Time Series Analysis and Forecast for Import Volume of Forzen Strawberries from 2010 to 2022

Mi Zhang

5/1/2022

Abstract

This paper intends to analyze the monthly import volume of frozen strawberries from January 2010 to February 2021 and forecasts the next 12 months of import volume comparing along with the actual data. Data is selected from the U.S. Department of Agriculture [1]. In this paper, three major time series model are performed including the non-seasonal ARIMA model with different metrics, the Holt-Winters Forecast model, and the seasonal ARIMA model with different metrics. However, non-seasonal ARIMA models are not so appropriate due to failure of accuracy performance for residuals analysis; and the Holt-Winters forecast model may not be consistent for the long-term forecast. Therefore, the final model is the seasonal ARIMA model of ARIMA(1,0,1)(0,1,1)[12]. The forecast for the next 12 months is plotted along with the actual import volume and results match well.

Introduction

In recent years, the United States has become the top importer of strawberries in the world, and 99% of strawberries are imported from Mexico[3]. On the other hand, frozen strawberries have a longer expiration date but the same nutrient values as fresh strawberries[4]. Especially during the time of the Covid-19 pandemic, people who are under quarantine value frozen fruits and vegetables more than before. There are many reasons for the importation of strawberries, such as limited US production, the seasonal gap in production, or foreign products having a lower price. As we can see in figure 1, there is an increase in volume in March, April, and May every year, proving that the major reason for the importation of frozen strawberries may be due to the season gap. In this project, we are going to use the data from the U.S. Department of Agriculture [1] to perform a time series analysis and forecast for the monthly import volume of frozen strawberries from the year 2010 to 2022. Some models include ARIMA, SARIMA, and Holt-Winters.

Data Preparation

The original data consists of a mixture data of fresh strawberries, frozen strawberries, and preserved strawberries, each column is not labeled properly, and the raw data has longer data from the year 2000 to 2022. The first thing I did was manually select data about frozen strawberries, shorten the period from January 2010 to February 2022 only, and label each column properly. On the other hand, the date of month and year are in different columns, so I created a separate column named "date" for the proper format of the date and drop all unnecessary columns. Table 1 displays the final format of my data. Before carrying out the model fitting process, I also transform my data into time series specific data with a frequency of 12 and separate data into the training set and testing set. Details of data clean can be found in the "MA585_REPORT.Rmd" file on my GitHub[5].

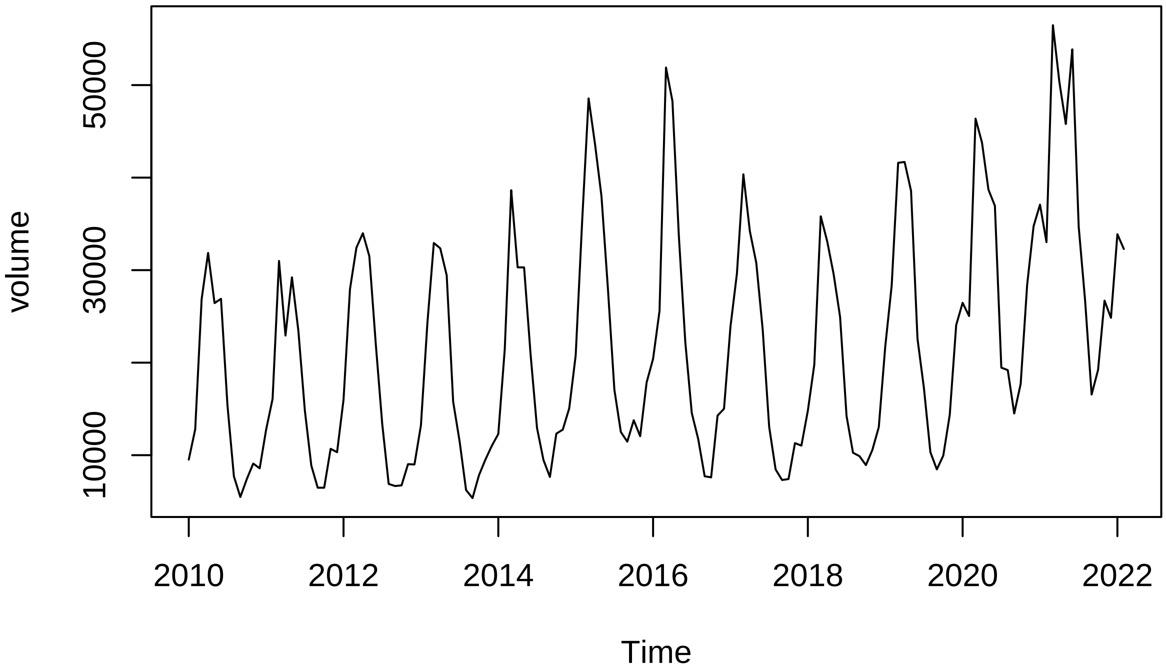

Figure 1: Import Volumn of Forzen Strawberry from 2010 to 2022

Table 1: Import Volumn of Forzen Strawberries from 2010 to 2022

date	volume
Jan 2010	9518
Feb 2010	12791
Mar 2010	26860
Apr 2010	31862
May 2010	26435
Jun 2010	26906

Models and Validation

Stationality, Transformation, and Differencing Tests

Before performing any model fitting, I performed an Augmented Dicky-Fuller test to check if my data is stationary or not and see if there is any difference needed. As the result shown below, the p-value of 0.01 suggests that I can reject the null hypothesis of non-stationarity and there is no differencing needed. However, I also tried log transformation to stabilize the variance. The top row of Figure 2 shows two time-series plots of the import volume of frozen strawberries. We can see after log transformation (top right), our data has a more stabilized variance and there is a seasonal trend in the data. The ACF plot shows a sine-wave decay to zero gradually, and PACF shows a cutoff after initial lags.

```
##
##  Augmented Dickey-Fuller Test
##
## data:  train
## Dickey-Fuller = -7.8996, Lag order = 5, p-value = 0.01
## alternative hypothesis: stationary
```

```
## Number of differences required for a stationary series is 0
```

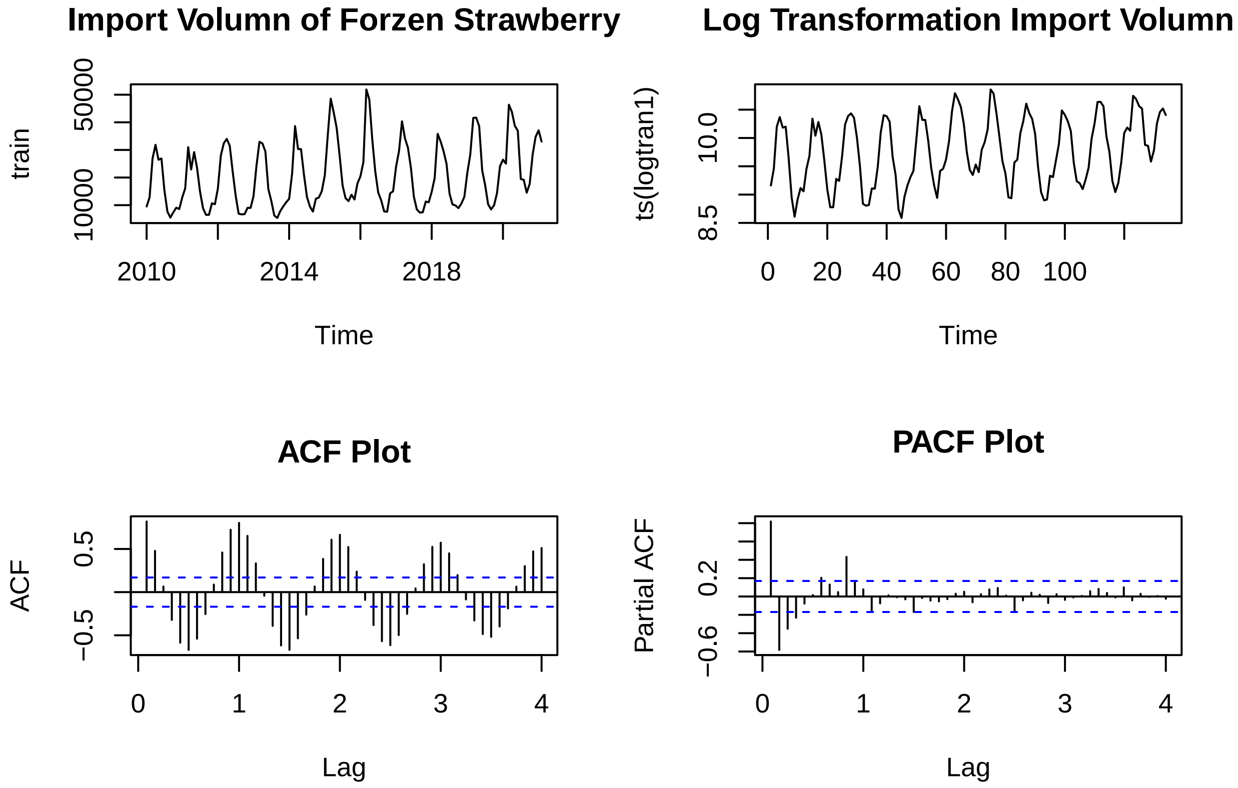

Figure 2: Log-transformation graph, ACF and PACF of log data

ARIMA

After observing the ACF and PACF plot for log data (Figure2), I decided to start fitting some potential non-seasonal ARIMA models. To start, I performed ARIMA Subset Selection aiding in model fitting as shown in Figure 3, and the potential models are AR(1), MA(1), ARMA(1, 1), and ARMA(3, 2).

The AICs results of ARIMA models are presented in Table 2. As we can see that ARIMA(3,0,2) has the lowest AICs, so further analysis is needed. Therefore, a time series diagnosis is performed and the result is shown in Figure 4. However, the result is not favorable as some lags are above the 95% confidence interval in ACF of Residuals and some lags fall below the 5% in p-values for Ljung-Box statistics. Therefore, another model should be considered.

Table 2: Comparison of ARIMA Models

ARIMA_Models	AICs
ARIMA(1,0,0)	91.27
ARIMA(0,0,1)	131.41
ARIMA(1,0,1)	62.68
ARIMA(3,0,2)	-30.13

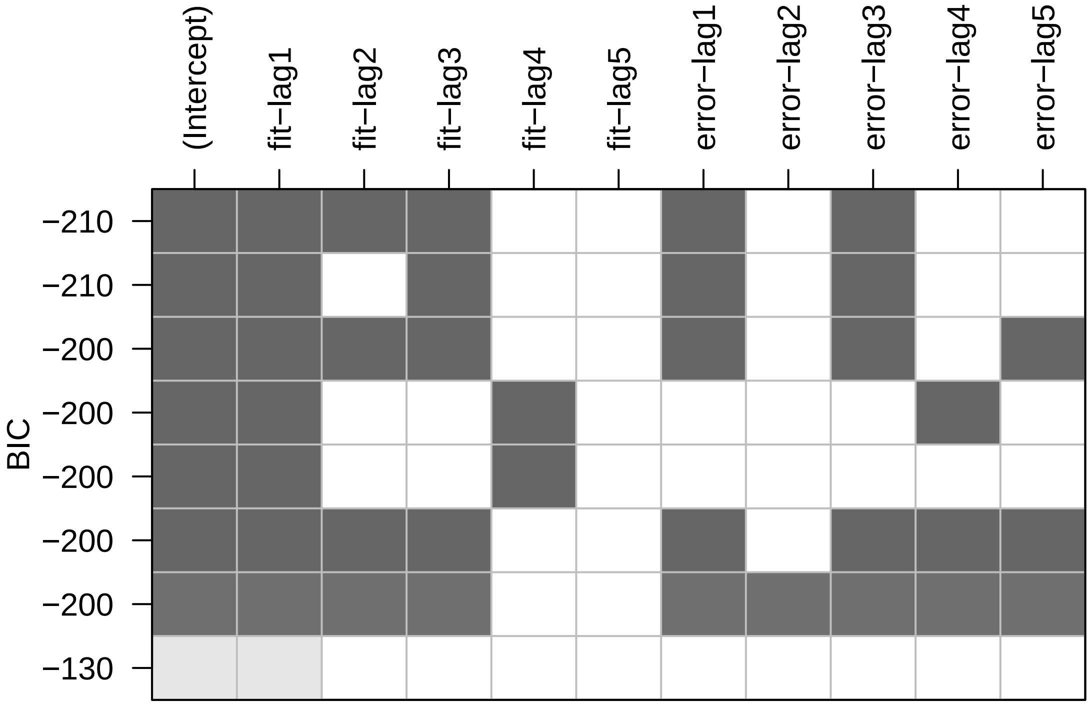

Figure 3: ARIMA Subset Selection

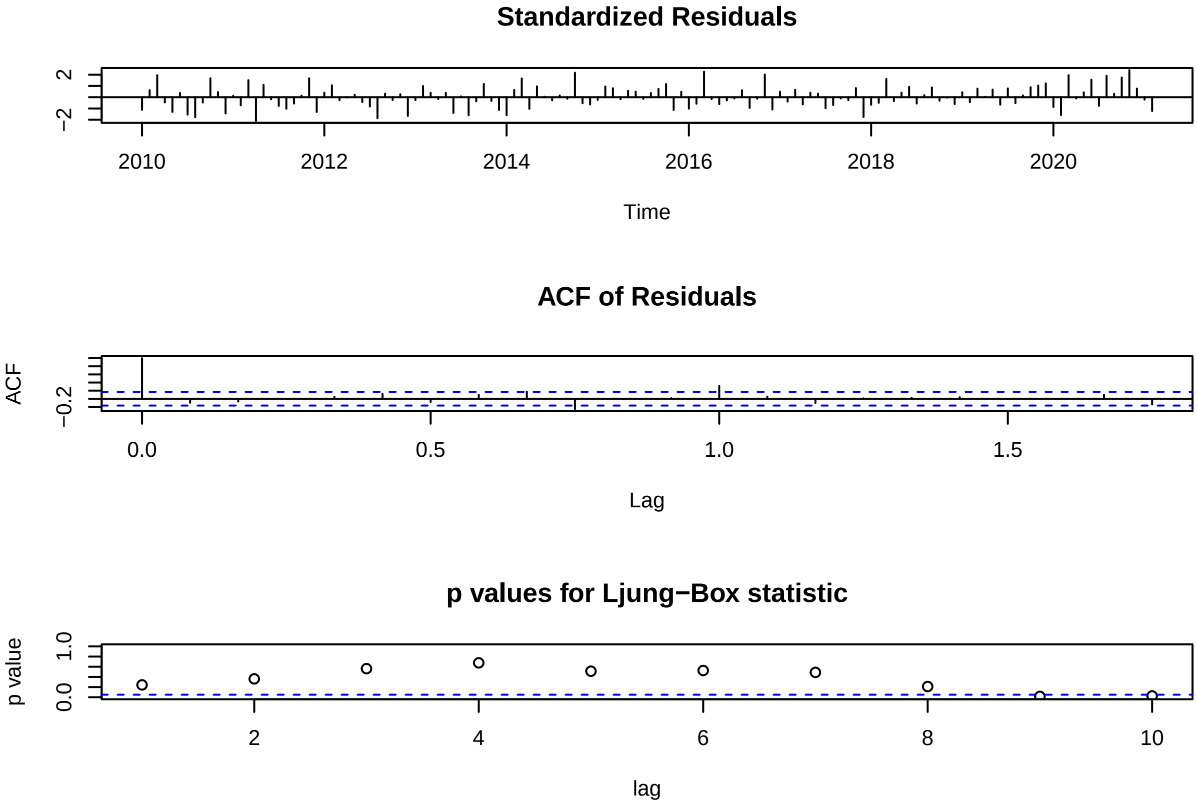

Figure 4: Time Series Diagnosis for ARIMA(3,0,2)

SARIMA

Since ARIMA models do not work well, I revisited ACF and PACF plots and noticed that the PACF plot shows a cutoff at lag 1 (lag12) which means there may be a seasonal pattern in the data. Therefore, I started to explore some seasonal ARIMA models. Potential SARIMA models are fitted and AICs results are presented in Table 3.

Table 3: Comparison of ARIMA Models

SARIMA_Models	AICs
ARIMA(0,0,1)(0,0,1)[12]	62.28
ARIMA(1,0,1)(0,1,1)[12]	-64.48
ARIMA(1,0,0)(0,1,1)[12]	-56.94

Since ARIMA(1,0,1)(0,1,1)[12] shows the lowest AICs value, I performed accuracy diagnosis for the residuals to confirm the result. As shown in Figure 5, there is no pattern in Standardized Residuals, none of the lags are significant in ACF of Residuals which means the residual is white noise, and all lags are non-significant in p-values for Ljung-Box statistics. Therefore, the result of ARIMA(1,0,1)(0,1,1)[12] is favorable.

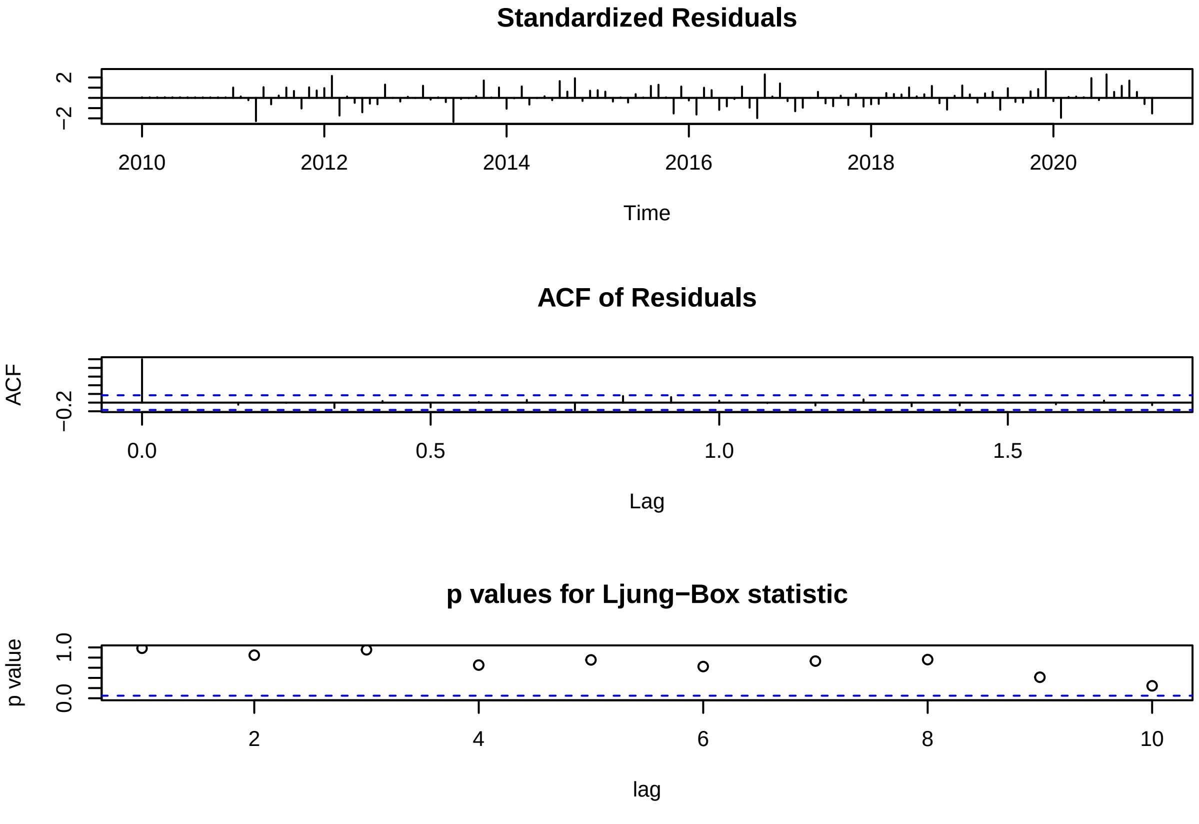

Figure 5: Time Series Diagnosis for ARIMA(1,0,1)(0,1,1)[12]

Holt_Winters Forecast

Since the data shows also both a linear trend and a seasonal trend, I think it is also reasonable to carry out a Holt-Winters seasonal forecast. Figure 6 shows a forecast of the data with a multiplicative seasonal pattern.

To check the forecast accuracy for next 12 months, I compared the forecast result with log(test) dataset. For the first ten months, the forecast matches well, but the last two months (January 2022 and February 2022) should be an uptrend instead of a downtrend. Therefore, Therefore, the Holt-Winters forecast may not work well for a long-term forecast and further evaluation is needed.

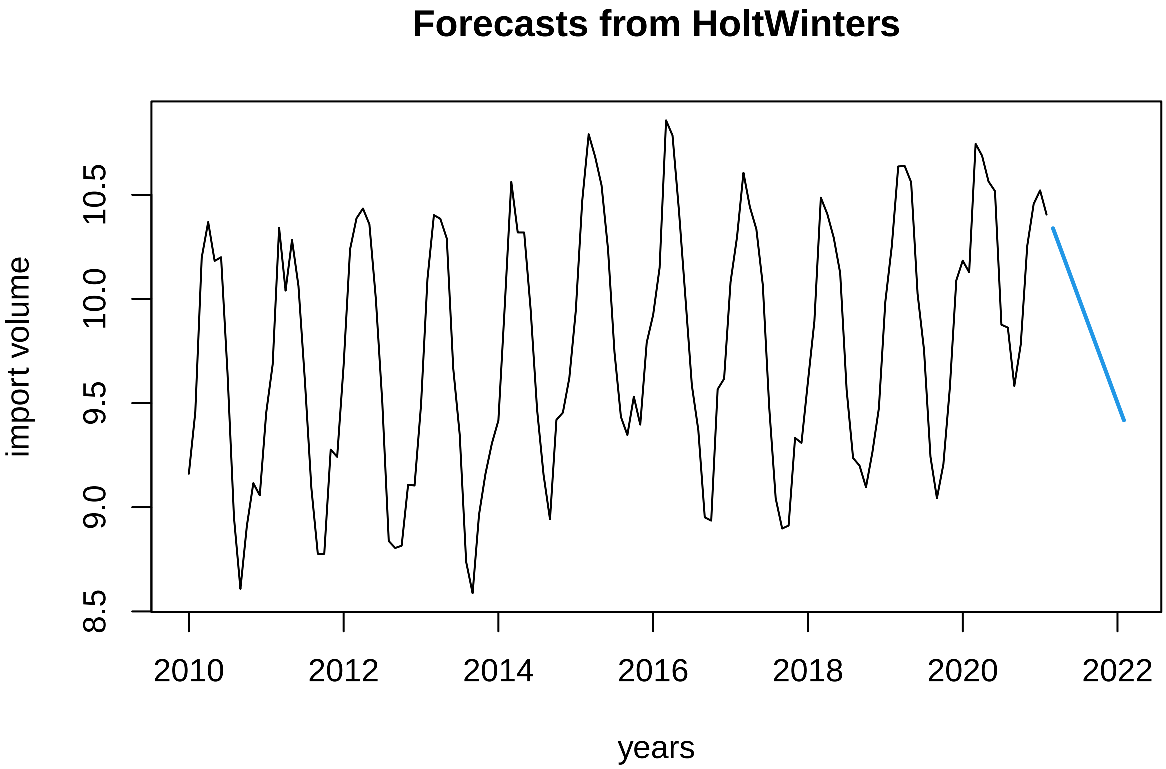

Figure 6: Mulitpicative Holt-Winters Seasonal Forecast

Final Model & Forecast

In order to compare all models that are performed in this report, a forecast performance accuracy summary is shown in Table 4. As we compare all measures of error, ARIMA(1,0,1)(0,1,1)[12] demonstrates lower forecast errors than ARIMA(3,0,2) and Holt-Winters'. Therefore, I performed a forecast for ARIMA(1,0,1)(0,1,1)[12] model for next 12 months as shown in Figure 7. In addition, I also plot log(test) which contains the actual import volume for the next 12 months along with the forecast. Overall, the forecast plot matches well with the actual import volume.

Table 4: Comparison of ARIMA(3,0,2), Holt-Winter, and ARIMA(1,0,1)(0,1,1)[12]

Forecast_Error	ARIMA_Model	Holt_Winters	SARIMA_Model
MAE	0.4874181	0.5374143	0.2042181
RMSE	0.5518245	0.5996087	0.2590818
MAPE	4.725703	5.116152	1.978883

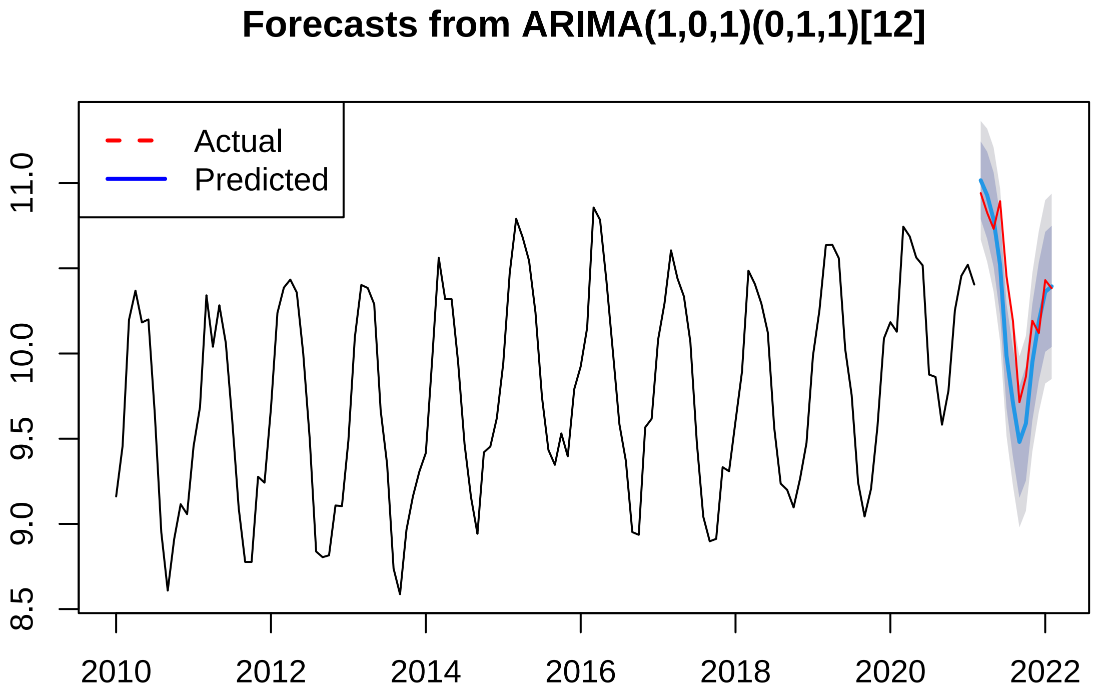

Figure 7: Forecast from ARIMA(1,0,1)(0,1,1)[12]

Result and Discussion

Recall the fact that the United States is the top importer of strawberries in the world, an accurate forecast of the import volume of frozen strawberries can play an important role in international trade. After performing non-seasonal ARIMA models, seasonal ARIMA models, and Holt-Winters forecast model and checking their forecast performance accuracy, ARIMA(1,0,1)(0,1,1)[12] model demonstrates the best result for forecasting the next 12 months of import volume of frozen strawberries. However, we also need to keep in mind that time series analysis and forecast can be impeded by real-world events, such as the Covid-19 pandemic caused labor shortages or natural disasters.

References

[1] https://data.ers.usda.gov/reports.aspx?programArea=fruit&top=5&HardCopy=True&RowsPerPage=25&groupName=Noncitrus&commodityName=Strawberries&ID=17851#P5f88596f8b7c43dea7b836500f2e9887_11_89iT0R0x1

[2] https://oec.world/en/profile/hs/strawberries-uncooked-steamed-or-boiled-frozen

[3] https://edis.ifas.ufl.edu/pdf/FE/FE97100.pdf

[4] https://foodingredients.treetop.com/fruit-ingredients/frozen-fruit/frozen-strawberries/

[5] https://github.com/lvesmile/Time-series-analysis

Appendix

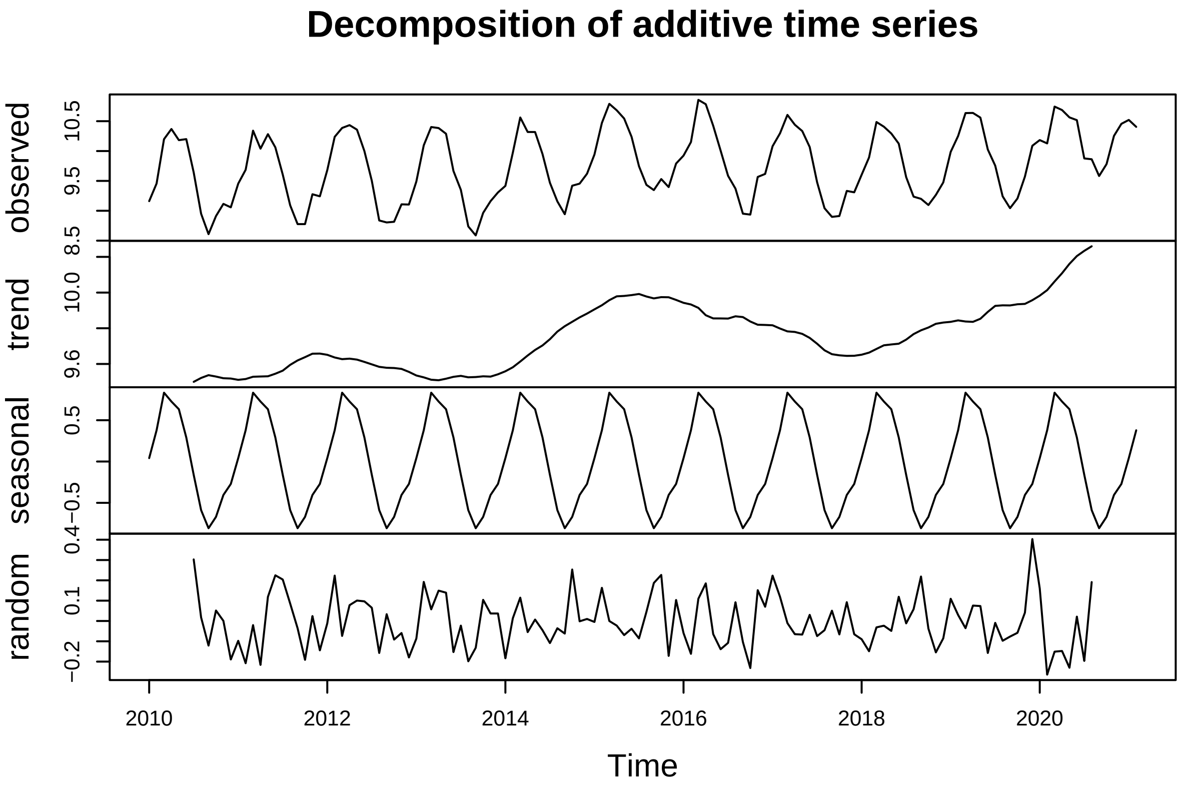

Figure 8: Decomposition of additive time series

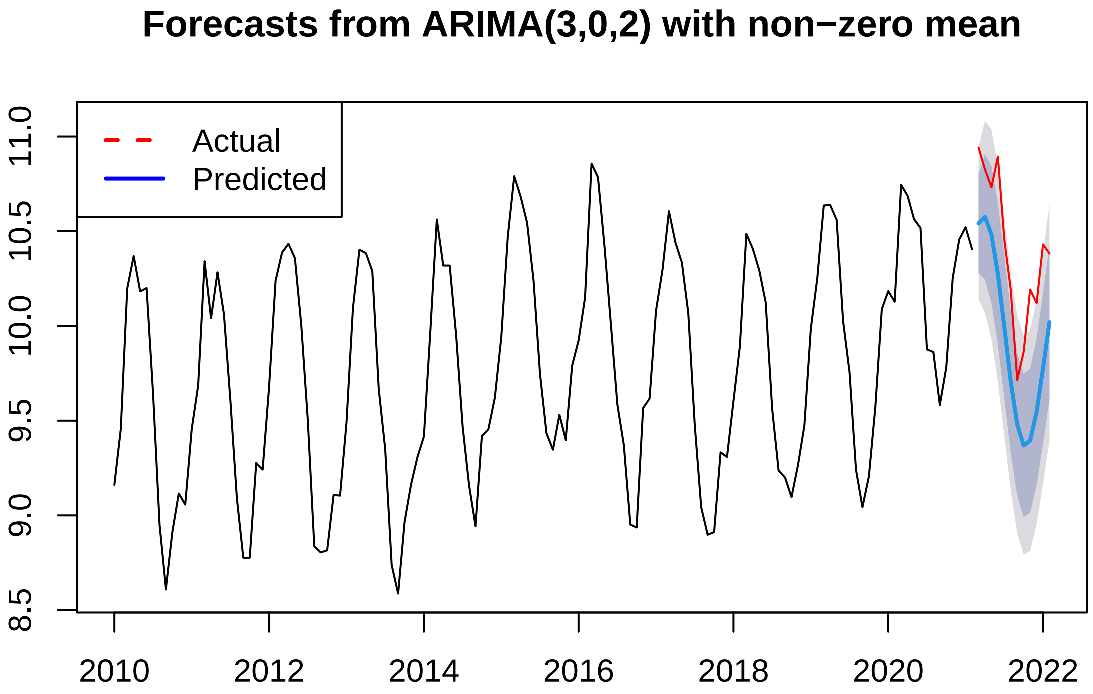

Figure 9: Forecast from ARIMA(3,0,2)