AR model.

While when the PACF quickly drops to zero after a certain lag, it indicates that the past values significantly influence the current value only up to that lag, and beyond that, their influence is negligible. A non-tailing PACF typically suggests that the order of the model (p) is limited.

4.2 ARMA Model (AutoRegressive Moving Average model)

The ARMA model (AutoRegressive Moving Average model) combines the AutoRegressive (AR) and Moving Average (MA) models. The AR part represents the relationship between the current value and past values, while the MA part represents the relationship between the current value and the random error terms. It is commonly used for modeling stationary time series.

$$ARMA(p,q): \quad X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \mu + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots - \theta_q\varepsilon_{t-q} + z_t$$

4.3 ARIMA Model (AutoRegressive Integrated Moving Average model)

If the residual series is non-stationary, the ARIMA model (AutoRegressive Integrated Moving Average model) should be used. The ARIMA model adds a differencing step to the ARMA model to handle non-stationary time series. In the ARIMA(p, d, q) model, p is the order of the AR part, d is the number of differences, and q is the order of the MA part.
Parameter Selection: The selection of p and q is generally done by observing when the ACF and PACF converge. For the selection of d, one can perform an ADF (Augmented Dickey-Fuller) test for stationarity on the differenced series or observe the ACF and PACF. When the differencing order is sufficient, both ACF and PACF should abruptly stop at lag 0.

5 Modeling Steps

Since time series data can be thought of as consisting of a stationary random error term with seasonal fluctuations (S), plus a trend (T), and then plus an event (C), the logic behind modeling is to first decompose the trend and seasonal fluctuations, model the remaining residuals, and then add the trend and seasonal components back in to form a complete prediction.

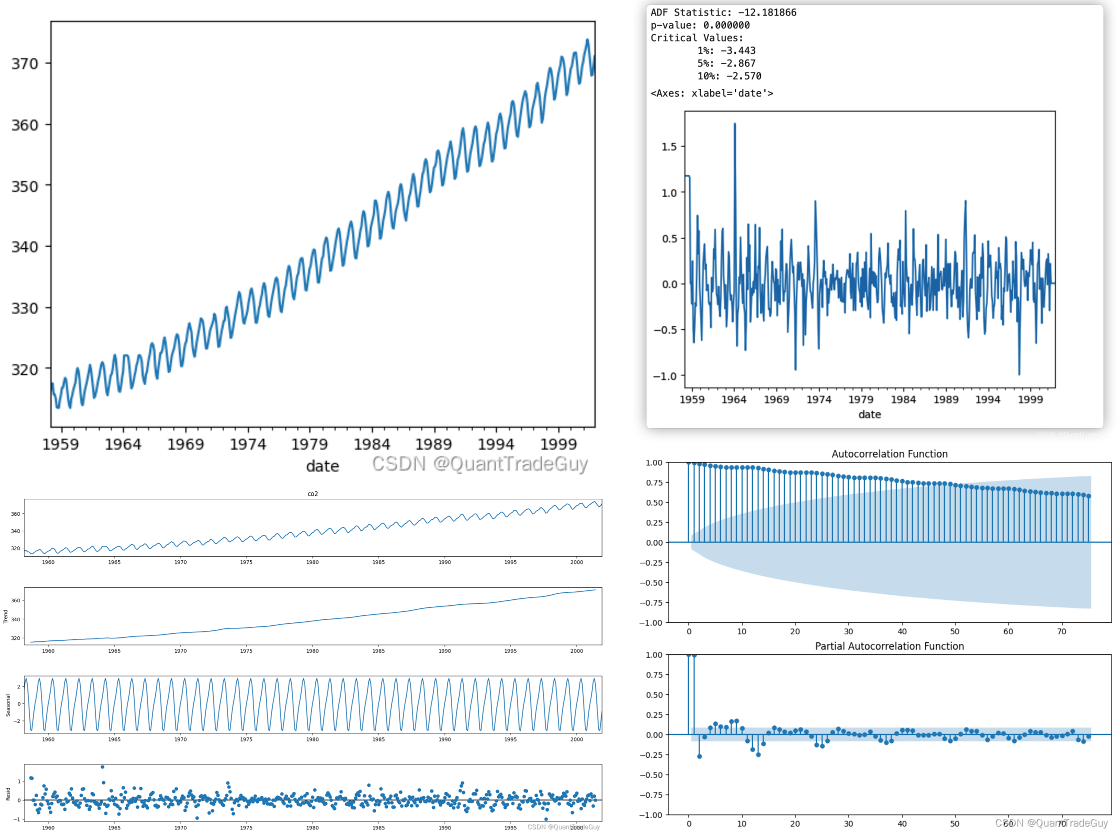
1. Preprocessing: After obtaining the data, first check for missing values. Missing values are crucial for the accuracy of time series predictions. If missing values are present, you can consider forward/backward filling, or resampling the data by time (e.g., resampling daily data to monthly data).

2. Decomposing Trend and Seasonality: After preprocessing the data, decompose the time series into its trend and seasonal components.

3. Observing Residuals: Check the remaining residuals for missing values, then perform an ADF (Augmented Dickey-Fuller) test to check for stationarity. Since the null hypothesis of the ADF test is that the time series has a unit root (i.e., it is non-stationary), if the test results reject the null hypothesis, the residuals are stationary. If the residuals are non-stationary, apply differencing to obtain the differencing order d. Observe the ACF and PACF of the residuals to determine q, p.

4. Using the ARIMA Model and Making Predictions: Implement the ARIMA model using the identified parameters and make predictions.

5. Evaluation: The model's performance can be evaluated using metrics such as Mean Squared Error (MSE) or trend consistency.



```
ADF Statistic: -12.181866
p-value: 0.000000
Critical Values:
        1%: -3.443
        5%: -2.867
        10%: -2.570
<Axes: xlabel='date'>
```

(a) After resampled and decomposed

(b) ACF, PACF and residual

6 Project 1: Carbon Dioxide Concentration Prediction

1. Import and examine the data, and there are 59 missing values.
2. Next, resample the data.
3. Decompose trend and seasonality.
4. Conduct a stationarity test on the decomposed residuals and observe the ACF and PACF.
5. Evaluation: The model's performance can be evaluated using metrics such as Mean Squared Error (MSE) or trend consistency.
6. Train the model. Based on previous observations, set p=q=1 and take the first-order difference.
7. Predict.

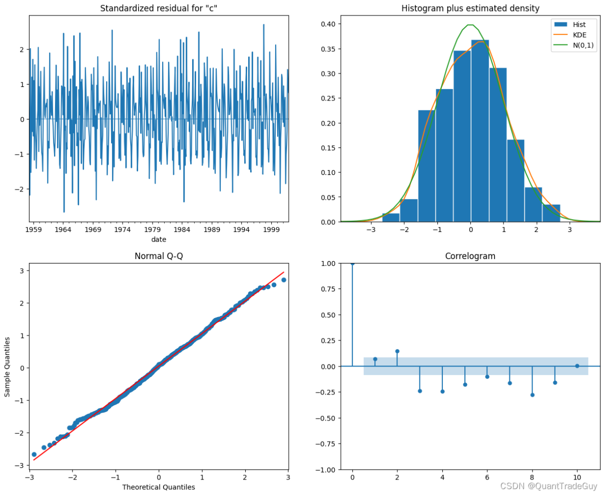
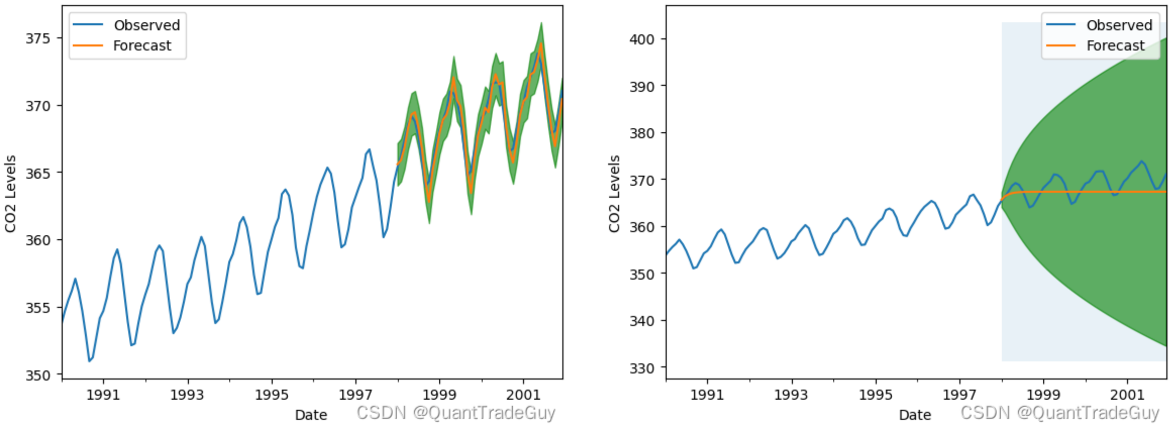


Figure 3: Parameters



(a) Real Data

(b) Simulated Data

Figure 4: Prediction using real (left) and simulated (right) data

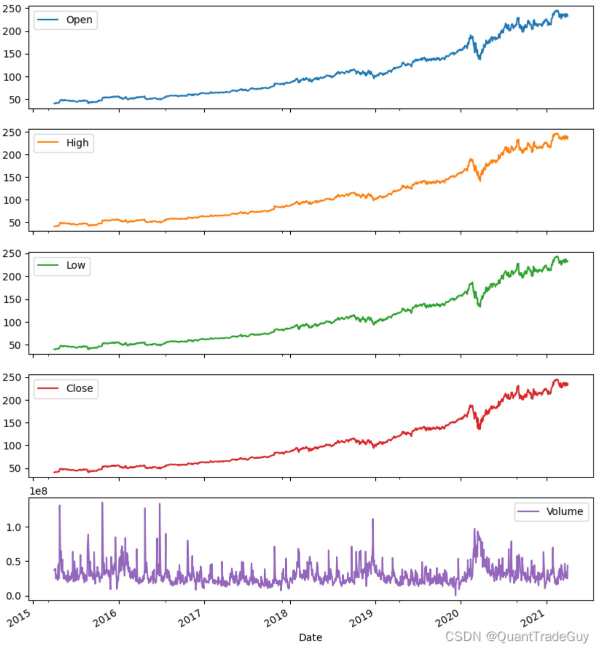


Figure 5: The stock prices

7 Project 2: Microsoft Stock Prices Prediction

1. Import and examine the data,. There are no missing values.

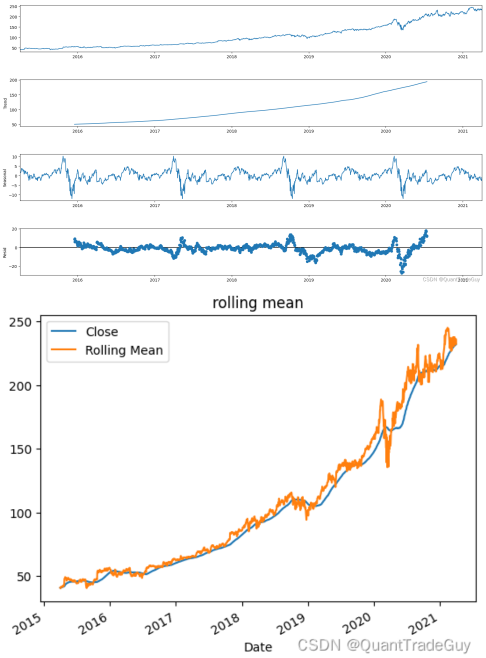
2. Next, resample the data by quarter.

3. Observe the correlation between the various data. Decompose trend and seasonality.

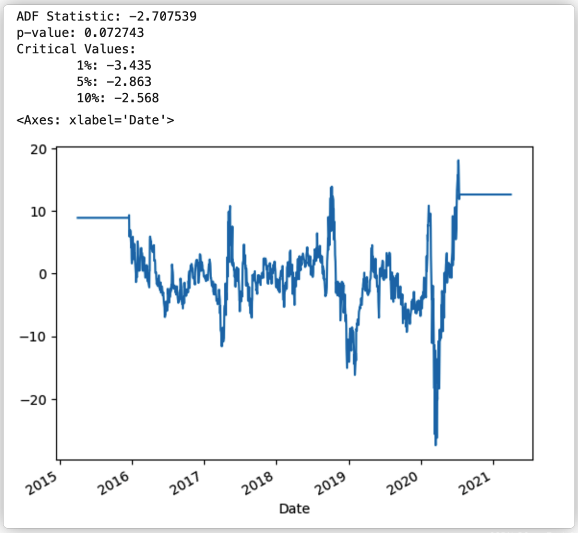
4. Then perform the ADF test on the residuals and find that the residuals are not stationary. Plot the ACF and PACF of the closing prices.

5. Train the model. Based on previous observations, set p to 3, q to 2, and use second-order differencing.

6. Predict.



(a) Decomposition



```
ADF Statistic: -2.707539
p-value: 0.072743
Critical Values:
        1%: -3.435
        5%: -2.863
        10%: -2.568
<Axes: xlabel='Date'>
```

(b) Residual

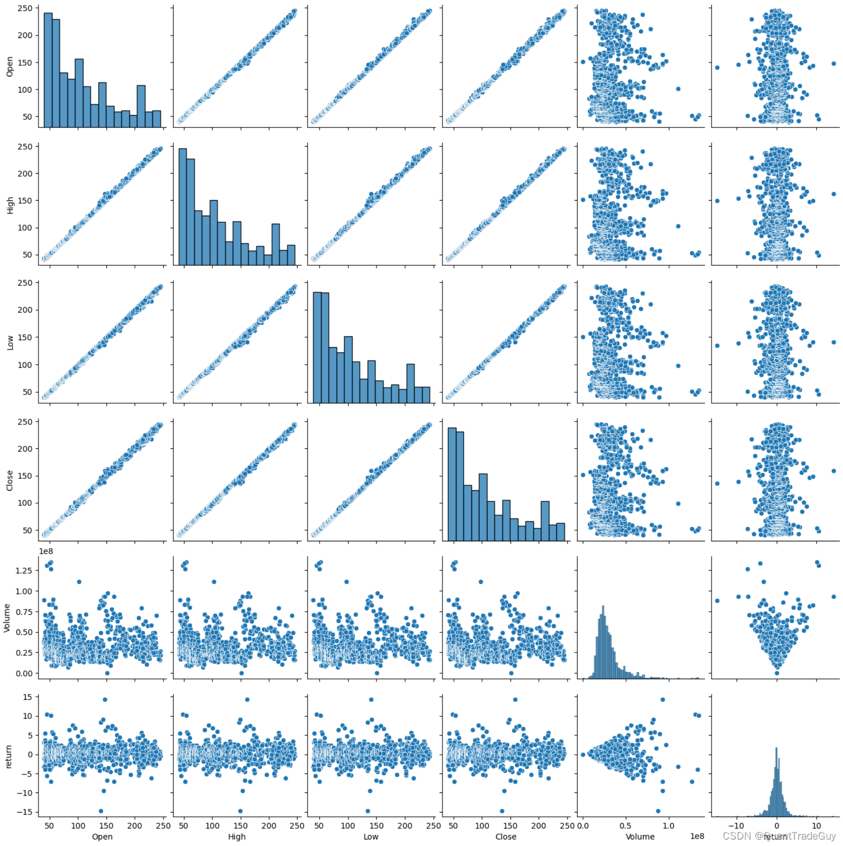
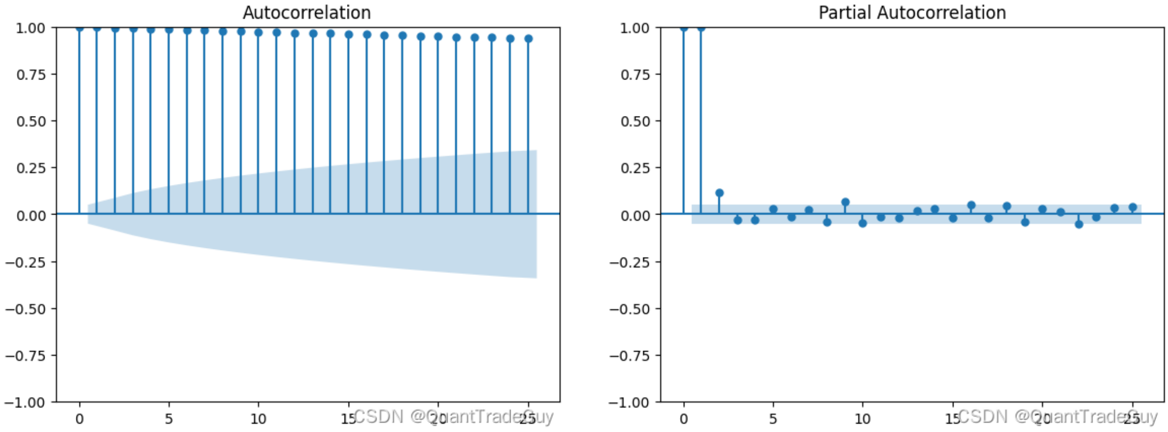


Figure 6: The features



(a) Average total cost with different schemes.

(b) Average time delay with different schemes.

Figure 7: Average cost with different schemes.

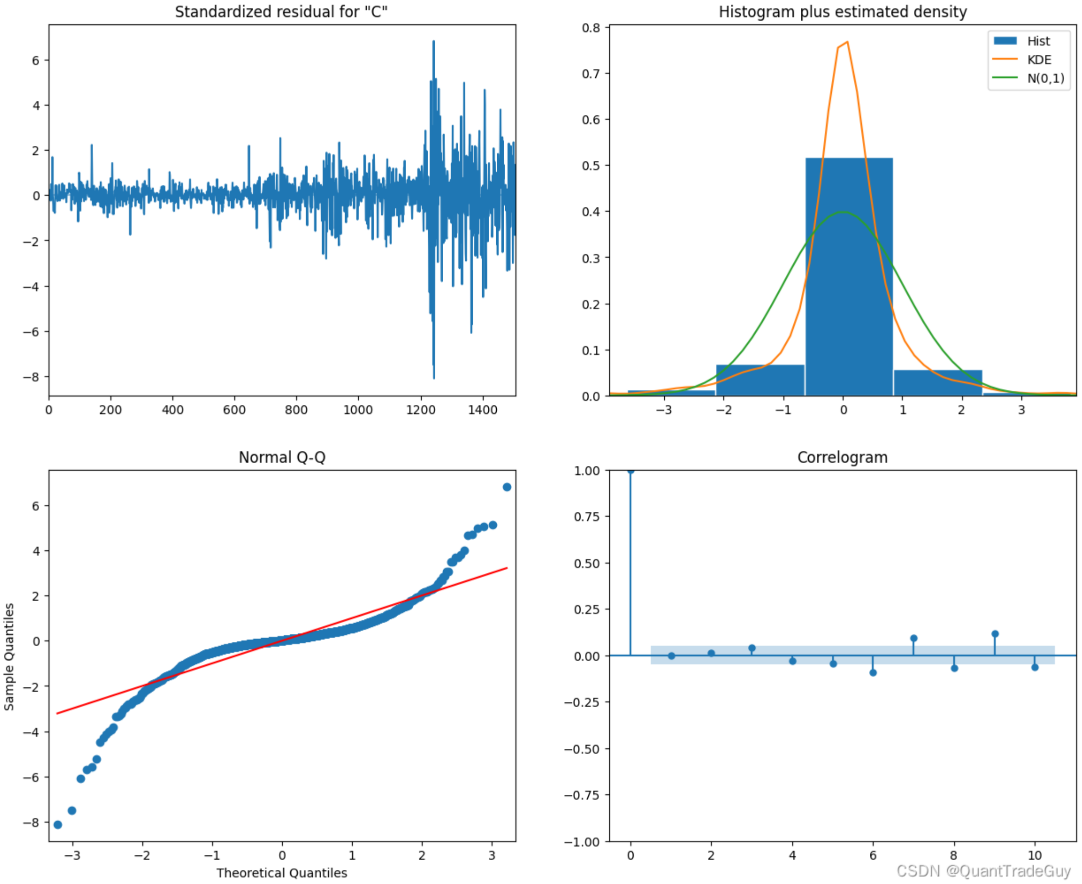
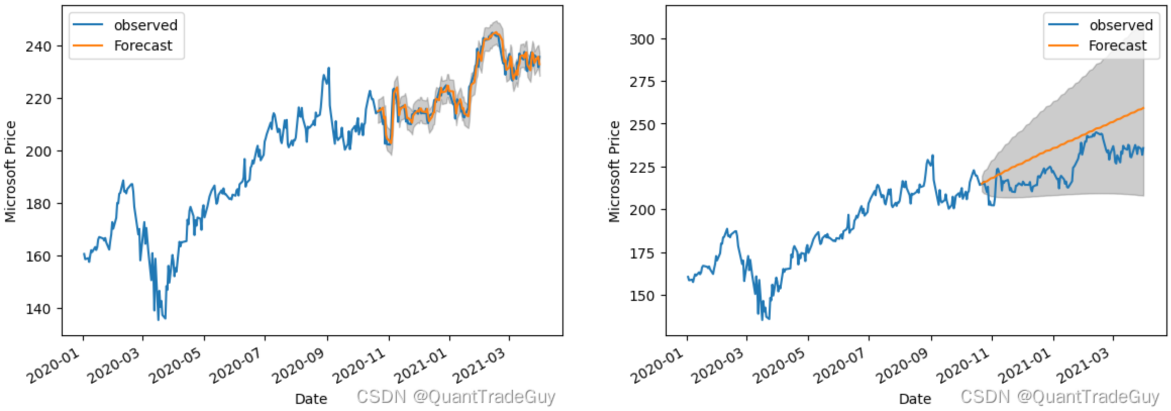


Figure 8: Parameters



(a) Real Data

(b) Simulated Data

Figure 9: Predict with real (left) and simulated (right) data.