U.S Retail Trade and Food Services TS Analysis

Wanning Wang

Introduction of the Data

- This data is from U.S. Census Bureau Monthly Retail Trade Survey.
  Source: https://www.census.gov/econ/currentdata/dbsearch?program=MRTS&startYear=2010&endYear=2019&categories=44X72&dataType=SM&geoLevel=US&notAdjusted=1&submit=GET+DATA&releaseScheduleId=
- Data Background:
  The Monthly Retail Trade Survey provides current estimates of sales at retail and food services stores and inventories held by retail stores.
  Source: https://www.census.gov/econ/overview/re0400.html
- Data Selection for Analysis:
  The data for analysis contains the recent ten years of data from 2010_Jan-2019_Dec; for ten years, monthly data can mainly capture the trend and seasonal patterns of the national retail sales behavior. The value is the total sales value of retail trade and food services estimates each month.
- Dataset Name:
  44X72: Retail Trade and Food Services: U.S. Total — Not Seasonally Adjusted Sales - Monthly [Millions of Dollars]

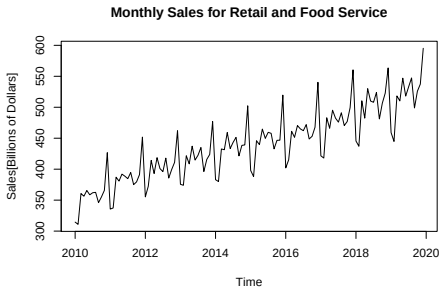
Purpose of Analysis:

- The purpose is to carry out a time series analysis in order to explore the data features to see the seasonal and trend pattern over time. At the same time, explore and compare the performance of different forecasting methods.

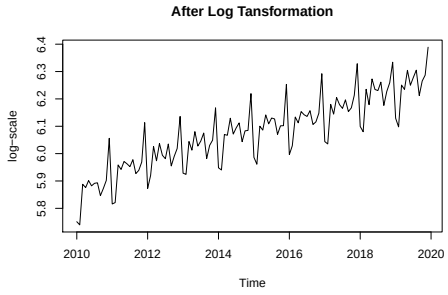
Final Model and Conclusions:

- From the result based on AICc and evaluations between different models, we choose SARIMA(0,1,1)(2,1,2)[12] as the final forecasting model. In this analysis, I introduced the method of how we chose the best ARIMA model derived from different model selection methods based on AICc. This analysis also carried out a comparison of the forecast performance between ARIMA and the non-Arima model.

1. Data Description and Transformation

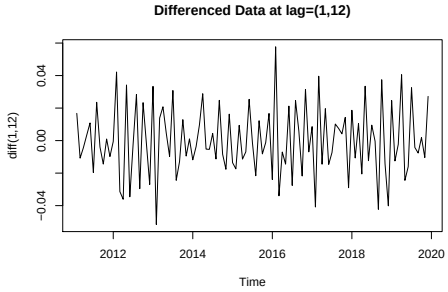


- As we can see from the data, the data has both trend and seasonal components. The variance is increasing as time goes on. This might due to the seasonal hilidays. To stabilize the data, we would like to take a log transformation of the data. After taking the log, the data looks with stable variance.



2. Identifying the ARIMA Model

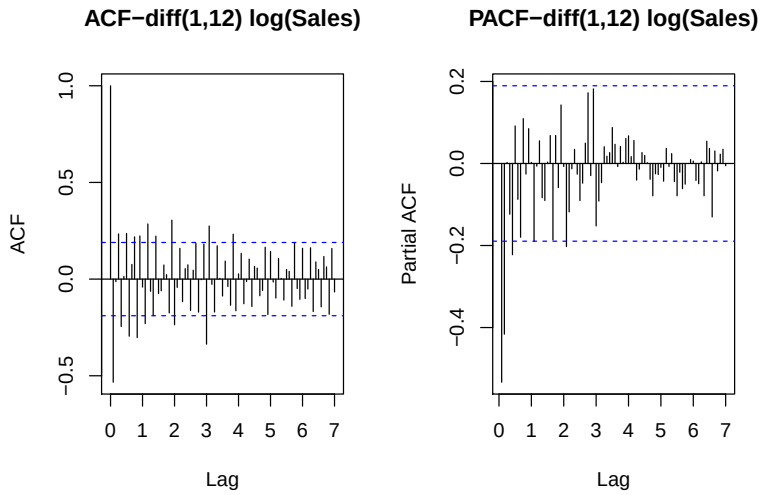
Take differencing to make data stationary



Test for stationarity

- Results from Augmented Dickey-Fuller Test:
  Dickey-Fuller = -7.2671, Lag order = 4, p-value = 0.01 alternative hypothesis: stationary
- From the plot, we can see the data after stabilized and removed trend and seasonal component, it shows white noise. From the result of the Dickey-Fuller Test, we can see p-value less than 0.05; thus, we reject the null hypothesis of non-stationary.

Plotting ACF and PACF



- From the ACF and PACF we can see that ACF shows strong correlation at seasonal lags, and PACF shows decaying and has clustering patterns at seasonal lags, so we consider a seasonal arima model.

Identify Order of the Model

To identify the non-seasonal ARIMA component: ACF cuts-off at lag 1; PACF seems cuts-off at lag2 also decays to zero. Thus, we conside p=0,1,2, q=0,1

To identify the seasonal ARIMA component: ACF decays to zero at seasonal lags, and PACF cuts-off at sesaonal lag 2. Thus, we consider P=0,2; Q=1,2 ; S=12

Hence, possible SARIMA model: $(1,1,1) \times (0,1,2)_{12}$,$(0,1,1) \times (0,1,2)_{12}$, $(0,1,1) \times (2,1,1)_{12}$, $(1,1,1) \times (2,1,1)_{12}$, $(2,1,1) \times (2,1,1)_{12}$, $(0,1,1) \times (2,1,2)_{12}$,$(2,1,0) \times (2,1,1)_{12}$

3. Model Building

Use Arima function fit model with different orders and seanonal orders, indentifing period=12 and lambda=0.

- fit <- Arima(retail_ts,order=c(0,1,1),seasonal = list(order=c(0,1,2),period=12),lambda=0)

3.1 Model Selection Based on AICc

SARIMA_Model	AICc	BIC
(0,1,1)(2,1,2)[12]	-594.79	-579.59
(0,1,1)(2,1,1)[12]	-587.74	-574.97
(2,1,0)(2,1,1)[12]	-586.82	-571.62
(1,1,1)(2,1,1)[12]	-586.07	-570.88
(2,1,1)(2,1,1)[12]	-584.58	-567.00
(0,1,1)(0,1,2)[12]	-582.96	-572.66
(1,1,1)(0,1,2)[12]	-581.95	-569.18

Lowest AICc: ARIMA(0,1,1)(2,1,2)[12]
As we can see from the result, SARIM$(0,1,1) \times (2,1,2)_{12}$ has lowest AICc. IF we based on AICc, this one would be our best SARIMA model.

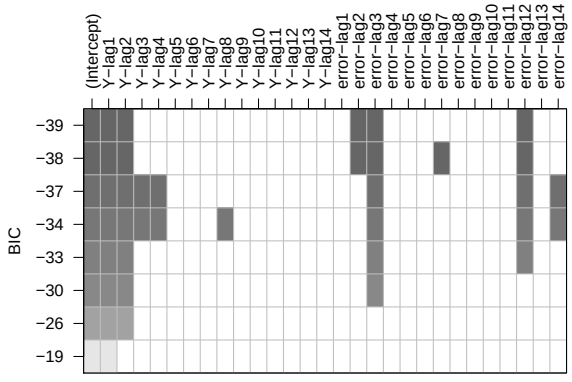
3.2 Auto Arima Model Selection

- fit.auto= auto.arima(retail_ts,lambda = 0)
  fit.auto$aicc

```
## [1] -582.9623
```

Interesting finding is that our SARIM$(0,1,1) \times (2,1,2)_{12}$ has lower AICc than using auto.arima function.

3.3 Subset ARMA Model Selection



From first row of the plot, we can see coefficients ϕ_1, ϕ_2 and $\theta_2, \theta_3, \theta_{12}$ are significant, so that we can also choose to fit a ARIMA(2,1,12) model with setting:

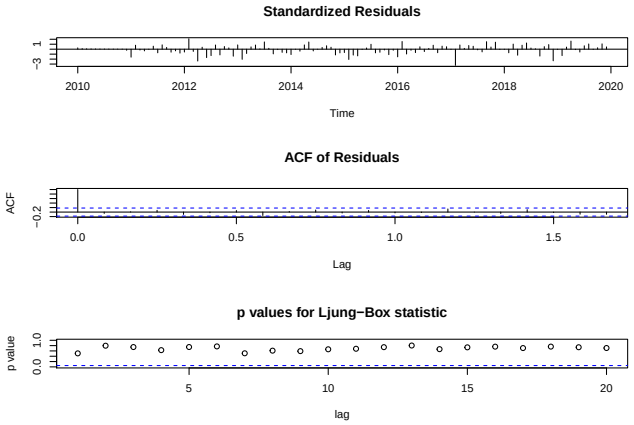
- seasonal=list(order=c(0,1,0),period=12)
  fixed=c(NA,NA,0,NA,NA,rep(0,8),NA)
  lambda=0

AICc of the subset model: -583.01.
The AICc of the model selected from subset model selection method is higher than $\text{SARIM}(0,1,1)\times(2,1,2)_{12}$.

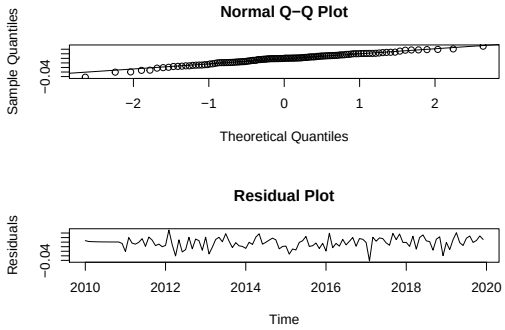
4. Model Check – Model Diagnostics

Residual checking:
We check the residual of $\text{SARIM}(0,1,1)\times(2,1,2)_{12}$ to see if it is adequate model. We can use the Ljung-Box test to see if the residual is white noise, and we can also check the normality of the residuals via Q-Q Plot.

- Ljung-Box test:



- Q-Q Plot & Residual Plot:



From the results we can see the residuals is white noise. The model is adequate.

5. Parameter Estimation

SARIM$(0,1,1)\times(2,1,2)_{12}$ **:**

```
## Series: retail_ts
## ARIMA(0,1,1)(2,1,2)[12]
## Box Cox transformation: lambda= 0
##
## Coefficients:
##          ma1    sar1     sar2     sma1    sma2
##      -0.6618  0.8253  -0.8196  -1.3168  0.8478
## s.e.  0.0718  0.1294   0.1043   0.3632  0.4890
##
## sigma^2 estimated as 0.0001613:  log likelihood=303.81
## AIC=-595.63   AICc=-594.79   BIC=-579.59
```

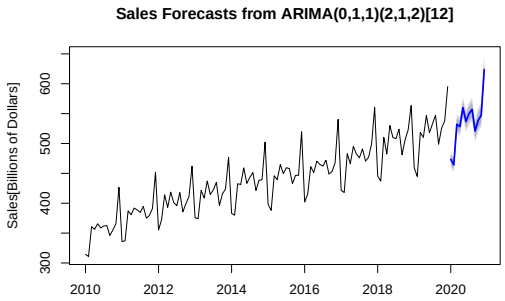
From the result we can see the coefficent for sma2 $\Theta_2 = 0.8478$ is not statistical significant since 0 is included the confidence interval of the estimate ($0.8478 - 1.96 * 0.489 = -0.11, 0.8478 - 1.96 * 0.489 = 1.8$; CI:(-0.11,1.8)). All other coefficients - the estimates of the parameters are statistical sigificant.

- Parameter estimates: $\theta_1 = -0.6618$, $\Phi_1 = 0.8253$, $\Phi_2 = -0.8196$, $\Theta_1 = -1.3168$, $\Theta_2 = 0.8478$

6. Forecasting

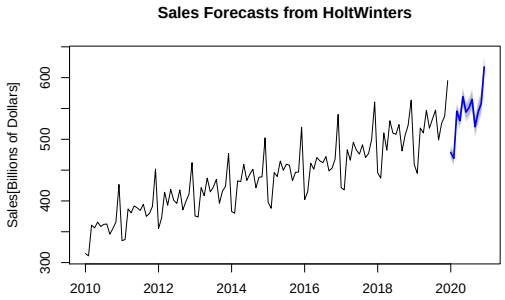
SARIMA Forecast

- We choose our selected model SARIM$(0,1,1) \times (2,1,2)_{12}$ based on AICc criteria to forecast future 12 month values.



Holt-Winters Forecast

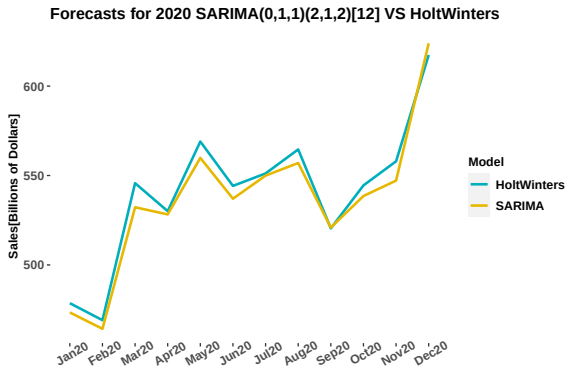
- We also choose to forecast using Holtwinters method to compare.



Both forecast method give similar results. They follow the trend and seasonal patterns.

Forecasts Value SARIMA Vs. HoltWinters:

- Compared the forecasts values from the two model (2020-Jan to 2020-Dec):



From the plot we can see SARIMA(0,1,1)(2,1,2)[12] model forecasts are lower than the HoltWinters forecast at some month. But generally, they are fairly consistent.

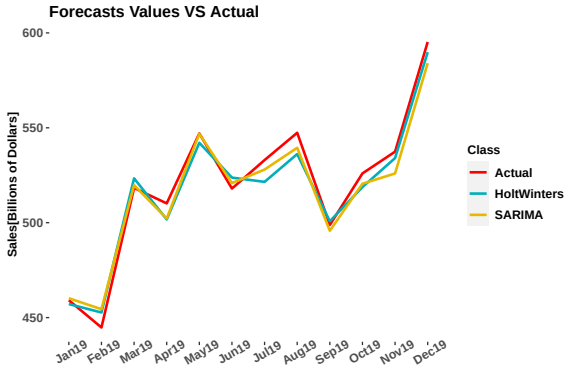
7. Evaluation of Forecast Accuracy

To compare their forecast accuracy, we use 2010-2018 data as training data and the last year(2019) as test set to compute the MAPE of the two models. Lower MAPE score means better forecast accuracy.

Model	MAPE
SARIMA (0,1,1)(2,1,2)[12]	1.073
HoltWinters	1.197

We can see SARIMA model has better forecast accuracy on the test data set.

- Visulize Forecasts from SARIMA, HoltWinters VS Actual Values



- Holtwinters is the forecast value derived from Hlotwinters model, SARIMA is the forecast value derived from SARIMA(0,1,1)(2,1,2)[12] model, Actual is the value of the test data set.

- As we can see from the plot, before Oct forecasts value from SARIMA(0,1,1)(2,1,2)[12] model is more fit with the actual values. However, Nov and Dec the Holtwinters performs better.

Order of Minimum AICC AR Model

```
## Series: residuals(fit.sarima.10)
## ARIMA(0,0,0) with zero mean
##
## sigma^2 estimated as 0.0001371:  log likelihood=363.41
## AIC=-724.82   AICc=-724.79   BIC=-722.03
```