Time Series Forecasting for Monthly Retail Trade and Food Service

Ranfei Xu

Summary

In the consumer industry that I am interested in, estimating future production/demand data based on historical data to prepare the supply chain is a crucial part of the company's operations. Because of the company's privacy, it is difficult to find the company's specific sales data, so I use the **Monthly Retail Trade and Food Service** data of U.S. published by the **United States Census Bureau** as the research object of this project. In particular, I focus on the **Restaurants and Other Eating Places** industry. And at last, I found that in my situation, Holt-Winters seasonal model performs slightly better than $SARIMA(0,1,1) \times (2,1,1)_{12}$.

The path of data is: https://www.census.gov/econ/currentdata/dbsearch?program=MRTS&startYear=2008&endYear=2022&categories=7225&dataType=SM&geoLevel=US&notAdjusted=1&submit=GET+DATA&releaseScheduleId=

Data Visualization

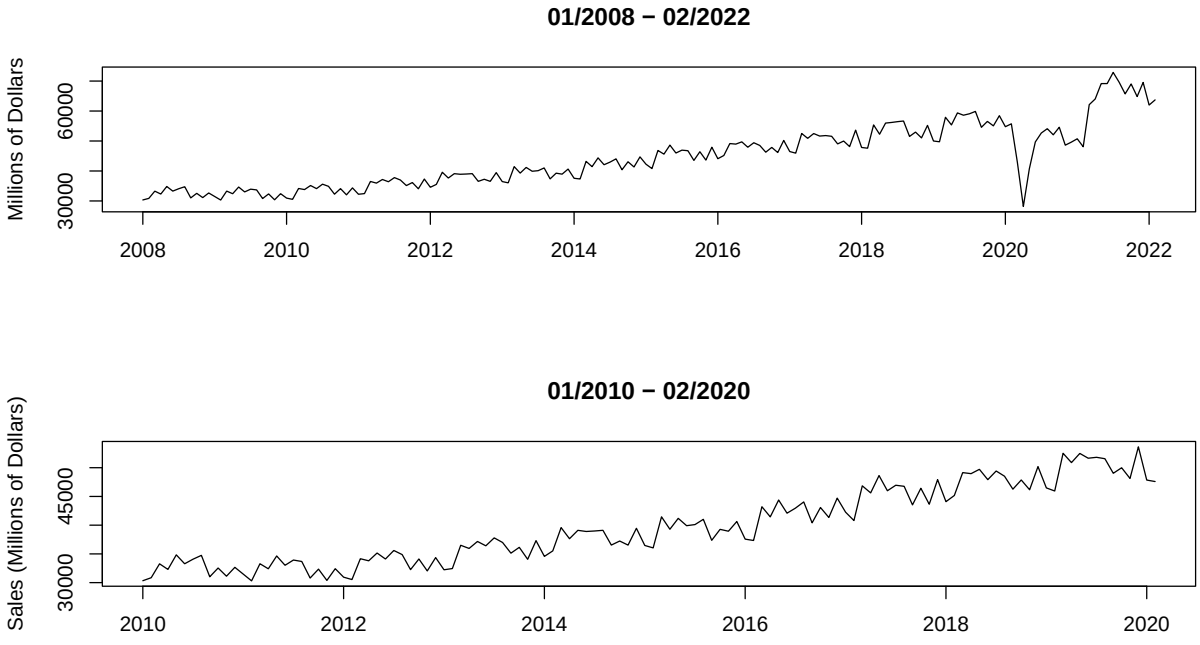

Figure 1: Plots of Time Seires

Data Preparation

The plots shown on the front page describe the data before and after I justified the length. Due to the impact of the pandemic, the data after 02/2022 is greatly affected by the policy which is hard to find the regular feature. And considering longer data is not always better than shorter data, so also to exclude the interference of premature data, I will describe the modeling procedure only using the data from **01/2010 to 02/2020**.

Data Processing

```
##
##  Augmented Dickey-Fuller Test
##
## data:  x
## Dickey-Fuller = -3.7255, Lag order = 4, p-value = 0.02495
## alternative hypothesis: stationary
```

The stable pattern of the variance of the data shown in figure[1] indicates that there is not necessary to apply the Box-Cox transformation to stabilize the variance. When carrying out the Dickey-Fuller test to check the unit root, the p-value is 0.02495 (smaller than 0.05) which confirms the process is stationary, which is a necessary condition for our analysis.

Classical Decomposition

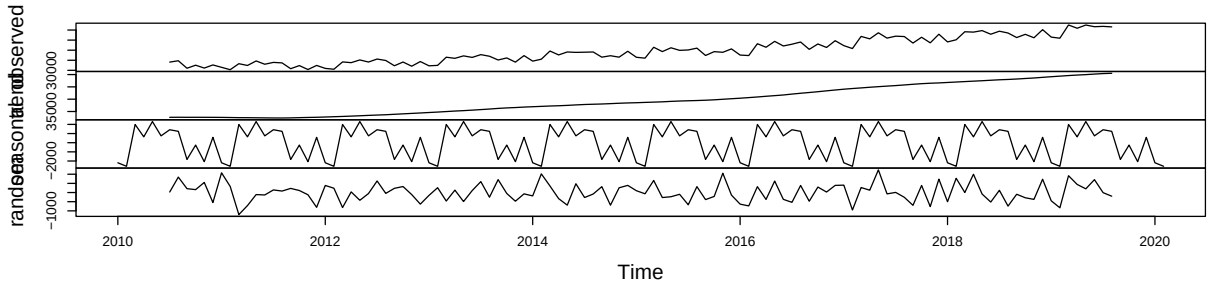

Figure 2: Decomposition of Time Series

Figure[2] shows the classical decomposition of the data. We can intuitively see that the trend and seasonal components are pulled out, which is consistent with the regular fluctuations of the PACF plot of raw data (Figure[3]). Thus, I choose the SARIMA model as my candidate model. And since the data is calculated monthly, I set the seasonal period as 12.

Modeling

I first split the data into training and test data sets for calculating the forecast accuracy later, and fit the model only based on the training set.

```
train <- ts(dt, start = c(2010,1), end = c(2018,12), frequency = 12)
 test <- ts(dt, start = c(2019,1), end = c(2020,02), frequency = 12)
```

Model-based Forecast: SARIMA

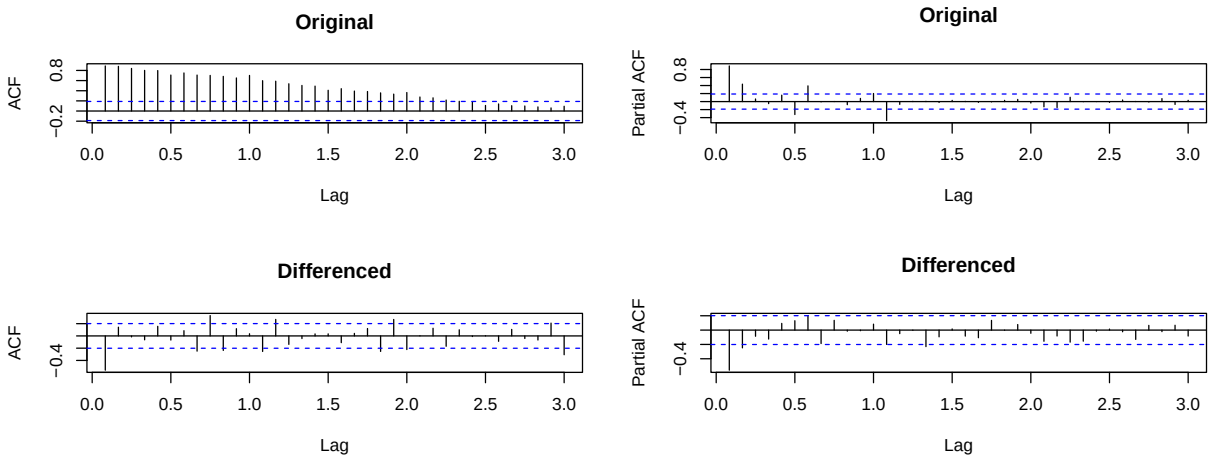

Figure 3: Correlation Plots before and after Differenced

The plots of original data show that ACF decay to 0 and PACF cuts off after lag 2 with a significant spike at near lag 12 which suggests a seasonal AR(2) component. The plots of data after eliminate trend and seasonal component show that ACF cuts off after lag 1 and PACF cuts off after lag 2, so I choose to compare the following models:

$SARIMA(0,1,2)\times(0,1,2)_{12}$ (returned from auto.arima()), $SARIMA(2,1,0)\times(2,1,1)_{12}$ (the initial model I choose), $SARIMA(2,1,0)\times(0,1,2)_{12}$, $SARIMA(2,1,2)\times(0,1,2)_{12}$, $SARIMA(0,1,2)\times(2,1,1)_{12}$, $SARIMA(0,1,1)\times(2,1,1)_{12}$

Model	AICc
SARIMA(0,1,2)(0,1,2)[12]	-502.625
SARIMA(2,1,0)(2,1,1)[12]	-503.8678
SARIMA(2,1,0)(0,1,2)[12]	-502.0279
SARIMA(2,1,2)(0,1,2)[12]	-499.2511
SARIMA(0,1,2)(2,1,1)[12]	-504.4763
SARIMA(0,1,1)(2,1,1)[12]	-504.7992

It turns out that $SARIMA(0,1,1)\times(2,1,1)_{12}$ has the lowest AICc. Then I check the coefficient estimations, and it turns out that the 95% CI does not across 0.

```
## Series: train
## ARIMA(0,1,1)(2,1,1)[12]
## Box Cox transformation: lambda= 0
##
## Coefficients:
##           ma1    sar1     sar2     sma1
##       -0.5665  0.4388  -0.4233  -0.6781
```

```
## s.e.   0.0753  0.1611   0.1141   0.1915
##
## sigma^2 = 0.0002547:  log likelihood = 257.74
## AIC=-505.47   AICc=-504.8   BIC=-492.7
```

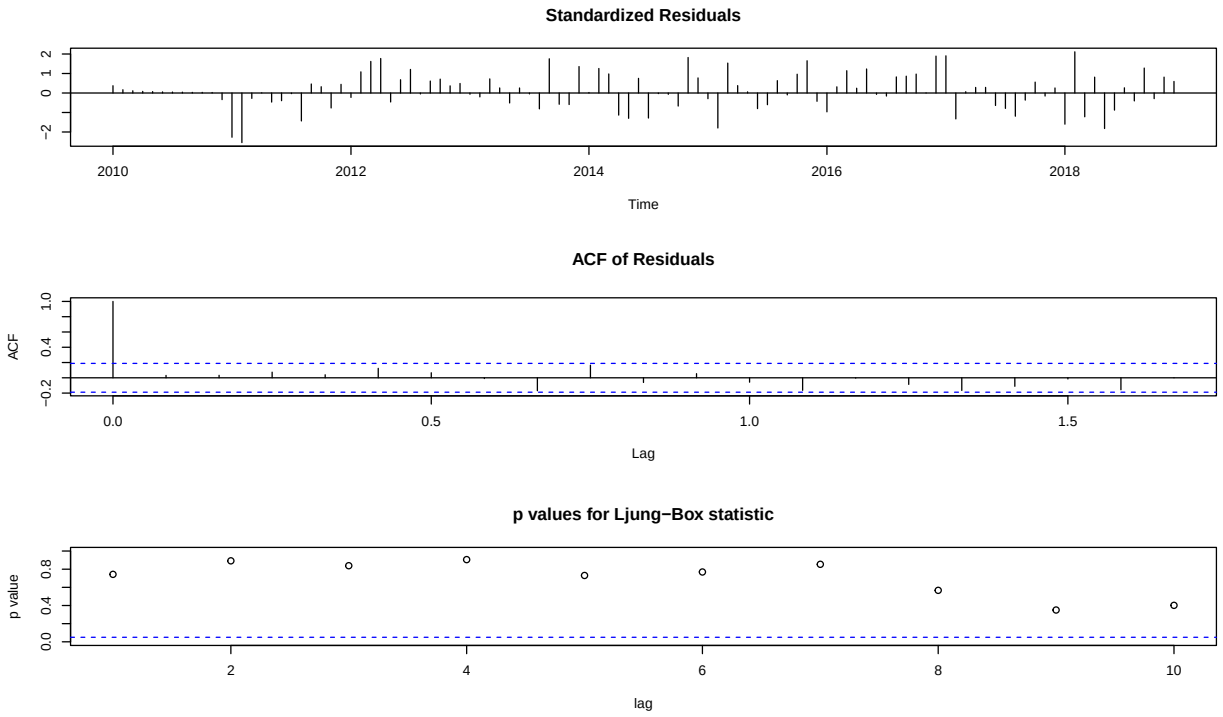

Figure 4: Diagnostics of SARIMA Model

I also use the Ljung-Box test to apply the model diagnostics process. Figure[4] shows the P-value of the Ljung-Box test from lag 1 to lag 120 are all above the dashed line (0.05), thus we can conclude that the residuals of the model is white noise and it's safe to use $SARIMA(0,1,1) \times (2,1,1)_{12}$ for forecasting. The forecast of 01/2019 - 02/2020 by using 01/2010 - 12/2018 is shown in figure[5].

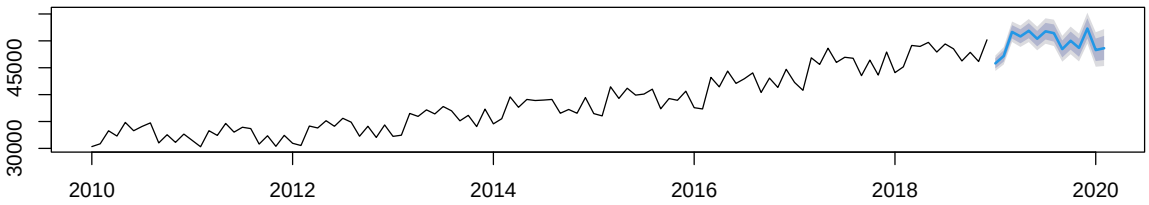

Figure 5: Forecast of SARIMA Model, Lag = 14

Exponential Smoothing Based Forecast

Since the time series has a seasonal component, another way to capture seasonality is using Holt-Winters seasonal method. After we fit the model and check the residuals, it turns out that the residuals is white noise, so it's safe to use this $\hat{X}_t(l) = \hat{a}_t + \hat{b}_t l + \hat{s}_{t+l-d}$ to predict the 14 values along with the forecast intervals and calculate the accuracy.

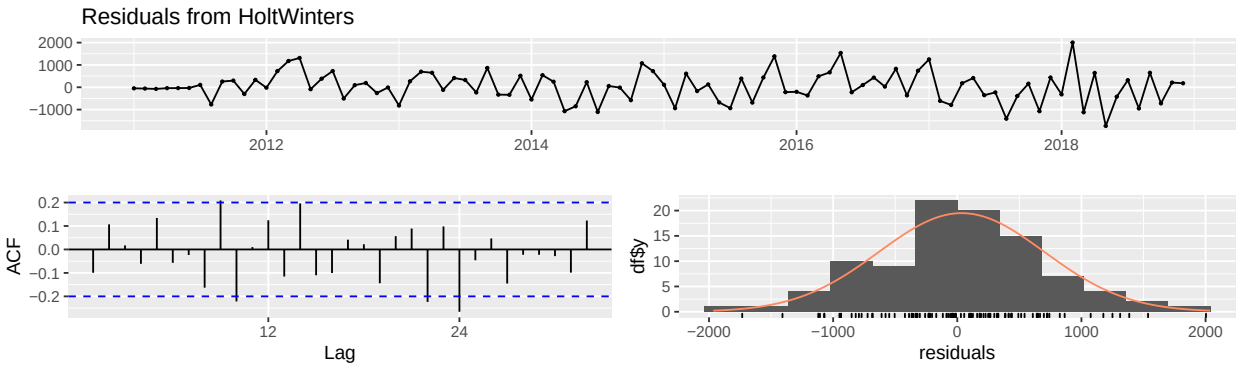

Figure 6: Diagnostics of Holt-Winters Seasonal Model

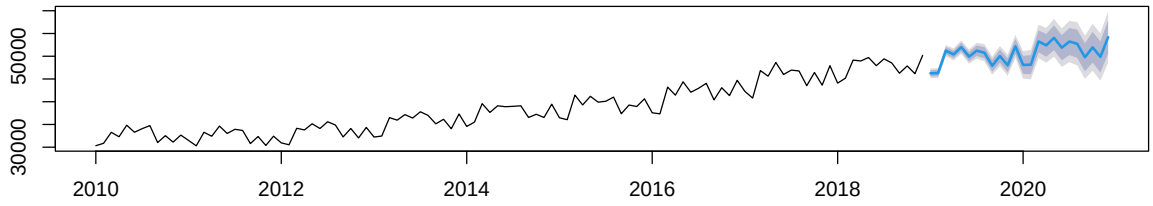

Figure 7: Forecast of Holt-Winters Seasonal Model, lag = 14

Model Comparison and Conclusions

Based on the table below, comparing the test accuracy between different forecast methods, we can conclude that Holt-Winter seasonal model performs slightly better than $SARIMA(3,1,0) \times (1,0,0)_{12}$. From my perspective, since the Holt-Winters seasonal method comprises not only the forecast equation but also three smoothing equations (for the level a_t, for the trend b_t, and the seasonal component s_t), maybe moderate complexity can lead to more accurate.

Criteria	ARIMA	HoltWinters
RMSE (Root Mean Squared Error)	17493.2347	17141.3930
MAE (Mean Absolute Error)	17464.5758	17141.3930
MAPE (Mean Absolute Percentage Error)	54.084376	52.991625

Discussion about Futher Study

Due to the pandemic, there were two significant sudden changes in the data after February 2020. Based on the real background, we found that these two-time points are related to the federal government's policy on restaurant opening status and the vaccine coverage.

Since we are still in the unpredictable and repeated pandemics, it is still difficult to find regularities in the data after 02/2020 (the first outbreak of the pandemic). Thus I have not done **Intervention Analysis** for the time being, but I am still look forwarding to continuing to study this area.