Time Series Final Report

Name: Dhanabordee Mekintharanggur, Student ID: T11902203

Time Series Background

Singapore is home to one of the world’s largest transportation hubs of Asia, the Changi Airport. Serving over 68 million air travelers, as of 2019, it is the 18th largest airport in the world. Not only that Changi is serving over 7,400 flights landing to and departing each week, but it is also a fascinating tourist destination of Singapore. With the multi-story lifestyle hub connecting its 3 terminals, Changi airport offers spectacular indoor gardens, hundreds of eateries, entertainment corners, and even movie theaters. With millions traveling in and out of Singapore each year, analyzing the number of passengers with respect to time can provide an interesting point of view about the airport's busiest time of the year, and provide some managerial insights through making forecasts of the passenger growth using time-series analysis methodologies.

The dataset for this report is obtained from https://www.kaggle.com/datasets/gohsoonheng/air-passengers-sgp-changi-airport-for-past-10yrs, which in turn extracted the data from the public dataset https://data.gov.sg. It contains information on the monthly number of passengers passing through Changi Airport between January 1980 to October 2019.

Exploratory Analysis

Trend and Seasonality

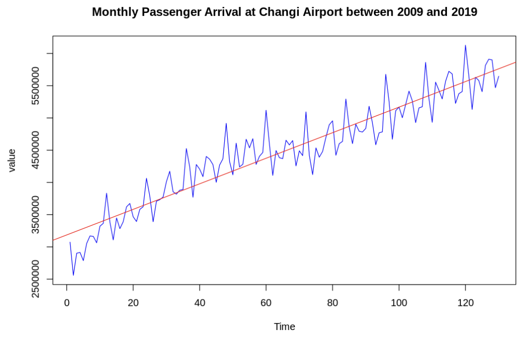


Fig 1. Original Data Plot

Upon visually inspecting the original data plot in **Fig.1**, a clear existence of the upward trend can be observed. Additionally, the plot shows a period spike after every 12 data points. It can be inferred that the amount of passengers traveling through Changi is seasonal but is not stationary. By applying the Augmented Dickey-Fuller (ADF) test, the null hypothesis that the time series has a unit root (non-stationary) is tested. The ADF test result yields a p-value of 0.05717, which exceeds the critical value of 5%. Thus, it fails to reject the null hypothesis and suggests that the time series is non-stationary, supporting the previous visual inspection.

Transforming

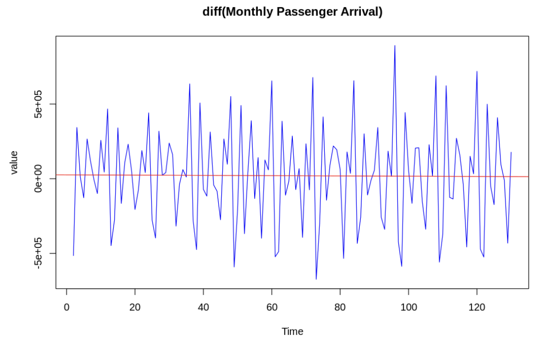


Fig 2. Differenced Data Plot

To effectively analyze a time series using the models such as SARIMA, a stationary time series is a requirement. Differencing is a process of computing the difference between two consecutive data observations. By differencing the time series, the mean can be stabilized by removing the effects of changes in the level of the data. Differencing reduces the existence of trend and seasonality. After differencing, the plot shows the data center at the static mean close to zero and the trend appears to be eliminated as shown in **Fig.2**.

Autocorrelation

Once the time series was made stationary by differencing, the next stop before fitting into a SARIMA model is to identify the relevant parameters by detecting the autocorrelation that exists in the differenced time series. From the ACF of the original data in **Fig.3**, it is observed that the autocorrelation is significant for multiple lags. Additionally, there is a very small spike periodically, suggesting that there is seasonality. From the overall picture, it appears that there is a large diminishing curving trend in the long term and a very small convex curve periodically. This suggests a major interseasonal AR trend and also possibly an intraseasonal AR trend. By looking at the ACF of the differenced data in **Fig.4** the seasonality becomes much clearer. It also reveals a major MA characteristic of spiking correlations every 12 periods.

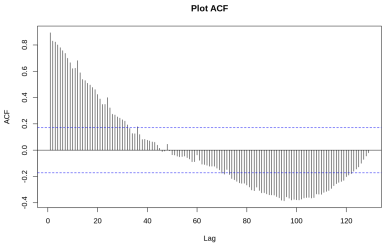


Fig 3. ACF of Original Data

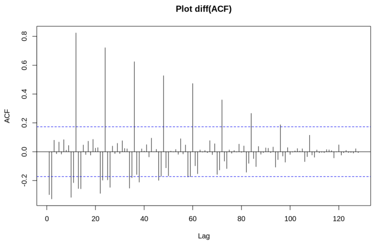


Fig 4. ACF of Differenced Data

Taking a look at the ACF of the differenced data with lag 12 in **Fig. 5**, the overall diminishing autocorrelation is observed, pointing to the conclusion that there is an interseasonal AR trend. The autocorrelation of the double differenced data in **Fig.6** suggests lag 1, 11, 12, and 14 are significant, and might be reasonable to investigate further.

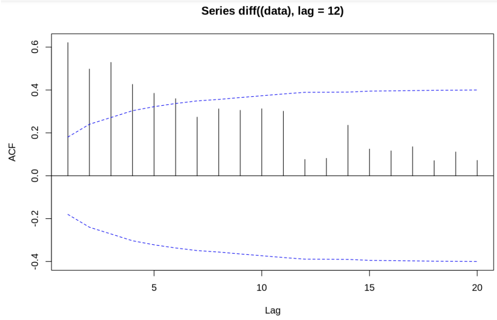


Fig 5. ACF of Differenced Data with Lag 12

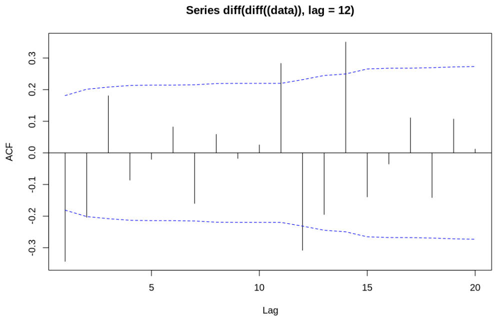
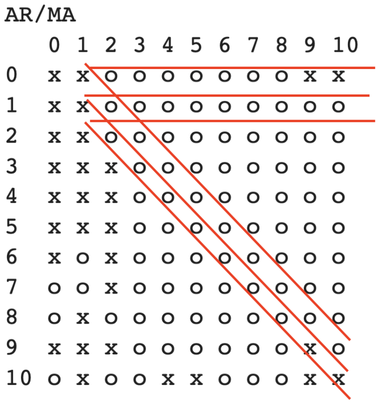


Fig 6. ACF of Double-differenced Data with Lag 12



AR/MA	0	1	2	3	4	5	6	7	8	9	10
0	x	x	o	o	o	o	o	o	o	x	x
1	x	x	o	o	o	o	o	o	o	o	o
2	x	x	o	o	o	o	o	o	o	o	o
3	x	x	x	o	o	o	o	o	o	o	o
4	x	x	x	o	o	o	o	o	o	o	o
5	x	x	x	o	o	o	o	o	o	o	o
6	x	o	x	o	o	o	o	o	o	o	o
7	o	o	x	o	o	o	o	o	o	o	o
8	o	x	o	o	o	o	o	o	o	o	o
9	x	x	x	o	o	o	o	o	o	x	o
10	o	x	o	o	x	x	o	o	o	x	x

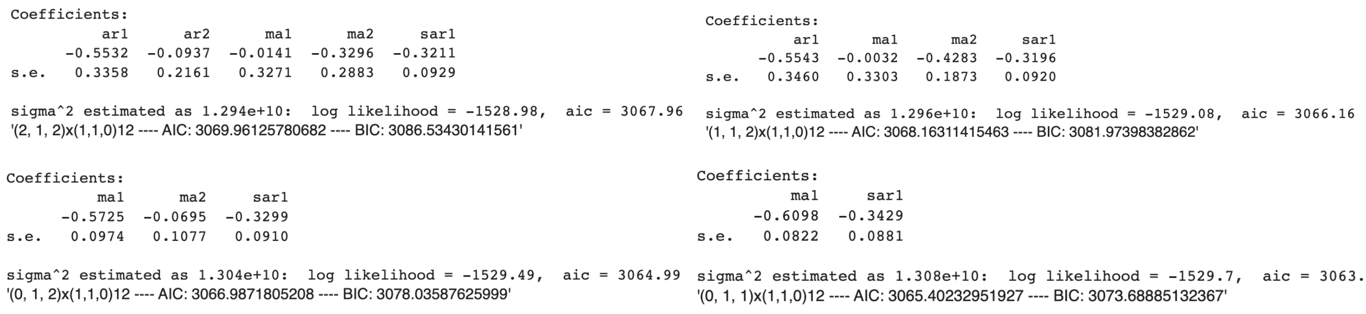
Fig 7. EACF Table

With the ACF plot, the AR and MA characteristics can be extracted. However, the ARMA characteristics are only hinted at. By also utilizing the EACF table, it is possible to identify the ARMA parameters that have significance. This is accomplished by comparing the p-value with the significance level of 5%. By observing where the corner of the triangle of significant values forms, it is possible to determine the likely AR and MA parameters. As shown in Fig. 7, there appear to have multiple possible corners for the triangle of significant p-values, as outlined by the red line. EACF appears to be strongly suggesting a MA(2) characteristic, but the AR part could be AR(0), AR(1), or AR(2). Further analysis is required to determine the best parameter for the SARIMA model to fit the data.

Model

With all the parameters determined from the ACF plots and the EACF table, the SARIMA model can be created. Since it is still unclear which parameters (p, d, q, P, D, Q) will yield the best result when using SARIMA to model the data. Thus, each set of parameters suggested by the EACF table will be analyzed. Using the Arima function of R to fit the model, the significance of each coefficient can be identified by comparing it to the standard error. The smaller the standard error, the more accurate the representation of each coefficient estimate. As shown in **Fig. 8**, the coefficient of each AR, MA, and SAR term are presented alongside its standard error. The coefficient is significant when its magnitude is more than twice its standard error. By this criterium, it can be found that SARIMA (2,1,2) x $(1,1,0)_{12}$ has insignificant coefficients ar2, ma1, and ma2 and is not a good fit for the dataset. Moving on to SARIMA (1,1,2) x $(1,1,0)_{12}$, only ma1 is insignificant, pointing in the right direction. With SARIMA (0,1,2) x $(1,1,0)_{12}$ only ma2 is insignificant, and the standard errors are much lower than in previous models. Finally, SARIMA (0,1,1) x $(1,1,0)_{12}$ is found to be the model where the coefficients are most significant and well beyond two times the standard error.

Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) are the two information criteria that measure the model performance while considering the model complexity. Both models punish the number of parameters used. The smaller the AIC and BIC values are, the better the model fits the data. Among the four SARIMA models analyzed, SARIMA (0,1,1) x $(1,1,0)_{12}$ has the lowest AIC and BIC values of 3065.40 and 3073.69 respectively, further confirming it as the best-fitting model for this dataset.



```
Coefficients:
          ar1      ar2      ma1      ma2     sar1
      -0.5532  -0.0937  -0.0141  -0.3296  -0.3211
s.e.   0.3358   0.2161   0.3271   0.2883   0.0929

sigma^2 estimated as 1.294e+10:  log likelihood = -1528.98,  aic = 3067.96
'(2, 1, 2)x(1,1,0)12 ---- AIC: 3069.96125780682 ---- BIC: 3086.53430141561'

Coefficients:
          ar1      ma1      ma2     sar1
      -0.5543  -0.0032  -0.4283  -0.3196
s.e.   0.3460   0.3303   0.1873   0.0920

sigma^2 estimated as 1.296e+10:  log likelihood = -1529.08,  aic = 3066.16
'(1, 1, 2)x(1,1,0)12 ---- AIC: 3068.16311415463 ---- BIC: 3081.97398382862'

Coefficients:
          ma1      ma2     sar1
      -0.5725  -0.0695  -0.3299
s.e.   0.0974   0.1077   0.0910

sigma^2 estimated as 1.304e+10:  log likelihood = -1529.49,  aic = 3064.99
'(0, 1, 2)x(1,1,0)12 ---- AIC: 3066.9871805208 ---- BIC: 3078.03587625999'

Coefficients:
          ma1     sar1
      -0.6098  -0.3429
s.e.   0.0822   0.0881

sigma^2 estimated as 1.308e+10:  log likelihood = -1529.7,  aic = 3063.4
'(0, 1, 1)x(1,1,0)12 ---- AIC: 3065.40232951927 ---- BIC: 3073.68885132367'
```

Fig 8. Model coefficients significance, AIC, BIC

Residual Diagnosis

Residuals are the difference between the observed values and the fitted values. By analyzing the residuals, it is possible to determine the effectiveness of the model in fitting the data. A model with a good fit should have residuals that are uncorrelated with zero means. From **Fig.9**, it can be seen that the model results in zero-mean residuals. However, at some lags, the ACF values are slightly higher than the threshold, indicating some correlation still exists. This calls for further investigation. This can be done by verifying if the model's residuals are normally distributed. By plotting the histogram of the residuals and the quantile-quantile plot, it can be seen that the residuals are indeed normally distributed. To support this, a Shapiro-Wilk normality test results in W = 0.98528 with p-value = 0.1746, indicating the confidence that residuals are truly normally distributed. With the models agreeing that the residuals have zero mean, low autocorrelation, and are normally distributed (except for a few lags with slightly significant ACF), it can be concluded that SARIMA (0,1,1) x $(1,1,0)_{12}$ is a suitable model for forecasting the data.

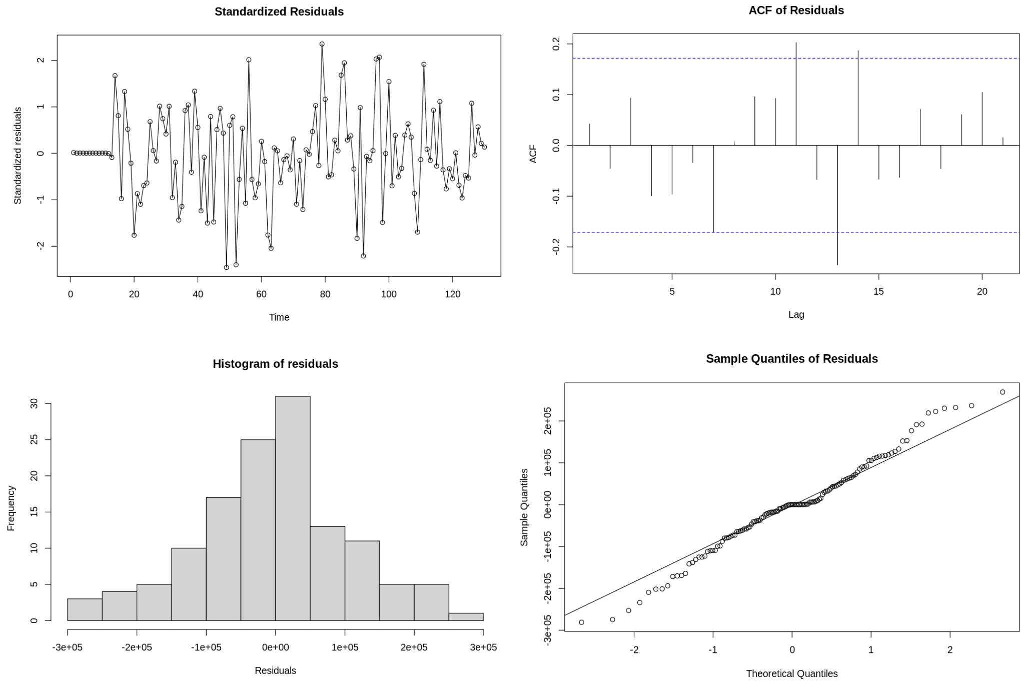


Fig 9. Investigative Diagnostics on the Residuals

Forecast

With the seasonal time series model determined, the plot of the forecast can be done by setting the n.ahead parameter to 24, to estimate the monthly Changi Airport passengers count for the next 2 years to come. The final forecast is shown in **Fig.10**. The forecast shows the upward trend that follows the original data and indicates the forecasted data points with the dark circle. The dotted line indicates the forecast limits that display the uncertainty of the forecast. It can be observed that the uncertainty does increase in the spread the more into the future it tries to forecast.

Nevertheless, the plot does keep up with the key features of the original data well. For instance, the sudden spike in passengers in December of each year followed by the stark decrease in January that follows and the higher highs and higher lows pattern that has been consistently forming over the years.

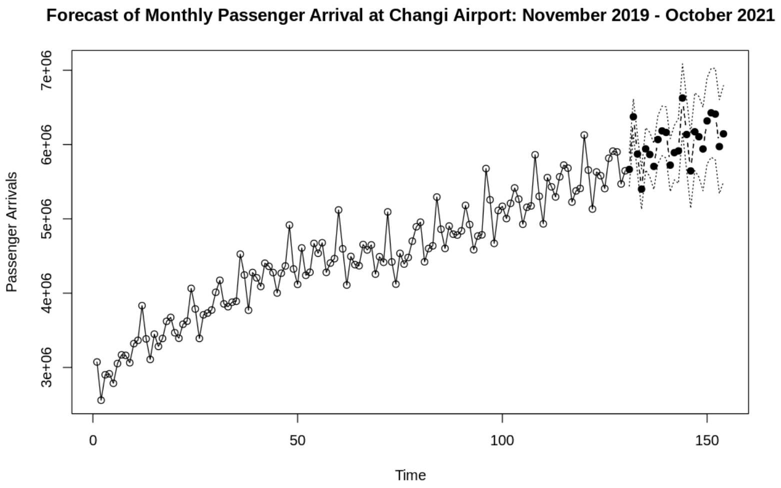


Fig 10. 24 Months Forecast of Changi Airport Passengers Count

Discussion and Conclusion

Analyzing the monthly passenger arrival at Changi Airport, as simple as the premise may seem, holds a lot of merits in itself. Travelers planning to visit Singapore could quickly identify the peak months during the end-of-the-year Christmas long holiday at a glance, or identify the time of the year in which there would be the least crowed. Managers of Changi Airport could make use of the data as one of its various passenger insights to better optimize their services and queue times such as the check-in or the security check process by employing enough workers to satisfy the demand all while avoiding the overhead cost.

Despite the promising outcomes reported, this forecast is still limited in some areas. For instance, to get deeper insights through passenger count, it would be beneficial to have data with shorter periods of the sample, perhaps daily or weekly to help make more fine-tuned decisions. In terms of the SARIMA (0,1,2) x $(1,1,0)_{12}$ model applied for the forecast, the ACF of the residuals still show some slightly significant autocorrelations between lags, albeit small. This result might call for further modifications of the model with perhaps the use of dynamic regression models to alleviate the problem of residual autocorrelation.

At the end of the day, this data might not be fully indicative of the actual passenger count in the real world. As we all know, the widespread effect of the COVID-19 pandemic, an event that took place after the available data points in this dataset, has crippled the travel and aviation industry. After all, the model might not be able to fully account for the effects of holidays, the economy, competitor actions, regulation changes, or other external variables like the COVID-19 global pandemic that wiped out more than half of air passengers count in 2020 and 2021. To ensure that time series analysis of data such as this one can provide a fruitful and insightful result, it is important to always keep data as up-to-date as possible. Just as with all-time series, the longer into the future the model forecasts for, the less accurate it becomes. Thus, it is important to have enough historical data to be used for modeling future values in a reasonable period of time.

References

https://www.kaggle.com/datasets/gohsoonheng/air-passengers-sgp-changi-airport-for-past-10yrs
https://data.gov.sg/dataset/civil-aircraft-arrivals-departures-passengers-and-mail-changi-airport-monthly
https://airssist.com/changi-airport-facts/
https://www.visitsingapore.com/travel-guide-tips/travelling-to-singapore/changi-airport-singapore/
https://www.statista.com/statistics/564717/airline-industry-passenger-traffic-globally/