Time series analysis on the stock price of Tesla Inc.

Wenhao Pan (3034946058), Ruojia Zhang, Mengzhu Sun, Xiangxi Wang, Mingmao Sun

November 13, 2021

Contents

1 Abstract

2 Introduction

Due to the increasing focus on carbon neutrality, the industry of replacing non-sustainable energy with sustainable energy has boomed in the past few years. Electricity, as a relatively environment-friendly energy, has been considered as replacement of some traditional energy, such as gasoline and diesel. Among all those enterprises pursing commercialized carbon neutrality, TSLA, as the largest electric car company, has been pioneering the fashion and aiming to transition the world to electric mobility. As the reflection of belief of the public, the stock price of TSLA has been sedentary for a period of time and has no evident increase until recent years. Therefore, we pick up the close price of TSLA stock of the recent 300 days to explore. In the following experiments, we utilize differencing, exponential smoothing, and fitting ARMA model, and combination of them to approximate the series.

3 Data Description

The TSLA stock price comes from Yahoo Finance (https://finance.yahoo.com). The stock price dataset consists of open price, close price, high price, and low price of a trading day. Since they has roughly similar trend, we choose close price to experiment on. The whole volume of data, which contains 2791 data points, has variance 39768.49, max price 1208.59, min price 4.01, mean price 112.4271. The recent-300-day data has variance 22697.1, max price 1208.59, min price 330.21, mean price 654.1912.

4 Exploratory Data Analysis

To obtain a comprehensive understanding of the data, we conduct explanatory data analysis (EDA) first. Figure 1(a) is the time series plot of all the given time points. We observe that the stock prices of Tesla before 2020 are averagely and considerably lower than those after 2020. The significantly different scales of different parts of the time series make it hard to visually examine the trend and seasonality pattern of the time series. Moreover, since we are majorly interested in the recent activities of Tesla, we do not have to analyze all the available data. Therefore, for the sake of interest and convenience, we decide only to analyze the last 300 time points, which cover the period from `2020-08-26` to `2021-11-02` excluding weekends. Thus, whenever we use the word "data" in the following analysis, we implicitly mean the time series of the last three hundred time points.

Figure 1(b) is the time series plot of the close prices of Tesla in the last three hundred trading days before and including `2021-11-02`. We first observe that our data is roughly homoscedastic based on Figure 1(b). To verify our observation, we try the square root and natural log transformations and see whether they effectively stabilize the variance of the time series. Their plots are below in Figure 2.

We can see that both transformations unnecessarily increase the variance of the time series before mid-November in 2020 and do not change the variance of other time series data. Although both transformations shorten the vertical distance between the maximum and minimum of the time series after Oct. 2021, the spike after Oct. 2021 is more like an increasing trend than a considerable fluctuation. In short, both transformations are redundant, and we do not need to use any variance stabilizing transformation.

Back to Figure 1(b), intuitively, the data is not stationary because of a nonlinear and generally increasing trend. The trend first increases until around Feb. 2021 and then decreases until around Mid-May. 2021. Finally, the trend increases again until the end of the time series. Nonetheless, we do not observe an obvious or significant seasonality pattern. It matches the intuition since the granularity of our data is day, and the structure of stock price data is too complicated to have a seasonality pattern.

In conclusion, based on all the previous discussions in EDA, we decide to construct possible models on the original time series data, including only the last three hundred time points.

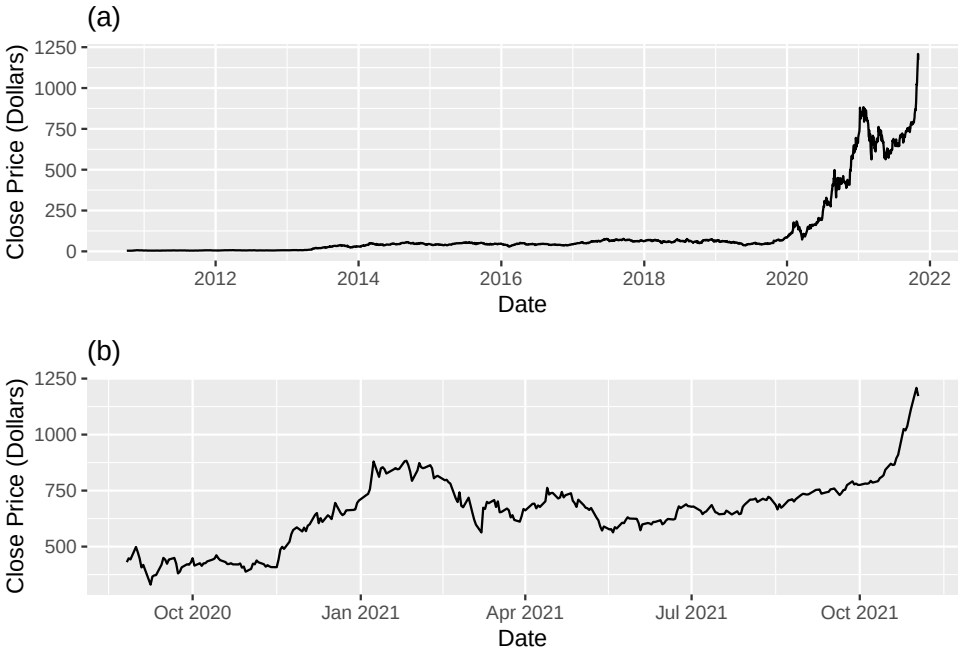

Figure 1: (a) Time series plot of all available trading days. (b) Time series plot of last 300 trading days

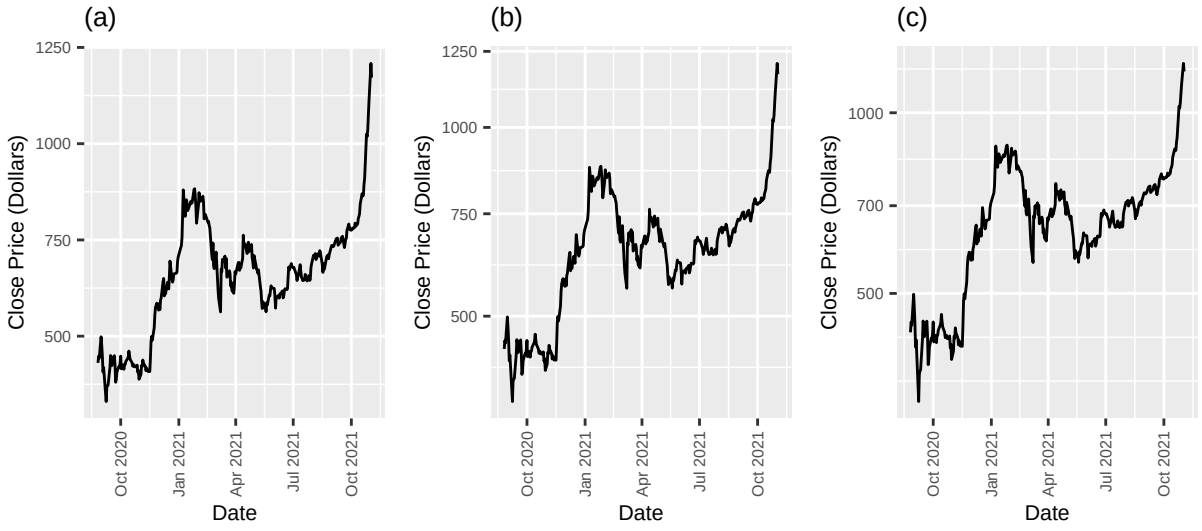

Figure 2: (a): Original time series. (b): Square root transfromed time seires. (c): Natural log transformed time series.

5 Model Construction

With a comprehensive understanding of our data, we start to experiment and construct different time series model. We choose and build two non-parametric signal models of the trend and seasonality in our data. We aim to make the residuals approximately weekly stationary. We do not consider any parametric trend model because we think the trend of the stock price data is too complicated to be modeled by a parametric model, such as a high-order polynomial. Certainly, we could use a 15 or 20 order polynomial, but it may overfit the training data and produce imprecise predictions. We do not consider a parametric seasonality model either because we do not find a clear seasonality pattern in our data by the EDA. Finally, based on each signal model, we provide two ARMA models or its extension, such as SARMA or ARIMA, to whiten the residuals of the signal model. Thus, we have four candidate models, and we will explain how we select a final model among them in the next section.

5.1 Non-parametric Signal Model: exponential smoothing

In this signal model, we choose exponential smoothing with weight $\alpha = 0.9$ and lag $k = 10$ and a seasonal differencing with period $d = 5$.

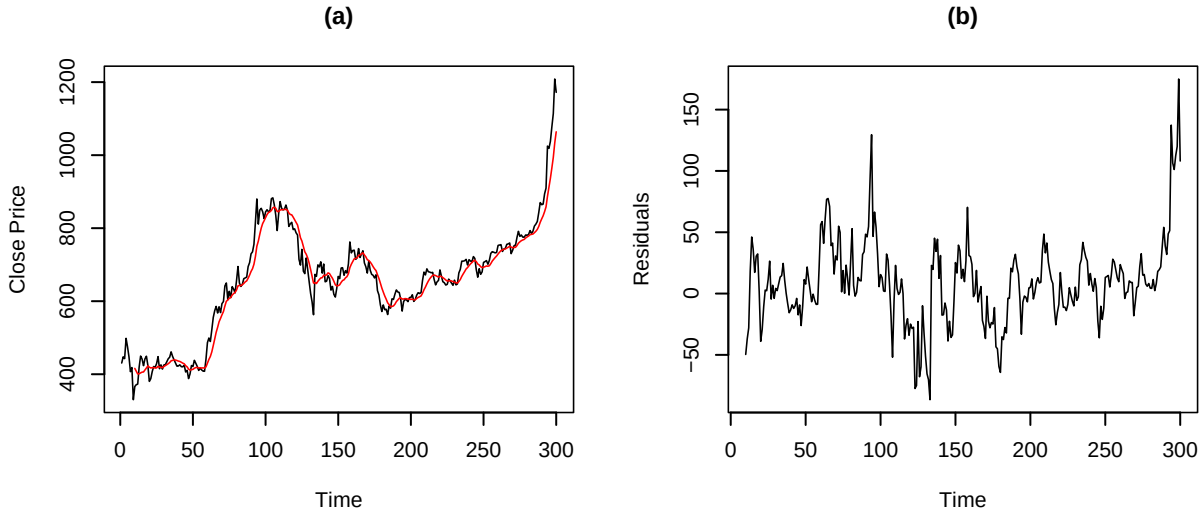

Figure 3: (a): Time series plot of the original data and fitted values. (b): The residual plot of exponential smoothing.

We experiment with different combinations of α and k with a careful consideration of overfitting issue. we choose $k = 10$ as the final value because we want to only use past two weeks, which are ten days in our data, to forecast. We choose $\alpha = 0.9$ as the final value because we think it best balances the smoothing effect and the capture of trend pattern among $(0, 1)$. Indeed, the smoothing line in Figure 4(a) fits the data in the way that we want. Note that we lose the first nine time points due to the computation process of the exponential smoothing.

However, the residual plot Figure 4(b) is fairly non-stationary, as it has cycling fluctuation pattern and still slightly nonlinear trend. It might be due to that we intentionally let exponential smoothing not fit the data perfectly. Next, We use the seasonal differencing with period $d = 5$, which is one week in our data, to further make the residuals more stationary.

Indeed, now the differenced residuals become more stationary. There seems to be a contradiction that recalling in EDA, we claim that there is not a clear seasonality in our data. However, the effect of the

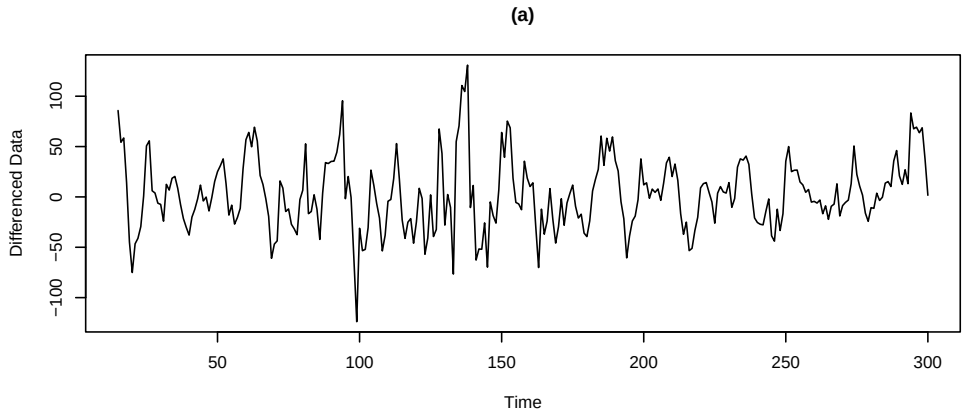

Figure 4: (a): Time series plot of the seasonal differenced (d = 5) residuals from the previous smoothing.

seasonal differencing here implies a possible seasonality with period $d = 5$. We think it might be due to that the seasonal differencing is actually removing the remaining trend left by the exponential smoothing instead of the seasonality. Nevertheless, We believe that the time series of the differenced residuals shown in Figure 5(a) is stationary enough for us to build ARMA models on it. For the convenience, we use "residuals" to represent "the seasonal differenced residuals" in the following two subsections.

5.1.1 ARMA 1A: $ARIMA(1,0,3) \times (0,0,1)[5]$

After acquiring a approximately stationary residuals from the signal model, we use ARMA model to further model the residuals so that we can predict future residuals reasonably. Then, summing the predicted residuals and signals gives us the prediction of the original time series, stock price. To obtain the intuition for choosing ARMA model, we use the sample ACF and PACF plots of the residuals shown in Figure TODO.

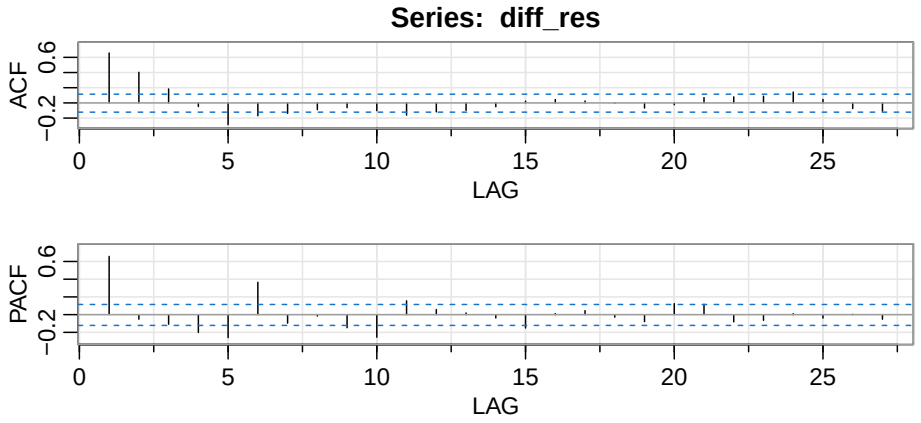

Figure 5: The sample ACF and PACF plots of the residuals from expo. smoothing

In the ACF plot, if we ignore lag 5 temporarily, we observe a sharp cutoff after lag 3, so $q = 3$ might be a reasonable choice. Similarly, if we ignore the lags around the multiples of 5 temporarily, the PACF plot has a cutoff after lag 1, so $p = 1$ might be a reasonable choice. To handle the spikes at lag 5 in ACF and lag multiples of 5 in PACF, choosing seasonal $MA(1)$ with period $d = 5$ might be reasonable. To increase the credibility of our choices, we apply `auto.arima` on the residuals and receive an $ARIMA(2,0,3) \times (0,0,0)[5]$ model,

which is similar to our candidate model ARIMA(1,0,3)×(0,0,1)[5], although all the Ljung-Box statistics of $ARIMA(2,0,3) \times (0,0,0)[5]$ have p-values smaller than 0.05.

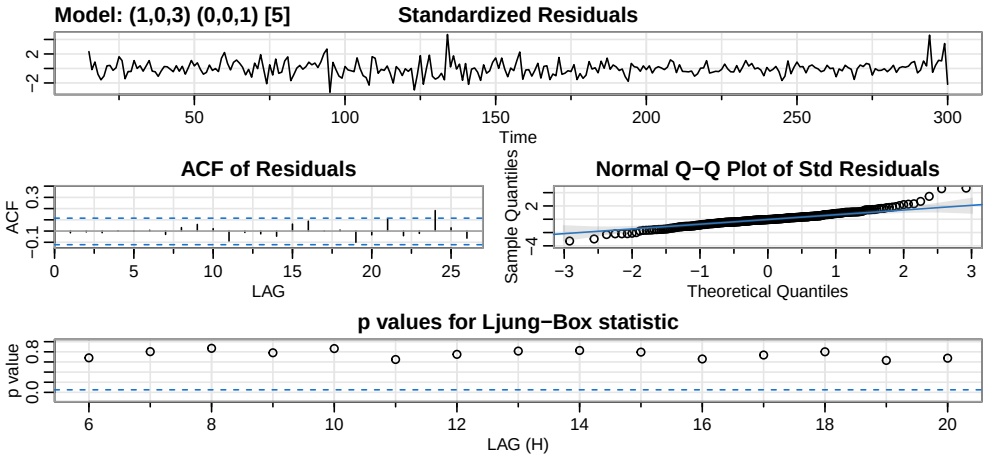

Figure 6: Model diagnostics plots of ARIMA(1,0,3)x(0,0,1)[5]

To assess our model, we visualize different model diagnostics methods in Figure TODO. The standardized residuals seem to have a i.i.d. standard normal distribution with a slightly more large values in magnitude. In the ACF plot of residuals, almost all the spikes lie within 95% confidence interval. The normal q-q plot implies a slightly fatter tail distribution than the normal distribution, but we can still claim the validity of the normality assumption. The p-values of all the Ljung-Box statistics are above 0.05, so we fail to reject the null hypothesis that the data was generated from our candidate ARMA model. In summary, all the model diagnostics methods imply that our candidate ARMA model is considerably reasonable. We call this model "ARMA 1A".

5.1.2 ARMA 1B: $ARIMA(1,0,1) \times (0,0,1)[5]$

We provide a second way to model the seasonal differenced residuals.

From the ACF and PACF plots of the residuals in Figure TODO, we can observe a negative spike at lag = 5 for both plots. If we do not ignore the spikes at some lags like we did in ARMA 1A, neither the ACF or the PACF plot has a reasonable cutoff. Thus, we need to choose both $p > 0$ and $q > 0$.

Recall that `auto.arima()` returns a $ARIMA(2,0,3) \times (0,0,0)[5]$ model but with bad diagnostics evaluation. Nonetheless, we can still treat it as a reference model and develop a better model based on it. Combining with the previous observations in the ACF and PACF plots, we experiment with different orders and look for the one with lowest AIC, AICc, and BIC values. See the code appendix for more details. We end up with $ARIMA(1,0,1) \times (0,0,1)[5]$ model.

The diagnostics plots in Figure TODO have a highly similar performance to those of ARMA 1A. Thus, we consider this model as a reasonable one and call it "ARMA 1B".

5.2 Non-parametric Signal Model: second-order differencing

In this model, we choose the second-order differencing to remove the trend. We observe that after the first-order differencing, there is still some trend pattern, such as the increasing one between 270 and 300, as shown by Figure TODO. This matches our previous analysis that the trend of our data is nonlinear in EDA. Thus, we take another differencing and acquire the second-order differencing data shown in Figure TODO.

The second-order differenced time series, denoted by Z_t, is more stationary than the first-order differenced time series. We can keep trying more higher-order differencings, but they may overfit our data. Therefore,

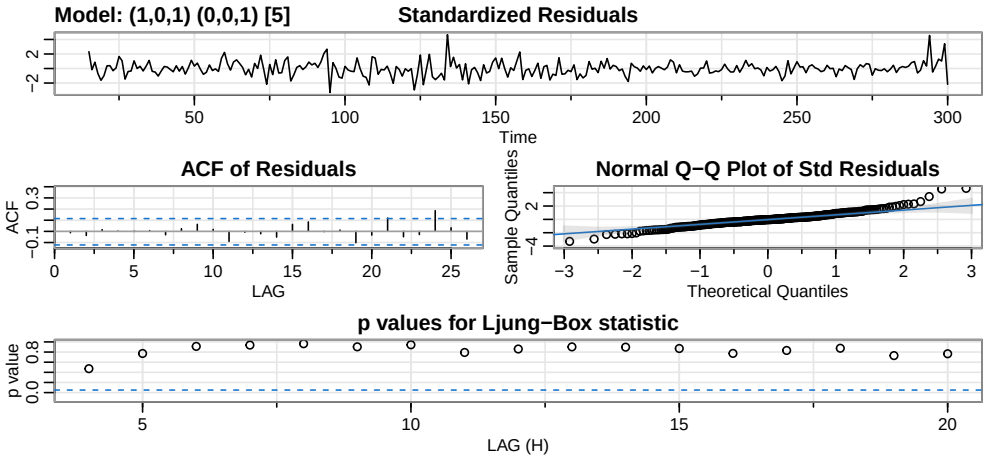

Figure 7: Model diagnostics plots of ARIMA(1,0,1)x(0,0,2)[5]

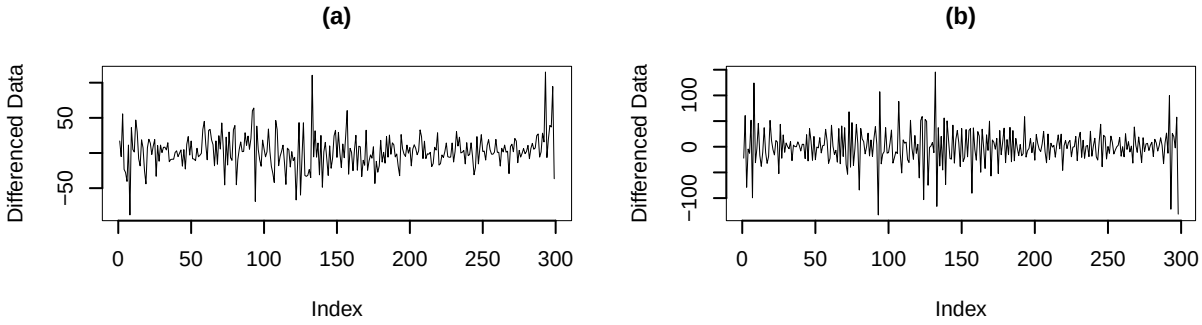

Figure 8: (a): The first-order differenced data. (b): The second-order differenced data.

we stop at the second-order differenced time series Z_t which is already stationary enough for us to build ARMA model on it.

5.2.1 ARMA 2A: $ARIMA(1,0,1) \times (0,0,0)[0]$

Based on Z_t, we continue our modeling by finding and fitting a suitable ARMA model on it. We first plot the sample ACF and PACF of Z_t as we did for the seasonal differenced residuals to obtain some intuition.

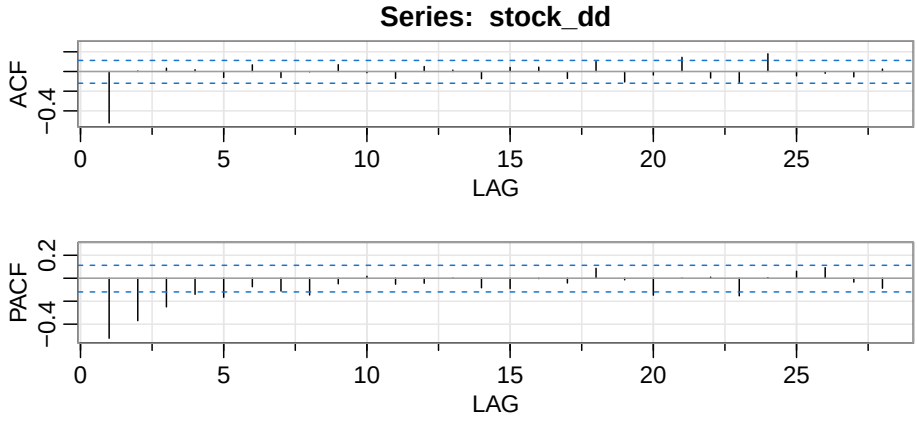

Figure 9: The sample ACF and PACF plots of Z_t

From the ACF plot of Z_t, we observe a sharp a cutoff after lag 1, so MA(1) seems to be a reasonable choice. Also, we observe an exponential decay in the PACF plot, indicating that we should use an $ARMA(p,1)$ model for some q. We use the `auto.arima()` to help us find the suitable p, and we acquire an $ARIMA(1,0,1) \times (0,0,0)[0]$ model on Z_t. Since we have done a second-order differencing with lag 1 on the original time series, essentially we use an $ARIMA(1,2,1) \times (0,0,0)[0]$ on the original time series. As usual, we plot the model diagnostics plots in Figure TODO to assess our model.

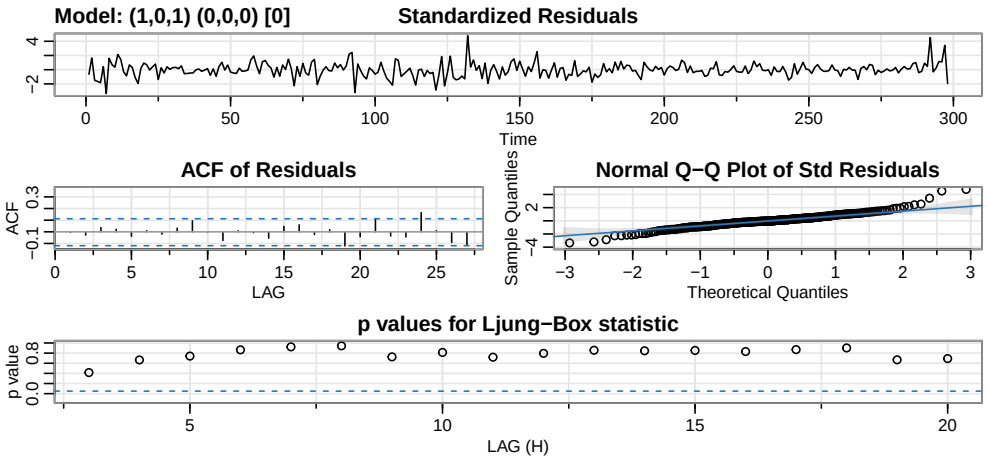

Figure 10: Model diagnostics plots of ARIMA(1,0,1)x(0,0,0)[0]

The performance of the diagnostics plots is similar to those of ARMA 1A and 1B models, which means our model choice here is reasonable.

5.2.2 ARMA 2B:

We provide a second way to model the second-ordered differenced time series Z_t.

In the PACF plot of Z_t in Figure TODO, the magnitude of PACF is decreasing significantly from lag 1 to lag 5. From the ACF plot of Z_t, we observe an evident cutoff of magnitude of ACF occurs at lag 1. Therefore, it is reasonable to fit Z_t with a MA model. However, clearly the ACF and PACF plot do not strictly match the theoretical ones of $MA(q)$ where $q \in \{4, 5, 6\}$ model, so we experiment with multiple combinations of $q \in \{4, 5, 6\}$ and of $p \in \{0, 1\}$. We choose $p \in \{0, 1\}$ because the model returned by `auto.arima()` has $p = 1$.

With consideration of cross-validation error, model AIC, model AICc, model BIC, the MA(4) model has the best performance. See the code appendix for more details about selection. We plot the model diagnostics methods in Figure TODO to further assess MA(4) model.

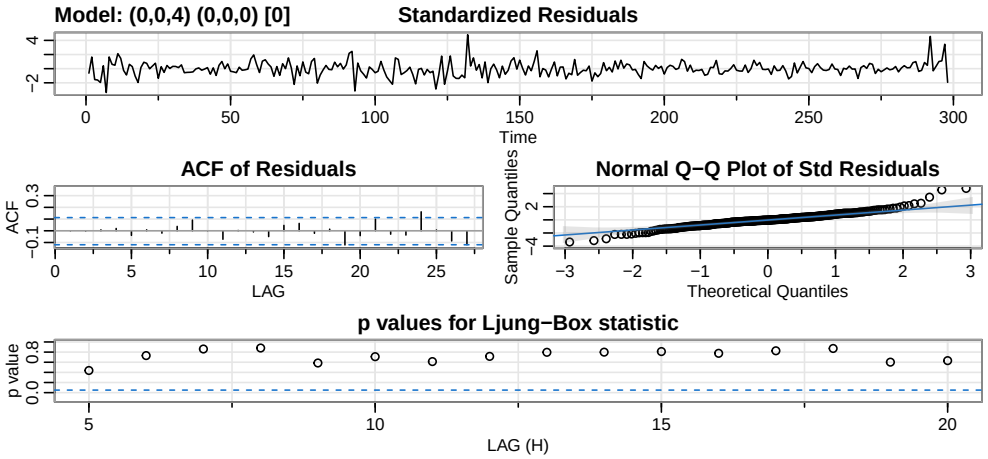

Figure 11: Model diagnostics plots of ARIMA(0,0,4)x(0,0,0)[0]

The performance of the diagnostics plots is similar to those of the previous three models, which means our model choice here is reasonable.

6 Model Comparision and Selection

Now, with four carefully chosen models, we need to select a final model that best fits the data. However, it is dangerous to blindly choose an overcomplicated model that can perfectly match the given data because the model may also fit the noise in the data as well. We only want our model capture the true underlying pattern of the data so that it can be generalized to unseen data well. Otherwise, our prediction might be too noisy and far away from the actual value. This issue is known as overfitting.

To lessen the overfitting issue as much as possible, we compute and compare the AIC, AICc, BIC, and cross-validation error of each model. The first three criterion measures the fitness of a model on the in-sample data while penalizing the complexity of the model. The last one measures the predictability of a model on the out-of-sample data. Then, we will select the model that has the best performance with a comprehensive consideration of all four criterions. See the code appendix for more details about the selection process.

we summarize the criterion of all the models into the following table:

Table 1: Criterion values of each model

Index	Model	CV.error	AIC	AICc	BIC
1	Expo. smoothing + ARMA 1A	5492.341	8.898896	8.899949	8.988378
2	Expo. smoothing + ARMA 1B	5521.305	8.886808	8.887305	8.950724
3	2nd-order Diff. + ARMA 2A	4507.891	9.231862	9.232136	9.281488
4	2nd-order Diff. + ARMA 2B	4461.713	9.243639	9.244328	9.318077

Interestingly, the models using exponential smoothing to remove the signal have higher cross-validation errors but smaller AIC, AICc, and BIC than the models using second-order differencing to remove the signal. It could be that our exponential smoothing model is more overfitting than the second-order differencing model. Because the average difference between cross-validation errors of second-order differencing models and those of exponential smoothing models is much larger than the the average difference between AIC, AICc, and BIC of two models, we will select one model from two second-order differencing models. Now, since the criterions of these two models are very close to each other, we will simply choose model 3 somewhat arbitrarily.

There is a caveat that the we extract AIC, BIC, and AICc of these four models directly from their `sarima()` fitted results. However, the exponential smoothing cannot be directly involved in `sarima()` while the second-order differencing can be. This discrepancy might bias the measured criterions. For the simplicity of this project, we ignore this discrepancy, although we recommend further exploration in it.

7 Final Model

7.1 Model interpretation

7.2 Prediction

8 Conclusion