STAT 8003 project

HU Jiamian 3035802768

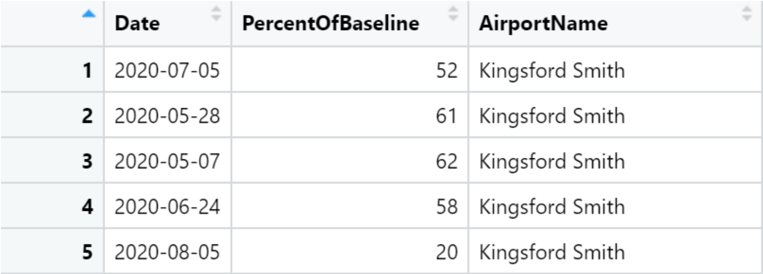
Introduction

The Covid-19 pandemic has influenced the world form multiple aspects, one of the most affected industry is commercial airline industry. A data set provided by Geotab recorded airport traffic from 28 different airports across the world from 16th Mar 2020 to 16th Oct 2020 (215 days). Traffic volume is recorded as a percentage of baseline period traffic. Baseline period is from 1st Feb to 15th Mar 2020. Original dataset can be extracted from here.

In this project, a time series analyses will be performed based on traffic data of airports and a model fitted will be chosen by AIC.

Data pre-processing and Observation

Original data size is relatively large and contains 28 airports with over 5000 records. Many of fields in the data set is not relevant to analyses like City, State and Centroid. I will drop data into from shown in Table 1 to do following analyses:



	Date	PercentOfBaseline	AirportName
1	2020-07-05	52	Kingsford Smith
2	2020-05-28	61	Kingsford Smith
3	2020-05-07	62	Kingsford Smith
4	2020-06-24	58	Kingsford Smith
5	2020-08-05	20	Kingsford Smith

Table 1: Data Sample

As shown in Table 1, data will only include Date, PrecentOfBaseline and AirportName as data fields to perform analysis. However, some airport traffic is highly depending on information that are not related to time series. One example is shown in Figure 1:

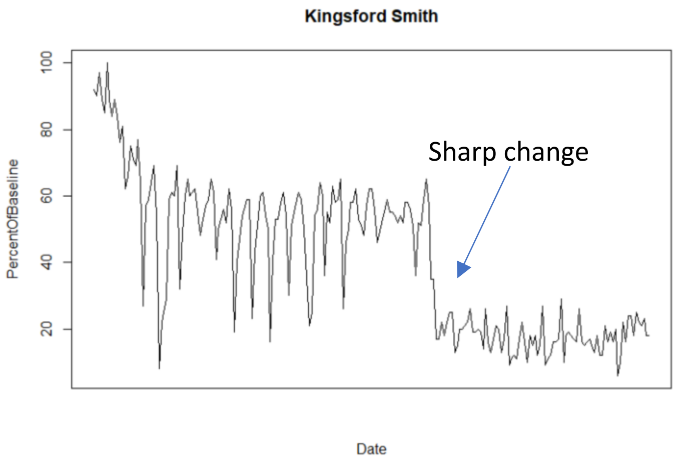


Figure 1: Plot of traffic of Kingsford Smith Airport

STAT 8003 project

HU Jiamian 3035802768

The traffic of Kingsford Smith Airport has a sudden drop during the period and the change my be caused by other events like local travel ban. After the sharp change in the airport also caused the variance of following period traffic decreased. To do analysis of traffic, the selected airport must have following properties:

1. No sudden and persistent change on variance.
2. No sudden and persistent change of traffic volume.
3. Clear seasonality.

According to above three rules, traffic data of Dallas/Fort Worth International Airport is selected for analyses. Plot of traffic over date is shown as Figure 2.

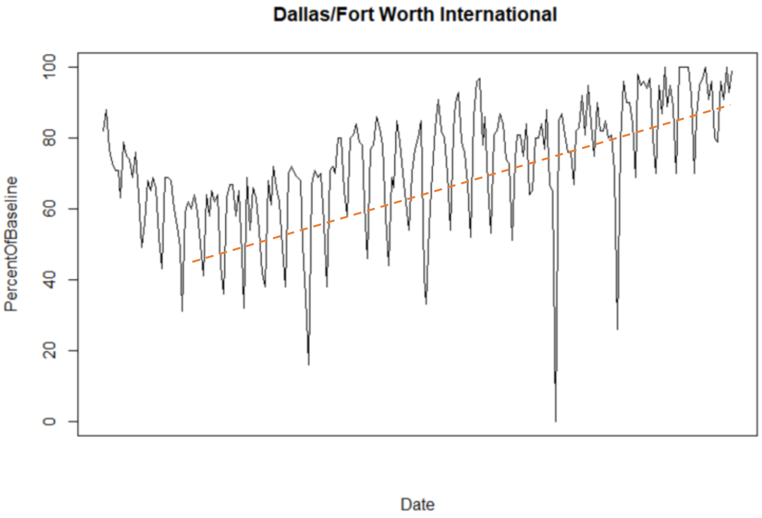


Figure 2: Plot of traffic of Dallas/Fort Worth International Airport

Traffic of Dallas/Fort Worth International Airport shows clear seasonality about a seasonal cycle of seven days and a clear trend of increase which indicates a transform of time series need to be done. The variance during whole period seems not changing too much. ACF and PACF of the traffic is shown below:

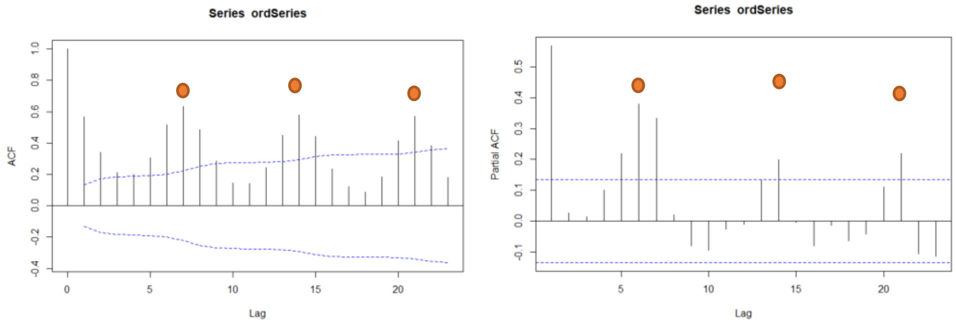


Figure 3: ACF and PACF of traffic of selected airport

Both ACF and PACF shows seasonality. PACF and ACF tends to reach local maximum for each 7-day period as marked by orange dot.

Next start doing Dickey-fuller test on data and getting data shown in Table 2:

Model	**Dickey-Fuller value**	**p-value**	**H_0**
ARIMA (0,0,0)	-4.892	<0.01	Non-stationary

Table 2: Dickey Fuller test on original series

Since p-value is smaller than 0.05, the null hypothesis that the original is not stationary cannot be accepted. The result indicates that differencing will not transform the time series to be non-stationary. Over differencing may be the main concern in proposing models.

Model Proposal & Specification

As shown in Figure 2, the time series shows an upper trend like the example shown in CO_2 level example in lecture note of chapter 8. The first specification of model will be like the example shown in the lecture. A first order differencing will be taken on the series and a seasonal differencing can be taken afterwards. ACF of differencing and seasonal differencing are shown in Figure 4:

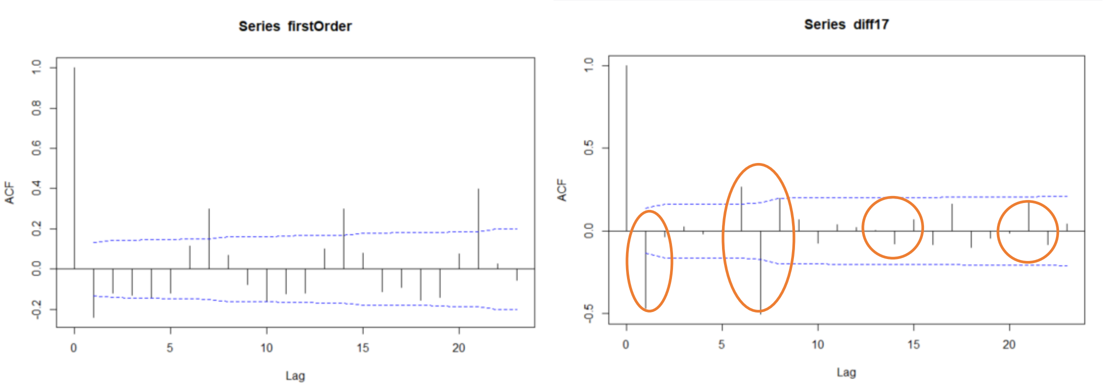


Figure 4: ACF plot of differencing transformation

Figure 4 shows that after taking difference and seasonal difference, the seasonal pattern seems drop dramatically. Only a weak SAR (1) X MA (1) was presented in the orange circles. Time plot of the differenced time series is shown in Figure 5:

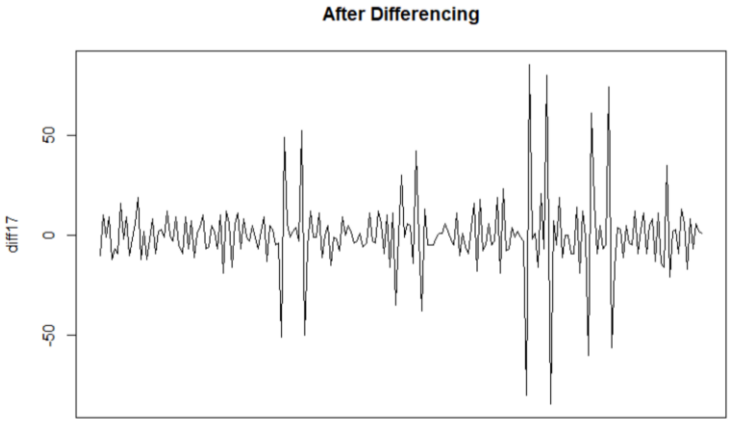


Figure 5: Time Plot of Differenced time series

At this step, the initial proposed model could be formulated as $ARIMA(0,1,1) \times (1,1,0)_7$. Based on the ACF plots and the 7-day seasonality, some other models may work well like $ARIMA(0,1,1) \times (0,1,1)_7$ and as

STAT 8003 project

HU Jiamian 3035802768

$ARIMA(0,1,1) \times (1,1,1)_7$ will be tested on the project. All test statics of possible models like AIC however will listed and compared.

Model Diagnostic

Model I: $\boldsymbol{ARIMA(0,1,1) \times (1,1,0)_7}$

Fitting $ARIMA(0,1,1) \times (1,1,0)_7$ model with maximum likelihood method, getting following parameters:

Parameter	**Value**
MA (1)	-0.9005
SAR (1)	-0.4887
AIC	1617

Table 2: fitted parameters of $ARIMA(0,1,1) \times (1,1,0)_7$ model

Doing Ljung-Box test on the residuals of fitted model with K from 1 to 14, results are summarized in the following model:

Lag	**2**	**6**	**7**	**8**	**13**	**14**
Statistic	2.18	3.08	14.64	15.38	16.01	38.43
χ^2**at 95%**	3.84	11.07	12.59	14.06	21.03	33.36

Table 3: Ljung-Box Test Statistic

As seen from Table 3, the test statistic show there are some inadequacy in the fitted model and the residual still have some seasonality left. The residual seasonality could be more clearly seen from ACF plot of residual below:

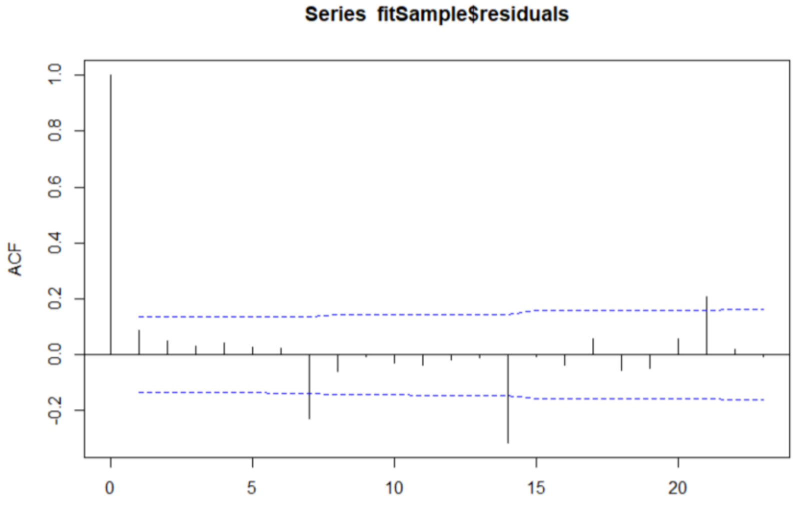


Figure 6: ACF Plot of residuals

As the residual ACF plot shows, ACF are significant at 7th, 14th and 21st lag, which is clearly a seasonality. Therefore, the model may not seem to fit the data at acceptable level. According to the observation, a SMA (1) process is added into Model I to catch the seasonality in ACF function.

Model II: $\boldsymbol{ARIMA(0,1,1) \times (1,1,1)_7}$

Fitting $ARIMA(0,1,1) \times (1,1,1)_7$ model with maximum likelihood method, getting following parameters:

Parameter	Value
MA (1)	-0.8212
SAR (1)	-0.0267
SMA (1)	-0.9157
AIC	1561.11

Table 4: fitted parameters of $ARIMA(0,1,1) \times (1,1,1)_7$ model

Comparing to Model I, absolute value SAR (1) factor drops dramatically from 0.4887 to 0.0267. The dropping trend indicates that the SAR (1) process may not be significant and could consider dropping SAR (1) in Model III.

AIC of Model II is also smaller than Model I. The signal shows model II is a better fitted model compared to Model I.

Doing Ljung-Box test on the residuals of fitted model with K from 1 to 14, results are summarized in the following model:

Lag	**2**	**6**	**7**	**8**	**13**	**14**
Statistic	0.72	2.34	2.56	2.65	5.53	5.68
χ^2**at 95%**	3.84	11.07	12.59	14.06	21.03	33.36

Table 5: Ljung-Box Test Statistic

As Ljung-Box Test statistic shows Model II is adequate at 95% confidence level. Plotting ACF and PACF:

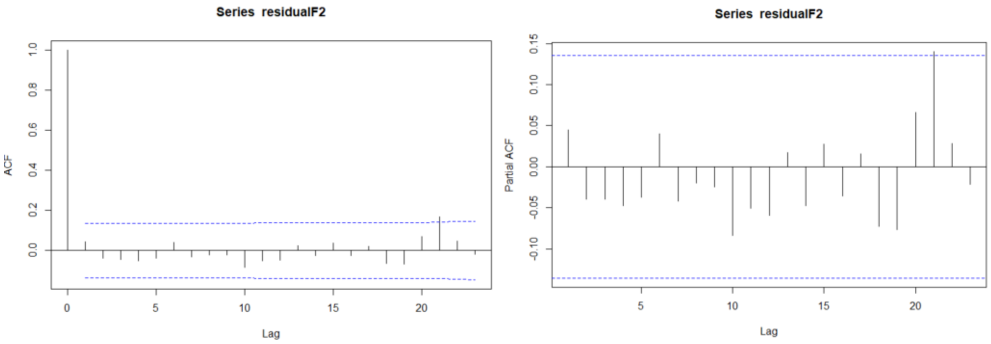


Figure 7: ACF and PACF of residuals of Model II

Resduial ACF and PACF plot shows the correlation between residuals could be condidered to be in-significant. Model II at current stage seems adequate. As stated in the obsvered changes in fitted parameters, SAR (1) in the model seems not significant. In Model III, the SAR (1) will be dropped.

Model III: $\boldsymbol{ARIMA(0,1,1) \times (0,1,1)_7}$

Fitting $ARIMA(0,1,1) \times (0,1,1)_7$ model with maximum likelihood method, getting following parameters:

Parameter	**Value**
MA (1)	-0.8226
SMA (1)	-0.9253
AIC	1559.22

Table 6: fitted parameters of $ARIMA(0,1,1) \times (0,1,1)_7$ model

Compared with Model II, AIC dropped due to less parameters. If model selection based on AIC, the Model III will be the best around proposed models.

Doing Ljung-Box test on the residuals of fitted model with K from 1 to 14, results are summarized in the following model:

Lag	**2**	**6**	**7**	**8**	**13**	**14**
Statistic	0.71	2.19	2.76	2.86	5.66	5.78
χ^2**at 95%**	3.84	11.07	12.59	14.06	21.03	33.36

Table 7: Ljung-Box Test Statistic

Similar to Model II, Ljung-Box test shows Model III is adequate. Plotting ACF and PACF of residuals, finding no significant correlation between residuals:

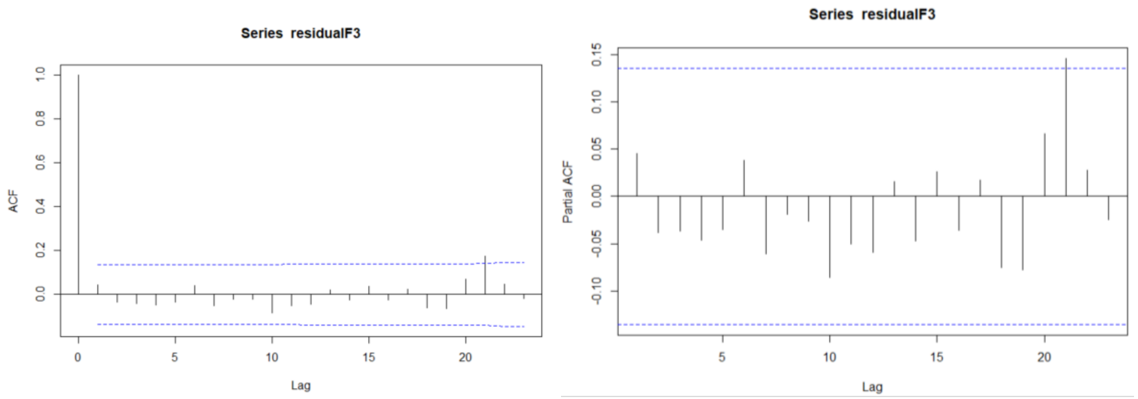


Figure 8: ACF and PACF of residuals of Model III

At this stage, no extra change of order of model could be proposed with clear reason obevered from data provided. However, it is not suciffient to draw conclusion that the Model III will be the situated model to model passanger traffic of the Dallas/Fort Worth International Airport. Considering the complex reasoning and intuition required to draw seasonal model from the ACF plot alone, a Brute-Force Search (BFS) algorithm will be introduced in next session. BFS algorithm will test all possibile combinations of $(p,d,q) \times (P,D,Q)$ with given order M. M is the largest number among $(p,d,q) \times (P,D,Q)$.

More on Model selection: Brute-Force Search (BFS)

Observing candidate models, they can be viewed as combination of ARIMA model (p,d,q) and Seasonal ARIMA model (P,D,Q). To find the most suitable model, it is posibile to do a BFS on the parameter space with p, d, q, P, D and Q ranged within 0 and a maxium number M. For example, if M equal to 2, (p,d,q) and (P,D,Q) can be ranged from (0,0,0) to (2,2,2). Detail is illstrated in following figure:

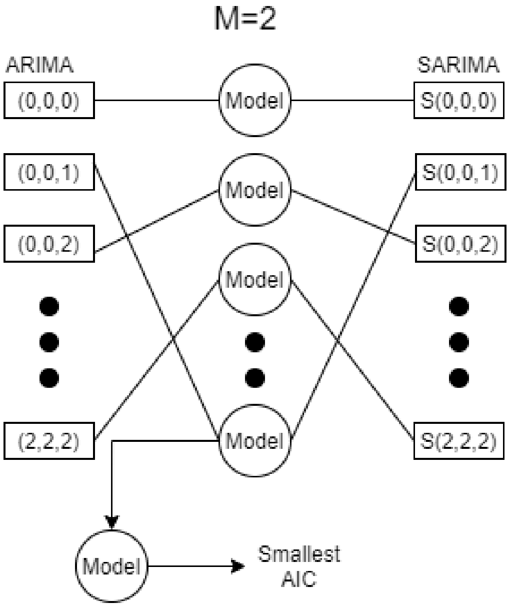
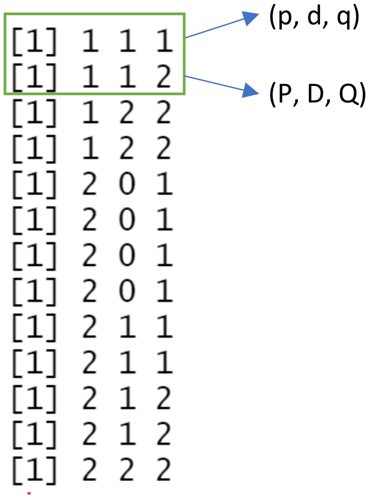


Figure 9: BFS

As proven in the lecture, when testing model it is dangerous to increase (p,q) at same time. In BFS, all possibilie (p,q) are tested and some of them failed as shown in Figure 10:



```
[1] 1 1 1
[1] 1 1 2
[1] 1 2 2
[1] 1 2 2
[1] 2 0 1
[1] 2 0 1
[1] 2 0 1
[1] 2 0 1
[1] 2 1 1
[1] 2 1 1
[1] 2 1 2
[1] 2 1 2
[1] 2 2 2
```

Figure 10: Failed model. (p,d,q) in odd line and (P,D,Q) in even line

After running BFS with M=2, the smallest AIC across all possibile models is 1559.22. The model with that AIC is exactly Model III. If based on AIC, Model III will be selected as best-fitting models among all possibile combinations of (p,d,q) and (P,D,Q) between (0,0,0) and (2,2,2).

Forecast

Five forcasted values and difference between real value and forecasted value are shown in following table

Series	**Forecast**	**Real value**	**Difference**
1 (12/10/2020)	89.885	96	-6.115
2 (13/10/2020)	98.181	91	7.181
3 (14/10/2020)	102.623	100	2.622
4 (15/10/2020)	97.560	93	4.560
5 (16/10/2020)	96.899	99	-2.101

Table 8: Forecasting value

The forecasting value are very close to real test value that did not involved in training. Prediction error in this case did not reach 10% in all predictions.

AIC is one of the most powerful model selection toll since it not only encourage model with largest MLE but also impose penalty on models with large parameters to avoid overfitting. Following table is showing relationship between three models' prediction error sum of squqres and AIC:

Model	**AIC**	**Error sum of squares**
Model I	1617	130.976
Model II	1561.11	126.171
Model III	1559.22	121.05

Table 9: AIC and Error sum of squares

As seen in Table 9, error sum of squares decreases as AIC decreases. The result proves AIC is a powerful indicator for selecting better model in this time series data.

However, the model fitted are only suitable for airport traffic data within the first criteria. When encountering non-stationary data like Sydney Kingsford Smith Airport, it is hard to find a time series like this to model it.

Conclusion

Passenger traffic of Dallas/Fort Worth International Airport can be fitted into $ARIMA(0,1,1) \times (0,1,1)_7$ model. The model proposed following experience provided in textbook may not be a good fit on data. However, with step-to-step analyses based on plotted ACF and PACF on fitted model, we can finally reach the model with better predictability. Proven by BFS, sometime increase (p,q) of ARIMA model could lead to disaster.

STAT 8003 project

HU Jiamian 3035802768

Reference:

1. Shin, T. (2020, October 19). COVID-19's Impact on Airport Traffic. Retrieved November 01, 2020, from https://www.kaggle.com/terenceshin/covid19s-impact-on-airport-traffic
2. Cryer, J. D., & Chan, K. (2011). *Time series analysis with applications in R*. New York: Springer.