Seasonal Time Series Model Identification for US retail Sales in the USA (1992,2018)

By: Shafim Khoda

Executive Summary

This project is an analysis of the seasonal time series model of retail sales in the United States. It is seasonal because the time series plot has recurring spikes around Black Friday and the holidays. This plot needed a transformation with a gamma value of about 0.5. Even though this wasn't the most ideal setting we continued differencing at this model and then proceeded to test multiple models that could have been a good fit. The models that were examined were SARIMA$(1,1,1)(1,0,1)_{12}$, SARIMA$(1,1,1)(1,1,1)_{12}$, and a SARIMA$(2,1,2),(2,1,2)_{12}$. The two best models from what we examined were the SARIMA$(2,1,2),(2,1,2)_{12}$ and the SARIMA$(2,1,1),(2,1,2)_{12}$ models. Since the models are too similar, the best model to pick would be the SARIMA$(2,1,1),(2,1,2)_{12}$ which has fewer terms. This is because of the principle of parsimony.

Background

Our seasonal time series observations are the monthly retail sales in the United States from 1992 to 2018. We chose this dataset because it is clearly a seasonal time series with more people spending during Black Friday and before Christmas than other parts of the year. We also chose this because we wanted to see if there would be a decline coming in the recent years to indicate a recession. We chose to go back to 1992 because we wanted to see the changes over time and see if the great recession of 2008 would be able to give us more information to predict future recessions. We did not have any missing observations in this set because the US Census Bureau continually updated the data.

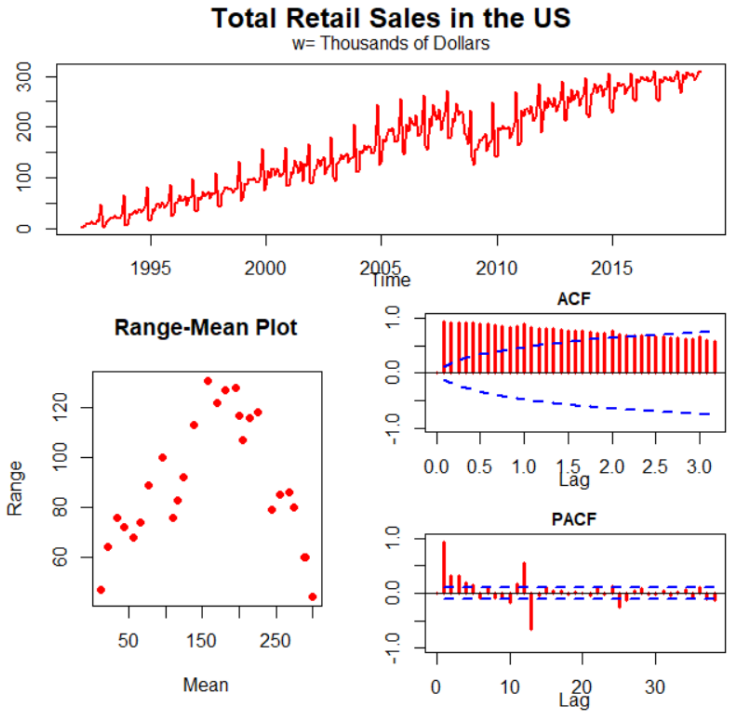


Figure 1 Original Iden Output No Transformation

As you can see from the original Iden output above the trend is constantly increasing with some seasonality peaks and valleys. Also, in 2008 you can see a distinct drop in the data because of the recession and nobody being able to spend as much money as normal. Then the data rebounds and starts increasing again with less seasonality and starts to flatten out. By looking at the range mean plot it appeared as though that we needed a transformation. After testing many values for gamma we settled with gamma=.5.

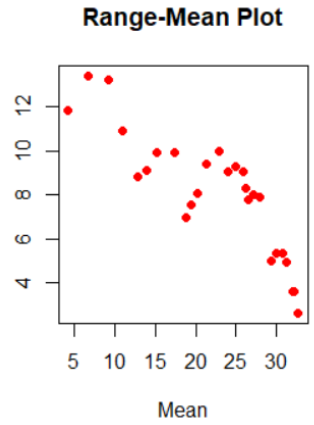
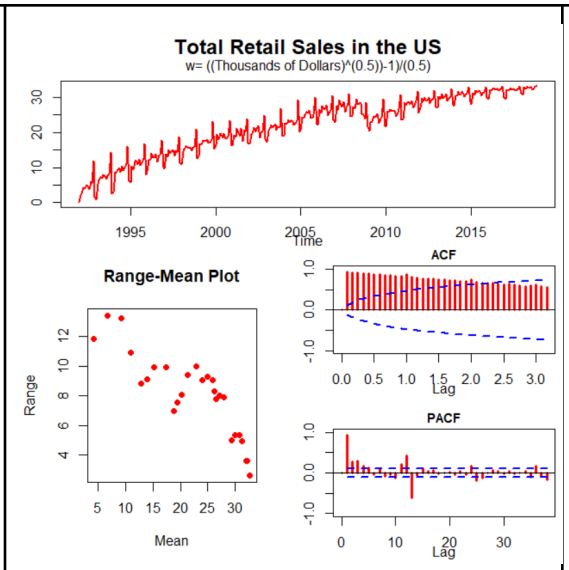
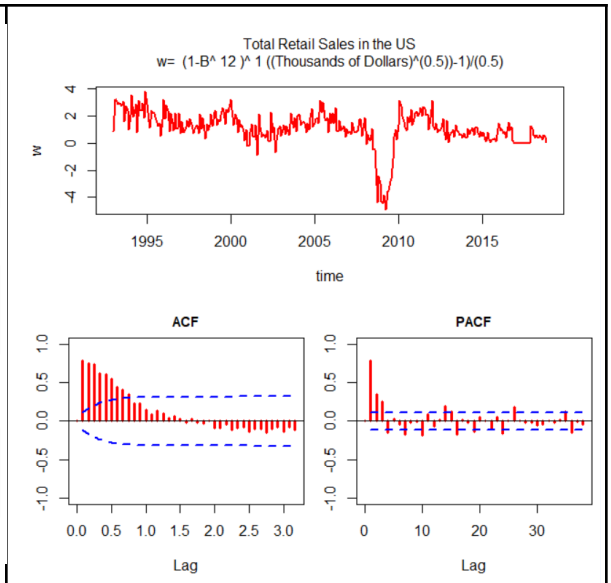
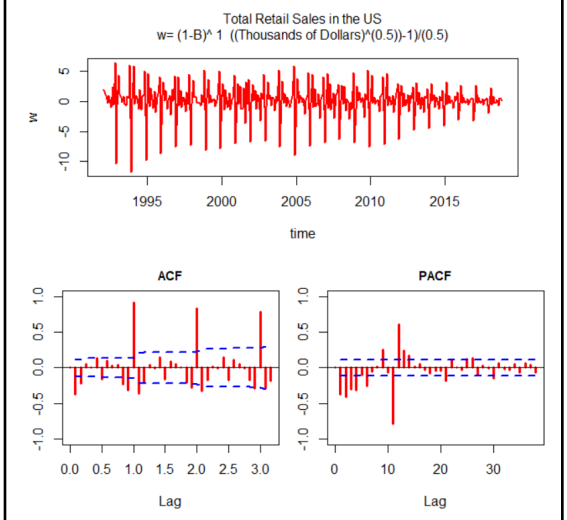
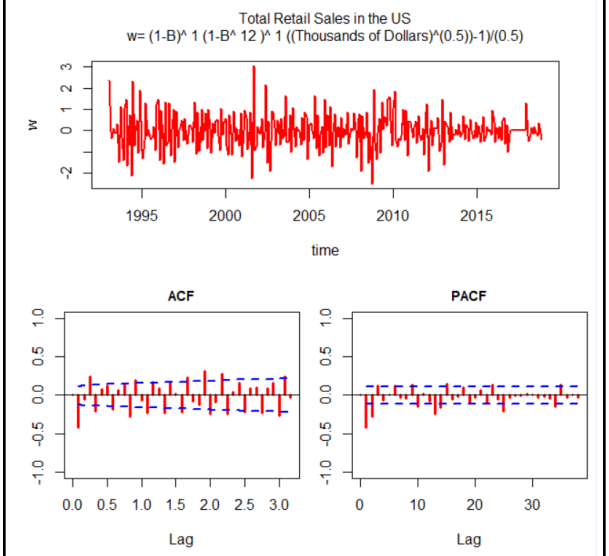


Figure 2 Transformed Range-Mean Plot

This is still far from ideal; however, this is the only transformation that had the weakest correlation between the range and mean. The reason that this is so difficult is because as seen in the original Iden realization vs time plot, there was increasing seasonality and the decreasing seasonality. So, we stuck to this transformation throughout our analysis and continued on to differencing.

Identification

Iden	D=0	D=1

d=0		
d=1		

We didn't want to take too many seasonal differences to introduce additional seasonality so by looking at the table above, we decided to test the following models: SARIMA(3,0,0),$(1,1,0)_{12}$ (ACF dying down and PACF cutting off after 3 and seasonal lags on ACF cutoff after 1 and PACF dies down), SARIMA(1,1,1)$(1,0,1)_{12}$ (both the ACF and PACF die down and seasonal lags of the ACF/PACF both die down as well), SARIMA(1,1,1)$(1,1,1)_{12}$ (both the ACF and PACF die down and seasonal lags of the ACF/PACF both die down as well), and a SARIMA(2,1,2),$(2,1,2)_{12}$ (both the ACF and PACF die down and seasonal lags of the ACF/PACF both die down as well).

	SARIMA Model $(p,d,q),(P,D,Q)_{12}$				
Model	(3,0,0),(1,1,0)	(1,1,1),(1,0,1)	(1,1,1),(1,1,1)	(2,1,2),(2,1,2)	(2,1,1),(2,1,2)
d	0	1	1	1	1
D	1	0	1	1	1
S	.3482	.3495	.3428	.3161	.3160
AIC_c	239.1967	278.8385	229.4315	197.1131	195.2299
-2log(Likelihood)	229.1967	268.8385	219.4315	179.1131	179.2299
Ljung-Box χ^2_{38}	p<.001	p<.001	p<.001	p<.01	p<.01

By looking at the table of values above for all of the models it looks as though the best models are the SARIMA(2,1,2),$(2,1,2)_{12}$ and the SARIMA(2,1,1),$(2,1,2)_{12}$. We came to this conclusion because they had the best AIC values and the best -2log(Likelihood) values. Also, the S values were lower than all the other models as well. However, the Ljung-Box p-values are greater, but they are still really low.

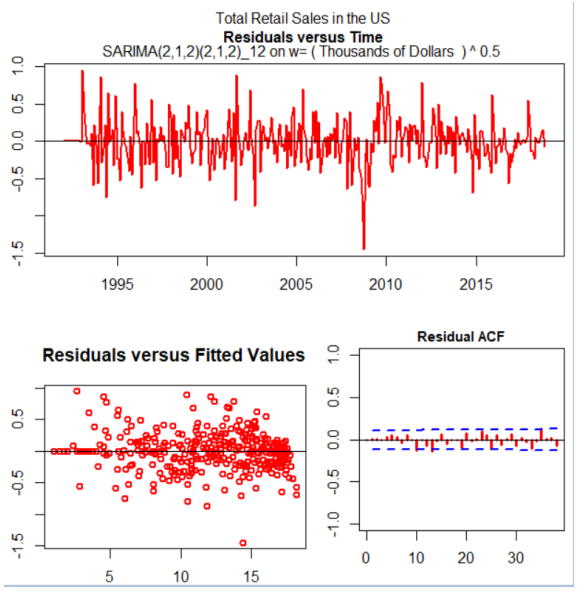


Figure 3 Esti Output for SARIMA (2,1,2),(2,1,2) Part 1

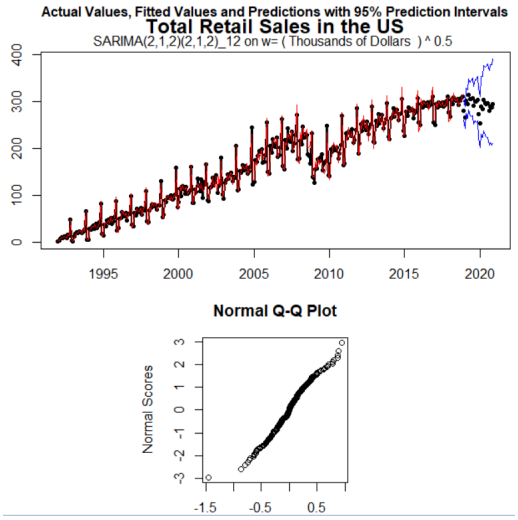


Figure 4 Esti Ouput for SARIMA (2,1,2),(2,1,2) Part 2

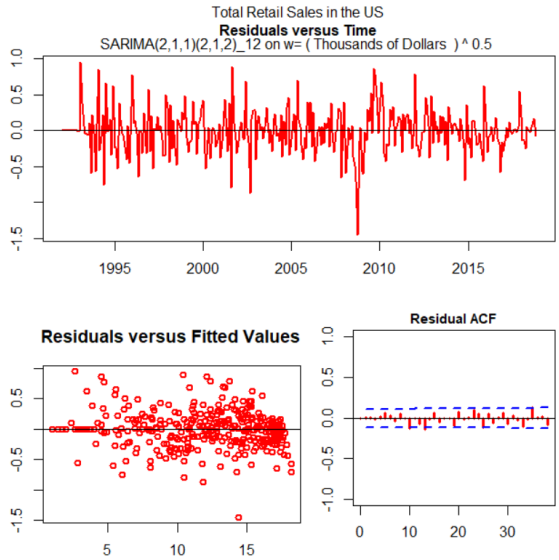


Figure 5 Esti Output for SARIMA (2,1,1),(2,1,2) Part 1

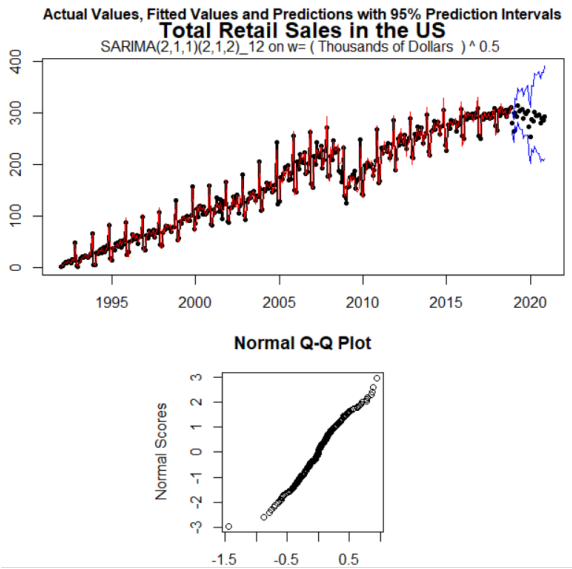


Figure 6 Esti Output for SARIMA (2,1,1),(2,1,2) Part 2

The residual ACF plots for both of these look pretty good with staying in between the bands. The forecasts for both models were very similar with both predicting the almost the same values and having similar prediction intervals as well. The models also indicate that retail sales will be on the decline soon. The normal Q-Q plots also look very similar but also, they both look pretty linear in relation which is a good sign.

Conclusion

While the latter 2 models are about the same, we want to choose the model with the least amount of parameters due to the principle of parsimony. Of the 2 models, SARIMA(2,1,2),$(2,1,2)_{12}$ and the SARIMA(2,1,1),$(2,1,2)_{12}$, the obvious choice here is the SARIMA(2,1,1),$(2,1,2)_{12}$ model because there are less terms in this. This model could be used to predict potential recessions and give the government a chance to proactively implement policies to help and avoid the recession altogether. Also, this could be used by investors to know when to sell their shares and when to buy them. With more data, hopefully the model will get more accurate to the actual values that are attained, however the prediction intervals will continue to grow do to continued variability.