|  |  |
| --- | --- |
| FORECASTING TRAVEL WITHIN AND OUTSIDE USA  Predict number of passengers (including both domestic and international) on inbound and outbound flight from the USA for coming 7 months. |  |

Contents

[Executive Summary 2](#_Toc58795023)

[Background, Objective and Result Summary 2](#_Toc58795024)

[Introduction 3](#_Toc58795025)

[Main Chapter 4](#_Toc58795026)

[1. Define Goal 4](#_Toc58795027)

[2. Data Collection 5](#_Toc58795028)

[3. Data Preparation 5](#_Toc58795029)

[4. Explore and Visualize 6](#_Toc58795030)

[5. Data Partition 10](#_Toc58795031)

[6 & 7. Apply Forecasting Methods & Evaluate and Compare Performance 10](#_Toc58795032)

[COVID Daily Forecast 10](#_Toc58795033)

[Passenger Enplanements travel Forecast: 18](#_Toc58795034)

[8. Implement Forecasts/Systems 32](#_Toc58795035)

[Conclusion 36](#_Toc58795036)

[Bibliography 37](#_Toc58795037)

[Appendix 38](#_Toc58795038)

# Executive Summary

## Background, Objective and Result Summary

Everyone is locked down in 2020 and the Travel industry is one amongst the many others which has taken a bad hit. U.S. travel and tourism made up approx. 2.8 percent of GDP and was worth approx. $2 trillion till 2019. COVID-19 has caused an unprecedented crisis for the tourism industry. Both domestic and international travels have ceased as national borders are closed and being at home is the new normal. Trends in travel for work and leisure have seen a drastic plunge in the year 2020. We further assume that travel is not likely to return to pre-crisis levels until at least a couple of more years.

Through this project we wanted to explore historic travel data for both domestic and international travel based out of the USA and predict the passenger numbers in flight for the coming 7 months. We tried to answer the following from our study:

1. How to predict numbers for an event that has never happened before? Hence, prediction of COVID cases was our area of interest.
2. How will this rare occurrence of COVID’19 pandemic influence passenger enplanements? Hence, we built several forecast models for predicting travel numbers for the next 7 months.
3. Whether a reduction in COVID cases will bring back the travel to normal or not? A dummy forecast scenario helped us partly answer this question.

We only considered an aggregation of inbound and outbound passengers on flight from the USA. We build several forecast models like Moving Average, ARIMA and Regression with trend to forecast COVID data and then utilize forecasts from the ARIMA model to utilize in one of the models for travel forecast. Also, we utilized several other models to forecast travel like Holt’s Winter model, Moving Average model and several two-level models like Holt’s Winter +Moving Average of residuals, Holt’s Winter + Auto Regression of residuals, Regression with Linear Trend, seasonality, COVID external variable and Moving Average of residuals.

We got two-level model with Regression (Linear trend, Seasonality, COVID external variable)+Moving Average of residuals as the best model with lowest MAPE i.e. 3.6 but because 2020 has huge residual values, it is forecasting negative #of passengers which is incorrect if you look at the use case. Hence, this model's output is logically incorrect, and we could not consider this model. We found Holt’s Winter with AR(1) to be the best model to forecast passengers considering the MAPE and overall model fit on data available so far.

Though we would consider this as the best forecast we would still want to revisit the data sometime early next year to be able to incorporate new COVID numbers into the Regression model and reassess it for travel forecast. Also, travel forecast is influenced by increasing COVID cases which in turn is dependent on a lot of other qualitative factors like closure of national borders, work from home by companies, loss of jobs, vaccine availability in early next year, inhibition of people to travel for leisure etc. Travel is likely to improve if COVID cases reduce possibly owing to the lockdown and vaccine emergence. Hence, if we refresh the model early next year, we will have an improved picture.

# Introduction

Two-thirds of the world’s aircraft fleet had been parked, and 18 airlines have filed for bankruptcy in a matter of months. COVID pandemic has led continents to close its borders leading to major hits to the travel industry. There are other factors like loss of jobs, inhibition to travel for leisure, lack of business travel owing to work from home by employers, etc which have influenced air travel within and outside the USA. McKinsey COVID-19 global air traffic demand scenarios suggest a slow recovery. It mentions that in one of the baseline scenarios, demand could be down 66 percent for 2020 and 47 percent for 2021, compared to 2019. Air travel demand recovery back to 2019 levels will likely not take place until 2024 globally— in line with the latest International Air Transport Association (IATA) forecasts. The McKinsey study made us curious and we wanted to develop our own prediction model for travel numbers.

Major objective for our study was to identify a model that would help forecast an unknown unseen scenario like COVID 19 pandemic which has no historical data and hence no trend and seasonality. We also wanted to understand if we could use this COVID forecast to forecast travel numbers. Further, assess how travel numbers can go back to pre-COVID levels like in year 2019.

We obtained passenger enplanement monthly data (domestic and International aggregated) from Bureau of Transportation Statistics and COVID daily data from a public database i.e. Our World in Data website. This was our starting step towards solving our use case.

# Main Chapter

We will present the solution to our case into below mentioned 8 steps:

## Define Goal

The objective of this project is to explore the historic inbound and outbound travel data based out of the United States of America between January of 2000 to August 2020 and to predict the flight travel numbers for the coming 7 months (till March 2021). The period we have taken into consideration also contains the unprecedented levels of fall in passenger enplanements number owing to the COVID-19 pandemic that started in the early months of 2020 and highly impacted the travel by March 2020.Through this project we wanted to explore and learn about the techniques to incorporate change in data trends caused by special events such as the pandemic.

## Data Collection

The main aim of this project is to forecast the travel numbers in the United States for the period between September 2020 and March 2021. The impact of COVID-19 in the decline of travel numbers in the US this year is significant and cannot be overlooked. In order to accommodate the effects of the pandemic on the travel we have used two datasets in the analysis. Passenger Enplanements to study the travel and COVID 19 testing numbers and positive case numbers to study the pandemic.

**COVID -19 data set:**

The COVID-19 data set we have used is the public data available in the [Our World in Data website](https://ourworldindata.org/history-of-our-world-in-data). The data relies on the data collected from John Hopkins University’s COVID dashboard and is updated regularly. For the purposes of the analysis in this project, we considered the daily data starting 31/12/2019 till 11/18/2020.

**US inbound and outbound travel data set:**

The data set used to predict the travel number in this analysis is from the [Bureau of Transportation Statistics](https://www.transtats.bts.gov/TRAFFIC/). The travel numbers reflect the actual monthly passenger enplanements captured between January 2000 to August 2020 in thousands.

## Data Preparation

The data sets that we considered for our analysis have a difference in their basic granularity. The COVID-19 data set gives us the COVID testing and cases daily. While the passenger enplanements data is of monthly granularity. To give an accurate forecast, the COVID-19 transmission rate needs to be captured daily. Also, the COVID-19 testing played an important role in capturing the number of daily cases. We chose to forecast the COVID cases at daily granularity for the next 120 periods and then aggregate it to monthly granularity to match the travel forecast.

Though capturing COVID time series data at its natural granularity helps us get a better forecast, the overall erratic nature of the numbers contributed to the noise factor to a great extent. The lack of seasonality or a definite trend in the COVID numbers posed a challenge in finding the best model that could be used for forecasting.

## Explore and Visualize

Data exploration is an important step in understanding the data and using the patterns to forecast better. Visualization of the data is a vital tool in this process.

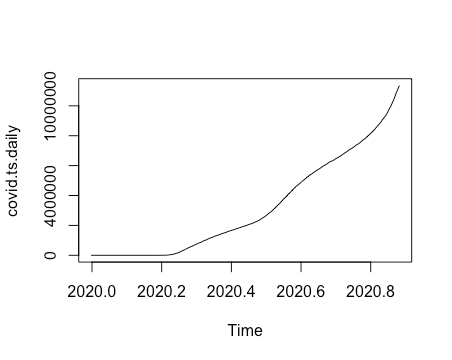


Figure 1: COVID daily time series

From the above visualization of the COVID data set, we can see that there is an exponential increase in the number of daily cases as the days progressed. There is an overall upward trend but there is no visible seasonality component within the data we have. Also, note that as this is an ongoing and recent event, we have only 11 months data and no historic data to compare it with.

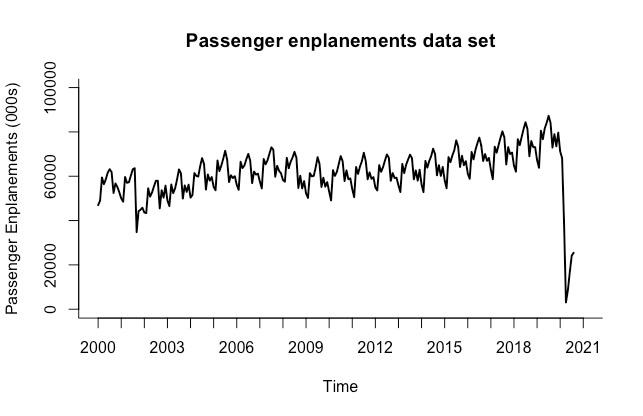


Figure 2: Travel - Passenger enplanements monthly time series

From the above time series plot for Passenger enplanements data set, we can see that there is a clear seasonality component and a fairly constant upward trend until the time lines of the pandemic in early 2020. This dip in the data is heavily influenced by the travel restrictions imposed during the months of March 2020 to June 2020 by various countries throughout the world.

**Predictability of the data:**

A time series data set is said to be predictable when its historical data can be used to make meaningful and accurate predictions into the future. This can be evaluated using the Acf() in R. If the correlogram output of the Acf() has statistically significant autocorrelation coefficients, then we can safely say that the data has patterns that are useful for making predictions.

Predictability of the COVID-19 daily data:

When we plot the correlogram for the COVID-19 data using the Acf() function, we can see that there is a significant presence of trend components in the data set. An upward trend is evident from the way the autocorrelation coefficients are in the first several lags. Also, there are no random variations in the data pattern as each data value is a discrete observation on the given day. Though there is a lack of clear-cut seasonality in the data, we can still safely assume that the data is in fact predictable with a heavy upward trend.

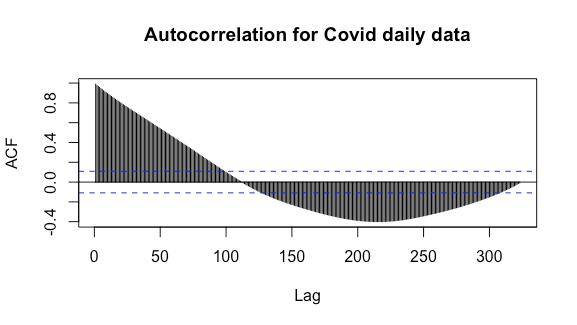


Figure 3: Correlogram for COVID daily data (max lag 365)

Predictability of the Passenger Enplanements - travel data set:

Based on the below correlogram output of the travel data set, we can see that there is a high autocorrelation in lag-1 implying a significant trend component. There are significant non-random autocorrelation coefficients in almost every lag except lag 6. Also, we have statistically significant autocorrelation coefficient for lag 12 which suggests that travel data has month over month seasonality. This is clear evidence that the travel data set is in fact predictable.

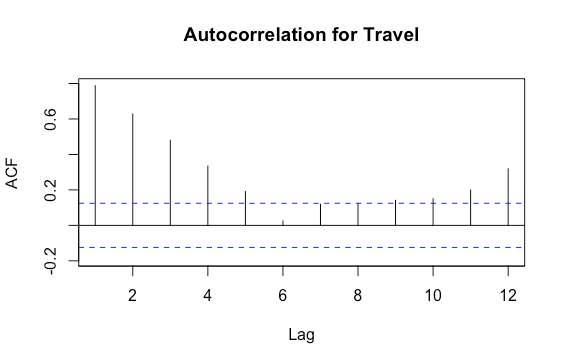


Figure 4: Correlogram for Travel - Passenger enplanements (max lag 12)

**Time series components for the travel data set:**

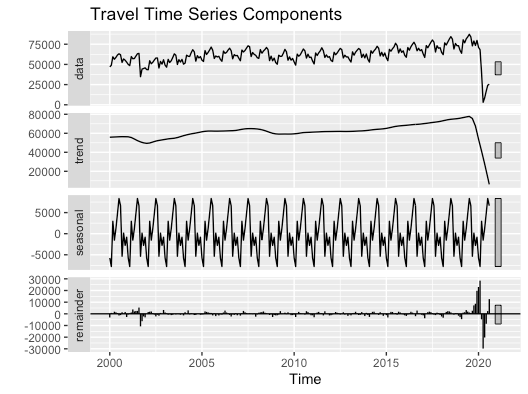
From the below plot of the time series components of the travel data set, we can see the presence of trend and seasonality in the data. Seasonality shows the cyclic behavior year over year as observed in the travel data. Passenger count increases in the first quarter then dips slightly followed by a huge peak in the fourth quarter around holiday season and then dips at the starting of next year. This variation is consistent on an annual basis. Further, the trend is slightly upward and remains constant over the period of observation until the timelines coincide with that of the onset of the pandemic. At this point the trend drops to an all-time low. This can be taken as a response to a special event that happens rarely. This drop also impacts the seasonality pattern of the data set.

Figure 5: Time series components for Travel - Passenger enplanements

## 5. Data Partition

Data partitioning is an important preliminary step before forecasting. Partitioning is done to test how a selected model performs when it comes to forecasting. The general practice is to develop a forecasting model using the training data and validate the model performance by forecasting the data for the validation period. This is then compared to the actuals in the partitioned validation data set to determine the quality of the forecast.

Data partition for COVID-19 data set:

The total periods in the COVID-19 daily data amounts to 324 days. The earlier period of t=1,2,3..,n=204 is designated as training period ( typically 70-80% of the whole time series data)

The later period t=n+1,n+2,...N=120 is designated as the validation period.

Data partition for Travel data set:

The total periods in the passenger enplanements travel data set amounts to 248.The earlier period of t=1,2,3..,n=228 is designated as training period .The later period t=n+1,n+2,...N=20 is designated as the validation period.

## 6 & 7. Apply Forecasting Methods & Evaluate and Compare Performance

Different forecasting methods helped us predict for COVID cases and passengers. Enlisted are the methods we used:

### COVID Daily Forecast

COVID data had an exponential trend and no seasonality as this is an unexpected event happening for the first time. Therefore, it was challenging to predict COVID cases for future time periods. We tried three models which focused majorly on trend and level components of the time series. Summarized below are the details of the models:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| # | Model | Accuracy (MAPE, RMSE or other measure) | Prediction Outcome Comparison | Reason for consideration/non consideration for Forecast | Model Rating for Forecasting  1 (worst) – 5 (best) |
| 1 | Regression with Quadratic Trend and External variable (# of Covid tests performed; Test cases were forecasted using Moving Average) | MAPE Inf  RMSE 216300.80 | We have a continuous increasing trend in the original data and hence the model predicts an increasing upward COVID case in the future periods for all models. | **Not Considered:** MAPE is Inf as we have CoVID cases as zero in the initial periods. If we compare RMSE then it has the maximum errors compared to other 3 models. | 1 |
| 2 | Moving Average (width = 7) | MAPE 12.32  RMSE 140373.20 | **Not Considered:** As Moving Average did not look as comprehensive as ARIMA. Even MA had a higher MAPE, so we did not consider it for COVID forecast. | 3 |
| 3 | Auto ARIMA  ARIMA (2,2,4)  where.  p = AR (2)  d = lag 1 difference done twice (yt – 2yt-1 – yt-2)  q = 4 i.e MA(4) for error lags | MAPE 3.20  RMSE 6330.92 | **Considered:** Flexible in incorporating all time series components and is a complex model hence, it helps in forecasting a new phenomenon like COVID which has no historical data and hence no seasonality. | 5 |

Table-1: Models used in the COVID-19 forecasting

**Consolidated view of all models as listed in the table above:**

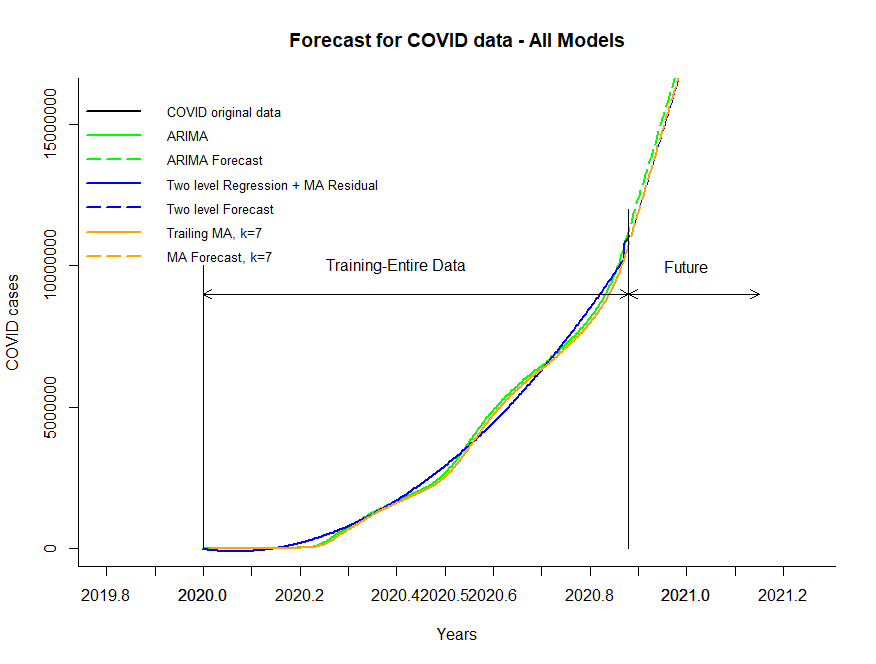
We can observe that almost all models are forecasting in sync. However, ARIMA and Moving Average models are following the original trend perfectly. We considered the ARIMA model for forecasting COVID because it is a more comprehensive model and is useful in predicting new events like a pandemic. Further, the ARIMA model had the minimum error profile.

Figure 6: Consolidated view of all forecast models for COVID

We will be discussing the details of only the top 2 models in our report.

1. Auto ARIMA

On partition dataset

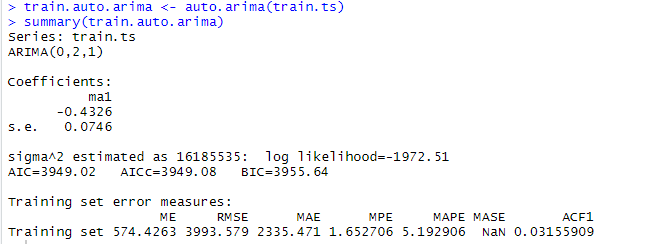


Figure 7: Summary output from R

ARIMA (p, d, q) model is used to forecast data with level and trend components – non-seasonal ARIMA model

ARIMA (0,2,1); where

p = 0, order (0) of autoregressive model AR(p) – number of autocorrelation lags included

d = 2, order (2) of differencing in AR model – indicates how many rounds of lag-1 differencing are performed to remove certain trend

d = 2: difference the series twice, each time of lag-1 (first difference of the first difference), (yt - yt -1) - (yt -1 - yt -2) = yt – 2 yt -1 + yt -2, e.g., y3 – 2 y2+ y1

q = 1, order (1) of moving average MA(q) – number of residuals’ autocorrelation lags included

***Model Equation on partition dataset: (yt - yt -1)- (yt -1 - yt -2) = - 0.4326et-1***

In the plot below, we can observe that ARIMA is not matching the numbers for validation while it is performing well for the training period. We would check the model forecast on entire dataset.

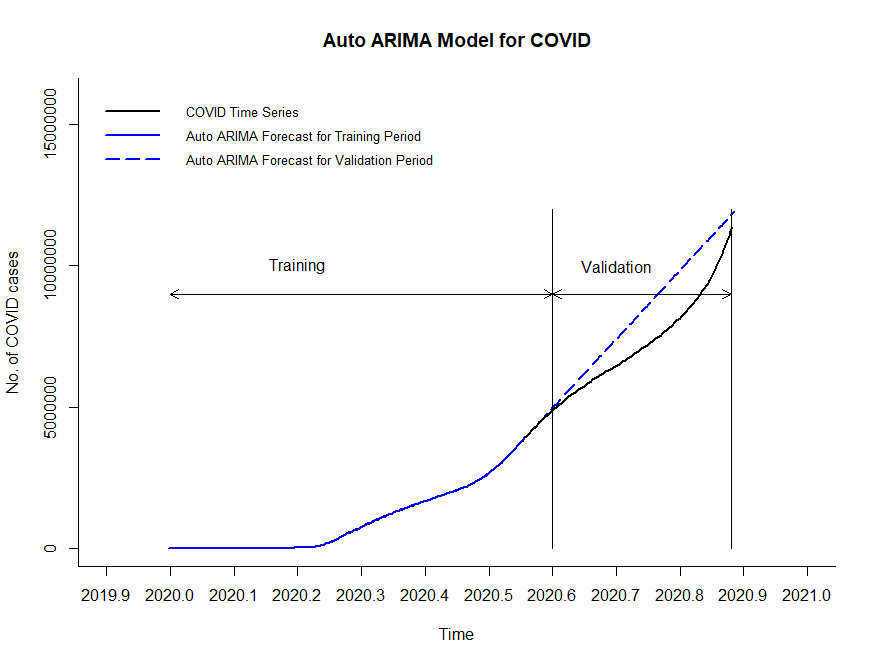


Figure 8: ARIMA output for training and validation period

ARIMA on entire dataset

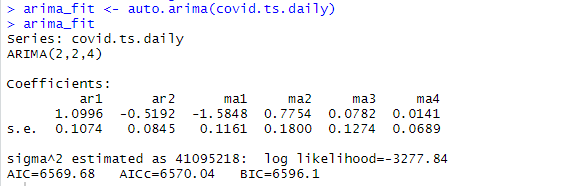


Figure 9: Summary output from R

ARIMA (p, d, q) model is used to forecast data with level and trend components – non-seasonal ARIMA model

ARIMA (2,2,4); where

p = 2, order (2) of autoregressive model AR(p) – number of autocorrelation lags included

d = 2, order (2) of differencing in AR model – indicates how many rounds of lag-1 differencing are performed to remove certain trend

d = 2: difference the series twice, each time of lag-1 (first difference of the first difference), (yt - yt -1) - (yt -1 - yt -2) = yt – 2 yt -1 + yt -2, e.g., y3 – 2 y2+ y1

q = 4, order (4) of moving average MA(q) – number of residuals’ autocorrelation lags included

***Model Equation on entire dataset:***

***(yt - yt -1)- (yt -1 - yt -2) = 1.0996 (yt-1 -yt-2) - 0.5192(yt-2 -yt-3) - 1.5848et-1 + 0.7754et-2 + 0.0782et-3 + 0.0141et-4***

Auto-ARIMA model on entire dataset has very different parameters compared to the ARIMA model on just training dataset i.e. ARIMA (2,2,4) vs. ARIMA (0,2,1) respectively. ARIMA is better able to assess the right parameters post look at the entire dataset. Hence, parameters are different.

In the plot below, we can observe that ARIMA is predicting an upward forecast for future periods.

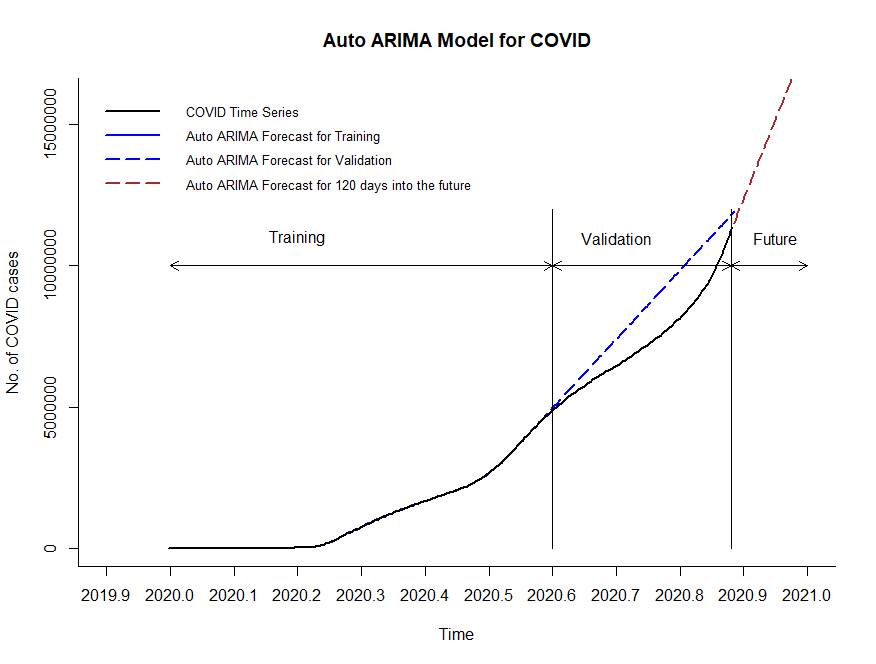


Figure 10: ARIMA output for entire dataset and 7 future periods



Figure 11: Accuracy output ARIMA

Also, model accuracy is high with 3.19 MAPE and 6330.92 RMSE.

1. Moving Averages of original dataset

We considered 7 days, 30 days and 60 days moving average width to forecast COVID cases. We can observe that 7 days moving averages completely synchronize with original data and hence had the minimum MAPE amongst all widths. It shows that COVID cases will keep on increasing exponentially into the future.

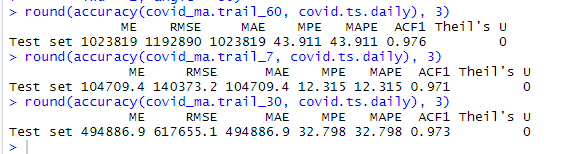


Figure 12: Accuracy output for moving averages with different widths

We can observe the same trend in the plot below. All three moving average models showcase that we will have more and more cases in the future 7 months.

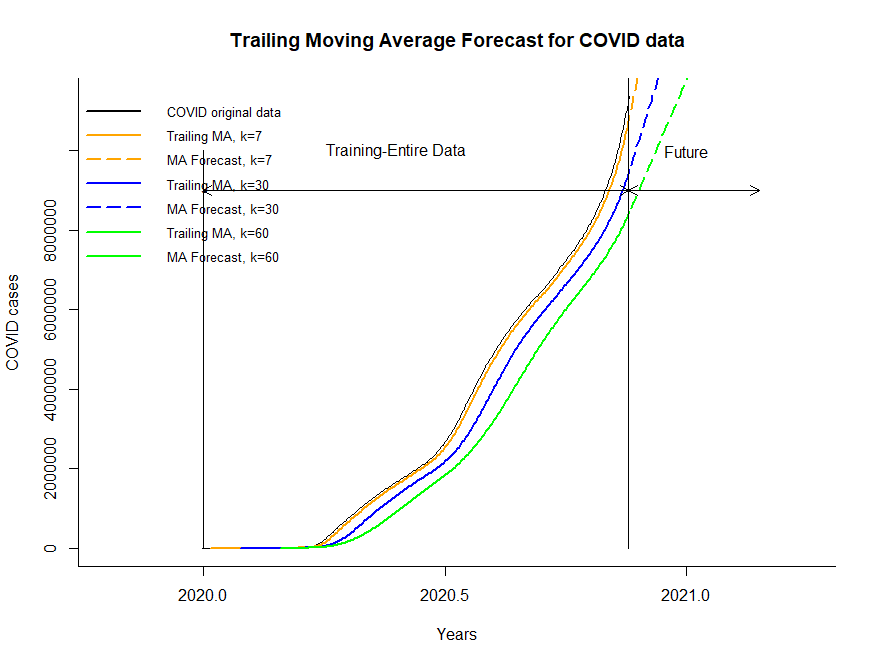


Figure 13: Trailing moving average outputs for entire data and future periods

To finally conclude and reemphasize the best model for COVID forecast, we would like to showcase only the top 2 models in the plot below. We have considered ARIMA forecast as the final forecast and rolled the numbers at monthly level so that we could use them as an external variable in forecasting passenger enplanements.

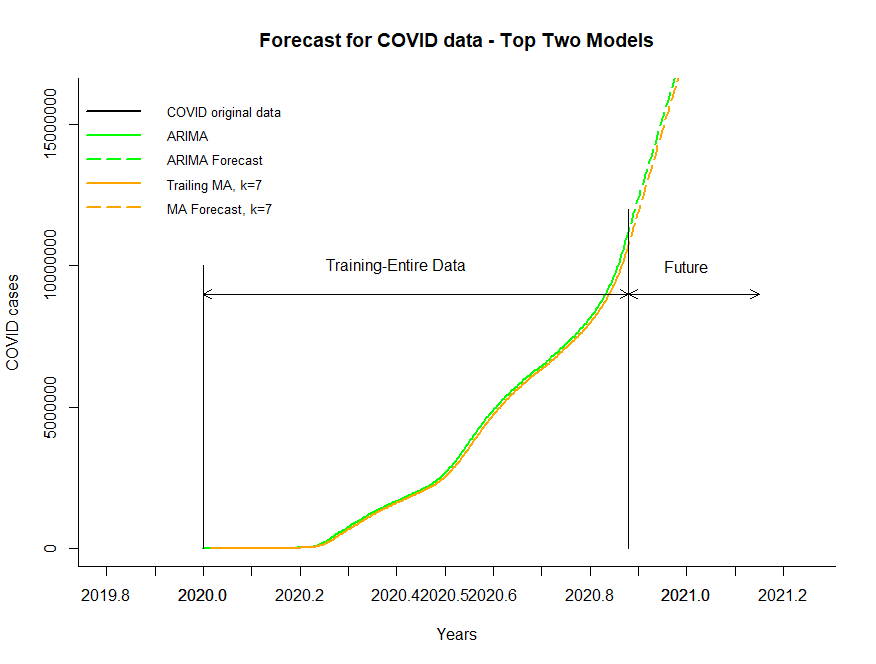


Figure 14: Consolidated view of top 2 forecast models for COVID

### Passenger Enplanements travel Forecast:

The passenger enplanements-based travel data set is relatively streamlined in comparison to the COVID-19 data. We were able to identify clear seasonality and a stable upward trend until the timelines coincided with the onset of the pandemic. We can see a sharp fall in the trend and disruption in seasonality starting early 2020. We needed a model which would capture the effects of the pandemic and forecast travel numbers as accurately as possible. We have used model-based forecasting models, data driven forecasting models, a regression-based model with an external variable (COVID-19 data) and two-level forecasting with moving averages and auto regressive models on residuals to try and identify the best model that would help us forecast better. Our observations are consolidated in the table below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| # | Model | Accuracy (MAPE, RMSE or other measure) | Prediction Outcome Comparison  (talk about forecast) | Reason for consideration/non consideration for Forecast | Model Rating for Forecasting  1 (worst) – 5 (best) |
| 1 | Regression with Linear Trend with Seasonality and External variable (Covid case rolled up at Monthly level) | MAPE 3.6  RMSE 2031.14 | Though the model considers the external factors that drive the special event, the forecasts are significantly impacted by the sharp decline in the trend. | **Not considered:**  The model forecasts are heavily influenced by the residuals. This gives negative forecast numbers for passenger enplanements. Though this makes sense for the model mathematically, this cannot be extrapolated into reality. | 1 |
| 2 | Moving Average (width = 2) | MAPE 6.5  RMSE 3350.99 | This model efficiently captures the drastic changes and is able to give relatively good short term forecasts into the future. | **Considered:**  Moving average(width=2) can be useful if there is a need to get good forecasts for short periods into the future. Parsimony is an important factor for considering this model. | 4 |
| 3 | Holt’s Winter (HW) ZZZ (A,N,A) | MAPE 7.23  RMSE 3838.15 | HW efficiently captures trend and seasonality.  During training and validation, the model was not exposed to the COVID related dip in the data, but it managed to capture the dip efficiently and forecast while using the full data set. | **Not considered:**  Though this model was able to capture the data patterns and provide meaningful forecasts, the MAPE was higher than the HW model with AR(1) for residuals. This shows that there may be some non-random variations in the residuals that are not fully captured. | 3 |
| 4 | Two-level Model: Holt’s Winter ZZZ  (A,N,A) + Auto Regression of residuals (1) | MAPE 6.44  RMSE 3691.22 | HW efficiently captures trend and seasonality.  During training and validation, the model was not exposed to the COVID related dip in the data, but it managed to capture the dip efficiently and forecast while using the full data set. The application of AR(1) on the residuals helped capture the leftover meaningful variations and helped forecast better. | **Considered:**  This model gave us the best forecast in terms of MAPE and forecast numbers. All possible data patterns such as the trend, seasonality, autoregression in the residuals are accommodated into the model’s considerations.  Though there is a need to do a second level forecast to capture the data in the residuals, it compensates in terms of reduced error profile. | 5 |
| 5 | Two-level Model: Holt’s Winter ZZZ  (A,N,A) + Moving Average of Residuals (width=2) | MAPE 7.0  RMSE  4355.08 | HW efficiently captures trend and seasonality.  During training and validation, the model was not exposed to the COVID dip in the data, but it managed to capture the dip efficiently and forecast while using the full data set. Using MA to capture the leftover patterns in the residuals helped adjust the forecast numbers. | **Considered:**  All possible data patterns such as the trend, seasonality and meaningful variations in residuals  accommodated into the model’s considerations.  The forecast numbers and the MAPE were also close to the best performing model. | 3 |
| 6 | Two Level Model: Auto ARIMA with MA on residuals | MAPE 11.85  RMSE 5960.65 | Auto ARIMA model+MA for residuals performed well in training but was unable to capture the essence of the dip in validation like most other models in the mix. However, when the full data set was used, the forecast was fairly consistent and captured the trend of the dip. | **Not considered:**  Though the model can capture the data patterns efficiently, the MAPE is significantly higher than our best performing models. The quality of forecasting may not be as good as the two-level forecasts with the HW models. | 2 |

Table-2: Models used in the Passenger Enplanements Travel forecasting

**Consolidated view of the above models:**

Each of the above models captured the data patterns in their own unique way. Linear regression with external variables had a shortcoming of the final forecast numbers. Holt’s-Winter model as a standalone entity captured the trend and seasonality of the data efficiently. When combined with AR (1) and MA for residual analysis, Holt’s-Winter model performed better than the other models. Auto ARIMA though is versatile and flexible in adapting to the data pattern, the error profile was not conducive.

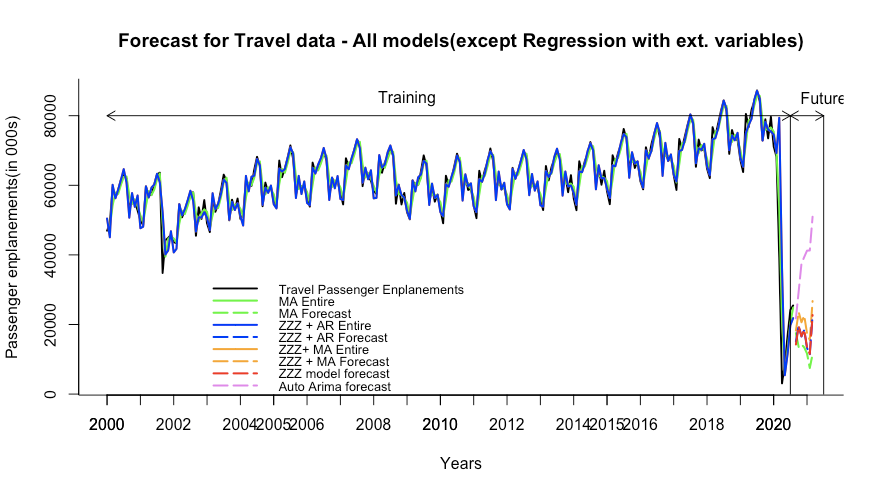


Figure 15: Consolidated output from all forecast models we tried on travel dataset

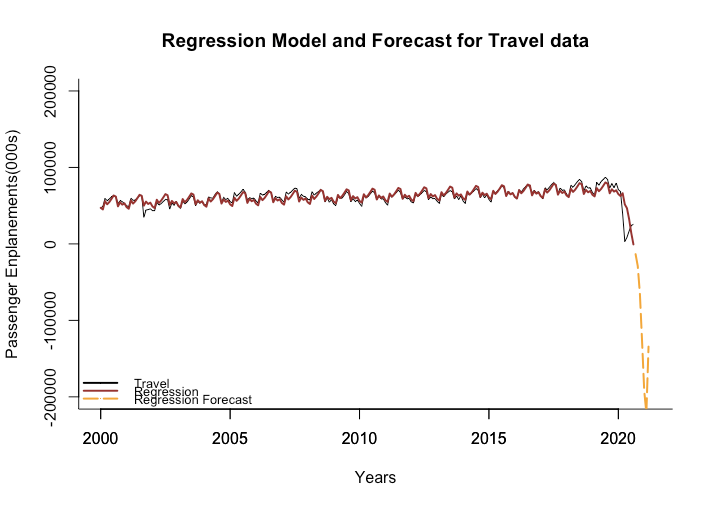


Figure 16: Regression model showcasing a huge negative passenger count forecast for future periods

We would be discussing further about the 3 models that performed well in the above mix.

1. Moving Averages of the travel dataset

We considered 2 months, 6 months and 12 months moving average width to forecast travel passenger enplanements. We can observe that 2 months moving averages completely synchronize with original data and hence had the minimum MAPE amongst all widths. The forecast for this width also captures the dip of the pandemic onset and the travel numbers increasing gradually.

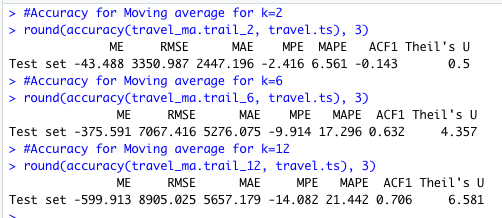


Figure 17: Accuracy output for trailing moving average from R

We can see the same trend in the below forecast plots for the trailing moving averages for the travel passenger enplanements.

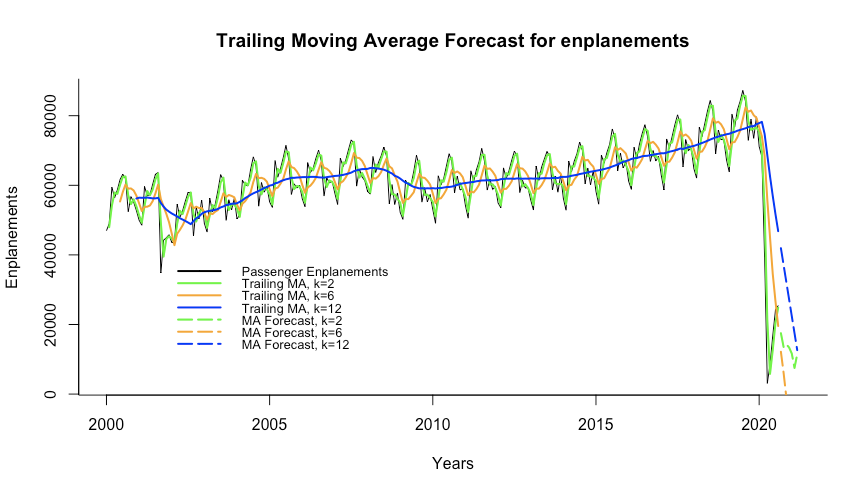


Figure 18: Trailing moving average output for different widths

1. Two-level Model: Holt’s Winter ZZZ (A,N,A) + Auto Regression of residuals (1)

This model has 2 levels of forecasting. The first level is the Holt’s Winter model and the second level is the application of the AR (1) model on the residuals.

Holt’s Winter on training data:

We have developed a Holt’s Winter model using the ‘ZZZ’ parameters in the ets () in R. The function has identified the appropriate parameters as follows.

alpha (smoothing constant for exponential smoothing) = 0.5665

gamma (smoothing constant for seasonality estimate) = 0.0001

seasonality for 12 months is also provided by the model.



Standard deviation sigma = 1740.789

Also, certain measures are provided to measure the entropy and quality of the models like AIC, AICc and BIC.

The plot for the training and validation forecast is as follows:

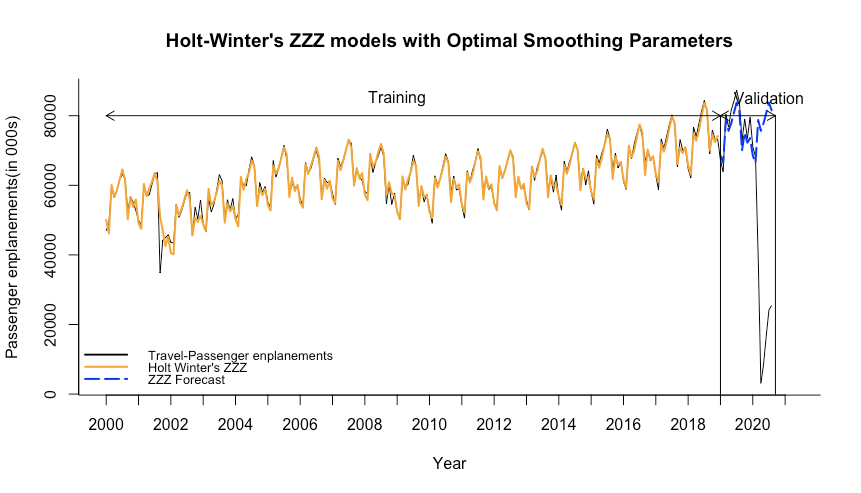


Figure 19: Holt’s Winter ZZZ model with Optimal Smoothing parameters

From the above plot, we can see that the model captures the trend and seasonality of the historical data very well. However, the sharp decline at the onset of the pandemic is unknown during the training period.

Holt’s Winter model using the entire data set:

The function has identified the appropriate parameters as follows.

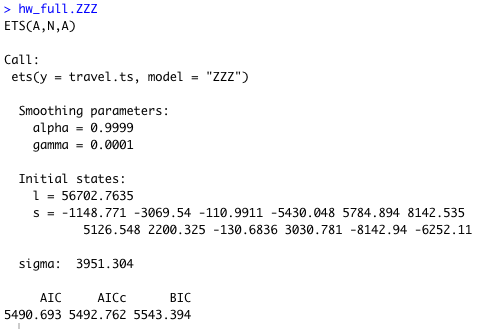


Figure-20: Summary output of HW ZZZ model from R

alpha (smoothing constant for exponential smoothing) = 0.9999

gamma (smoothing constant for seasonality estimate) = 0.0001

seasonality for 12 months is also provided by the model.



Figure-21: Seasonality provided by the HW ZZZ model in R

Standard deviation sigma = 3951.304

The plot for the forecast into the future using the historical data is as follows:

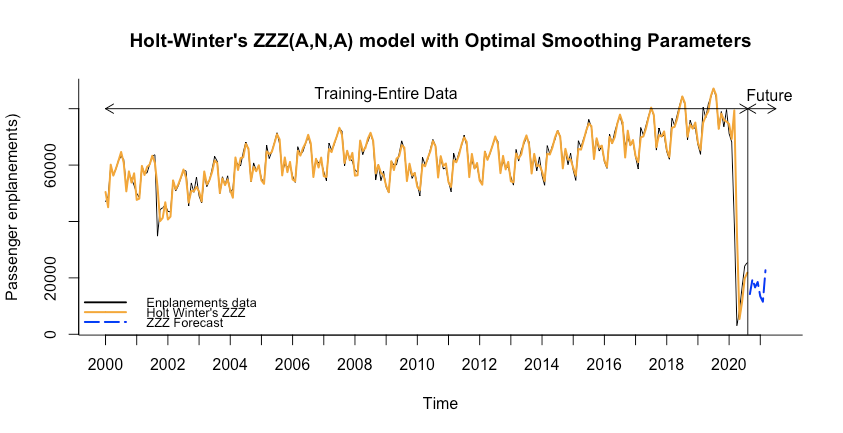


Figure-22: Holt’s Winter ZZZ (A,N,A) model with Optimal Smoothing parameters

From the above plot we can see that, when we forecast using the full data set, the model is able to capture the pandemic’s impacts on the trend and seasonality and forecast accordingly.

The accuracy is as follows:



Figure-23: Accuracy for HW ZZZ model with entire data set

Though the accuracy and forecast for the Holt’s Winter model is good on its own, there is scope for improvement in capturing the data patterns that could be hidden in the residuals. Hence,

application of AR (1) model on the residuals will be helpful.

To get an idea of the residuals and the autocorrelation for the residuals, the following plots would be helpful.

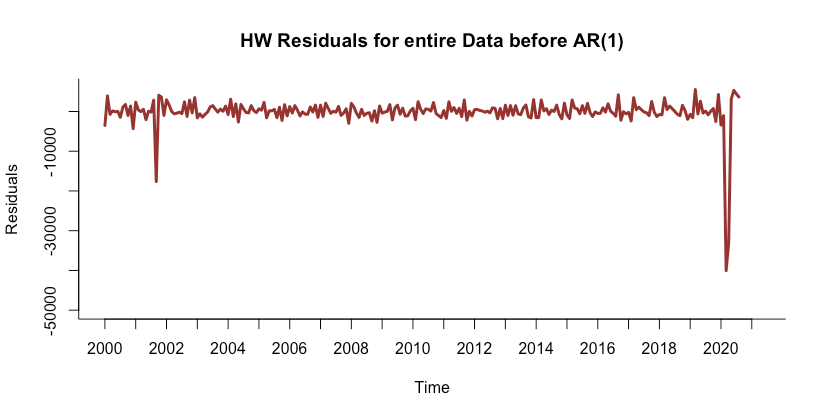


Figure-24: HW ZZZ model residuals for entire data before applying AR(1) on residuals

Though a major portion of the data patterns are captured by Holt's Winter model, there are some variations that we can see in the residuals plot.

The Correlogram shows the presence of non-random autocorrelation coefficients present in the first lag. This implies that the residuals may have some meaningful variations that can help us forecast better.

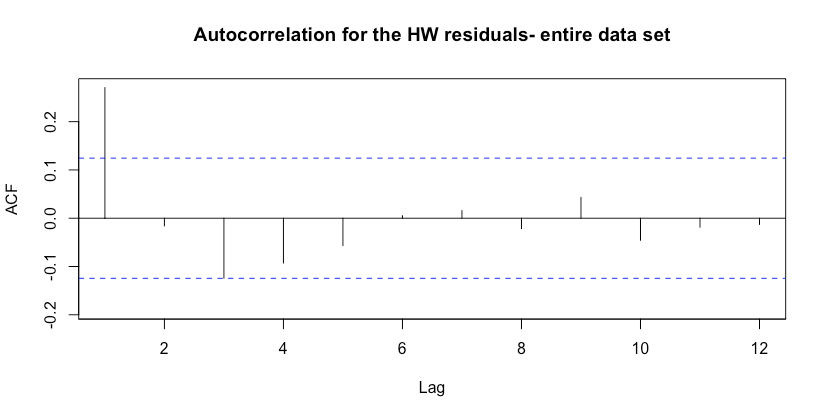


Figure-25: Autocorrelation for the HW ZZZ model residuals for entire data set

After the application of AR (1) model on the residuals of the HW model using the full data set, and forecasting them, we can get a combined forecast by adding the two-level predictions. The following plots help us see the adjustment in forecasting given by the addition of residual corrections.

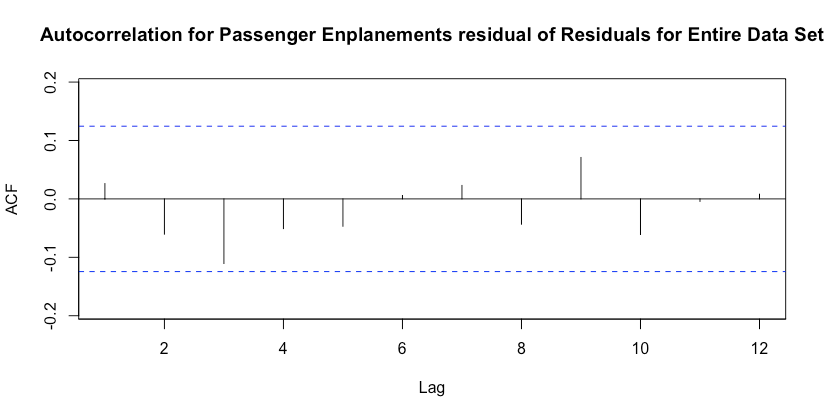


Figure-26: Autocorrelation for Passenger enplanements residuals of residuals for entire data set

After the second level forecast, there are no statistically significant autocorrelation coefficients present. The complete forecast for 7 periods into the using the entire data along with the residual corrections are as follows:

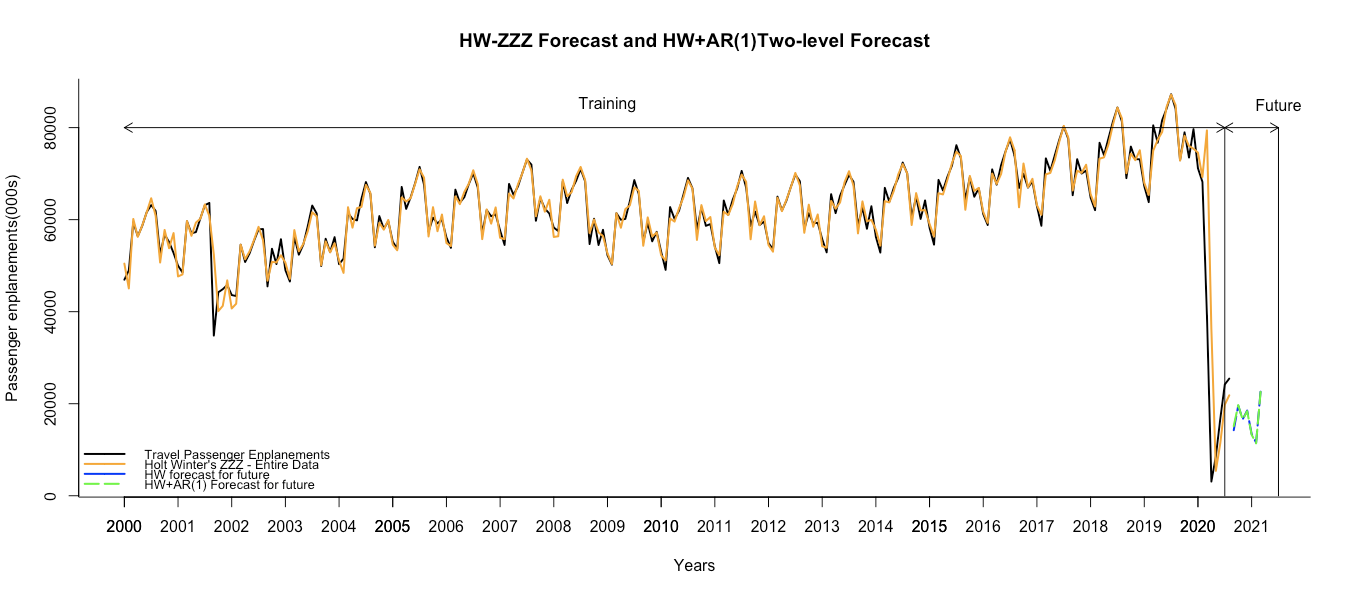


Figure-27: HW ZZZ + AR (1) two-level forecast

The accuracy for the two-level forecast using Holt’s Winter Model and the AR (1) on the residuals is as follows:



Figure-28: Accuracy for HW ZZZ + AR (1) two-level forecast

Application of Moving Average on the Holt’s Winter model residuals:

In a similar fashion as in the previous model, we analyzed the residuals of the Holt’s Winter model using Moving averages. This was an attempt to see if we could improve the forecast accuracy by capturing any leftover data patterns present in the residuals that do not have a strong autocorrelation factor. In our earlier trial on Moving Averages on the travel data, the window width of 2 was very effective in capturing the small variations in the data. Hence that width was chosen to be applied on the residuals.

After the application of the trailing MA of width=2 on the residuals, the residuals can be seen as follows:

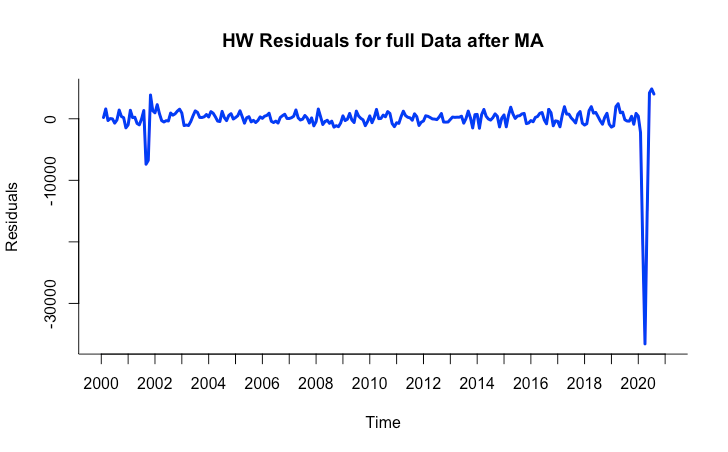


Figure -28: HW ZZZ model residuals after applying MA (width=2)

There is some reduction in the amplitude of the residuals after the application of moving averages. The residual forecast is also done for the 7 periods into the future.

The combined forecast for 7 periods into the future for Holt’s Winter model with MA for residuals using the entire data set can be seen in the following plot:

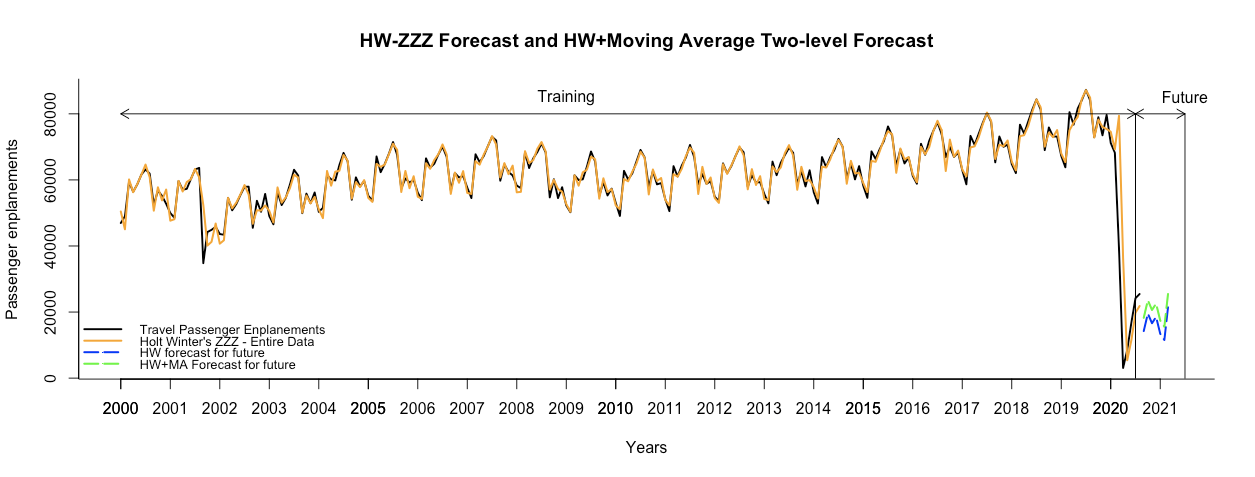


Figure -29: HW ZZZ + Moving average two-level forecast into the future 7 periods

The accuracy for Holt’s Winter model with MA for residuals is as follows:

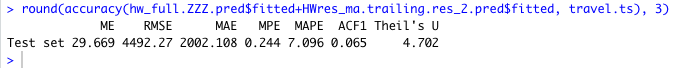


Figure-30: Accuracy for HW ZZZ + AR (1) two-level forecast

Comparing the forecast prowess of the top-3 models side by side:

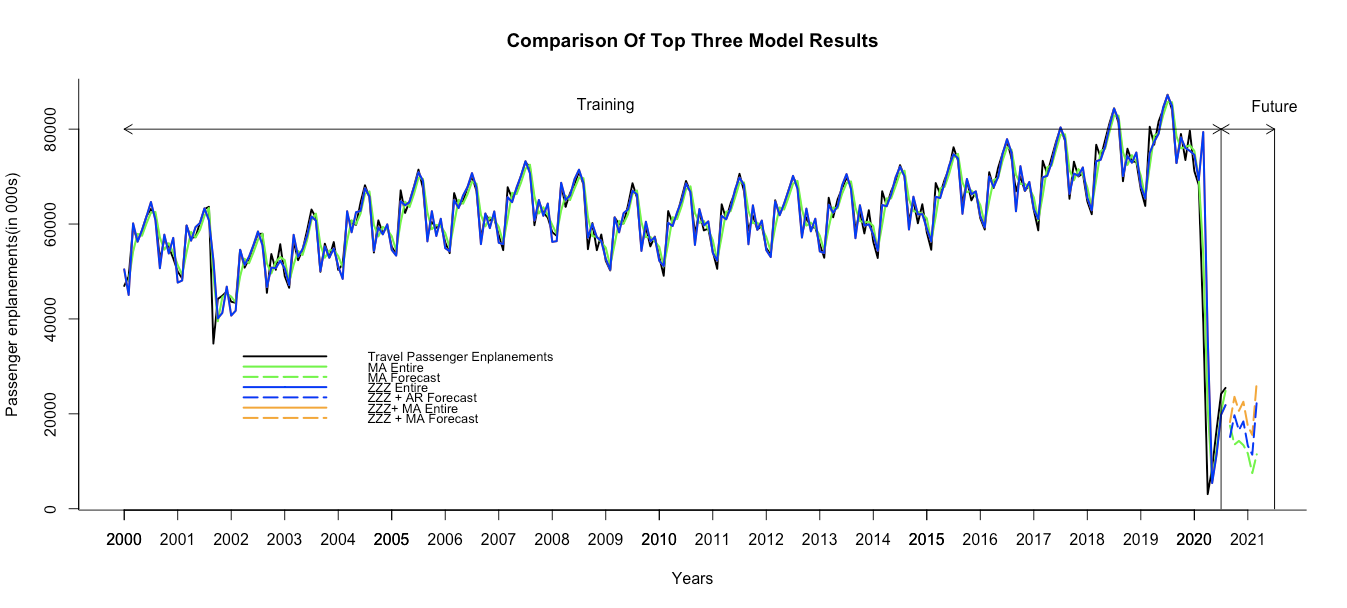


Figure-31: Comparison of the top 3 model results for forecasting travel passenger enplanements

Based on the MAPE, Two-level Model: Holt’s Winter ZZZ (A, N, A) + Auto Regression of residuals (1) emerges as the best model among the mix (refer to the line in blue).

## 8. Implement Forecasts/Systems

We would consider ARIMA as the best model for COVID cases prediction as it is more comprehensive and considers all-time series components while making a forecast. Also, it is known to work well for unknown events like pandemic.

For Travel, we would consider two level combined model i.e. Holt’s Winter (A, N, A) + Auto Regression (order 1) for residuals. This model was in close competition with Moving Average (width = 2) and another combined model i.e., Holt’ Winter (A, N, A) + Moving Average (width=2) for residuals in terms of accuracy performance metrics. However, there was one model which outperformed all others in terms of accuracy with MAPE as 3.6 which was the lowest of all 6 models we tried. In this model we included COVID cases as an external variable while predicting travel using a two-level model i.e., Regression model with Linear trend and Seasonality + MA for residuals. However, we would not consider this as the final model because the total forecast (Regression + Residual) generated has negative # of passengers owing to the huge dip in 2020 passenger enplanements which caused regression to predict large negative forecast values for the future period which gets reflected in the total forecast post adjustment with some positive values from residual forecast (as reflected in the table below and refer to Figure 16). As we understand, Passenger forecast cannot be a negative number hence, we had to discard this model for Travel prediction.

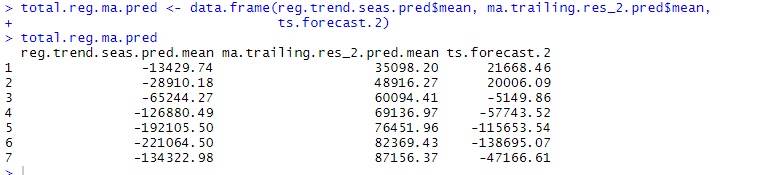


Figure – 32: Regression forecast with ext. variables showing negative forecast for future

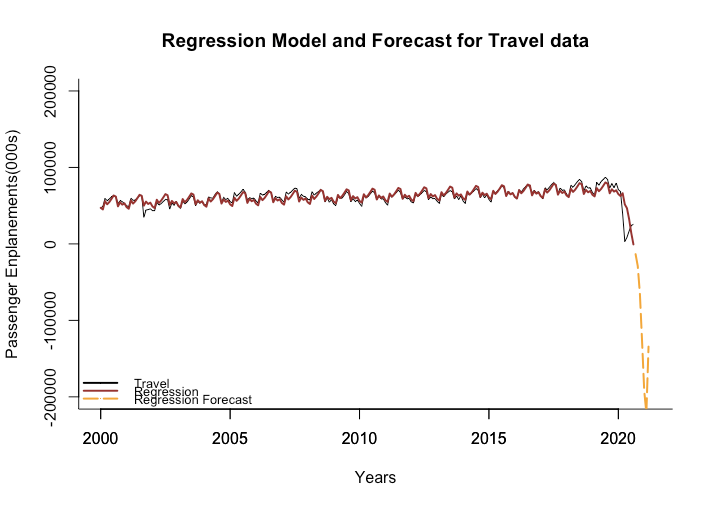


Figure-33: Regression forecast with ext. variables plot showing negative forecast for future

Finally, we considered HW+AR (1) as it has the second lowest MAPE/accuracy profile and it predicted positive passenger count which is a logically correct solution for our use case.

Out of curiosity, we wanted to find out - by when can the travel be back to normal like we had in year 2019. Since we do not have visibility on how COVID numbers would progress, if we predict it with the models at hand it will just show an upward trend. We need to have access to more variables like work from home numbers, better visibility to national border restrictions, job numbers, unemployment numbers, mass vaccine availability etc. These will have some role to play in predicting reduced COVID numbers. So, we foresee that we can update this model in March or April 2020, as we have news on possible release of vaccine in March. This new information while training the model will help us reassess the model, also we will have more data points in COVID dataset which will help models to predict better.

However, with the current set of data at hand, our model will show gradually increasing COVID numbers even if we ask for 4-year prediction. So, with that even the travel numbers will keep on following the trend as shown in green in the plot below.

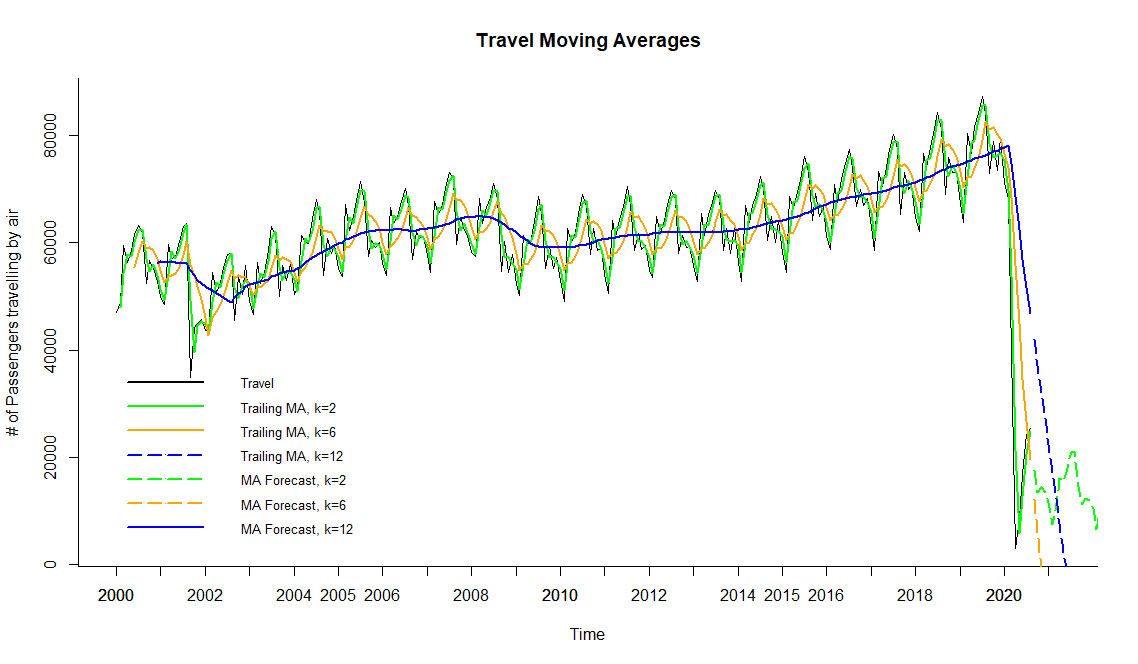


Figure-34: Travel Moving Averages for widths (k=2,6 and 12)

Additionally, in order to answer our question, “Whether a reduction in COVID cases will bring back the travel to normal or not?” - we hypothesized that if COVID cases decrease travel will show improvement. So, to test our hypothesis we created a dummy COVID forecast variable with reduced COVID cases starting January 2021 and you can observe that the travel is trying to be back to normal (refer to the plot below). As a next step to our study, we would like to monitor COVID cases and refresh our model to confirm our hypothesis. Though with the current limitation of restricted data, dummy forecast variables helped us prove our hypothesis right.

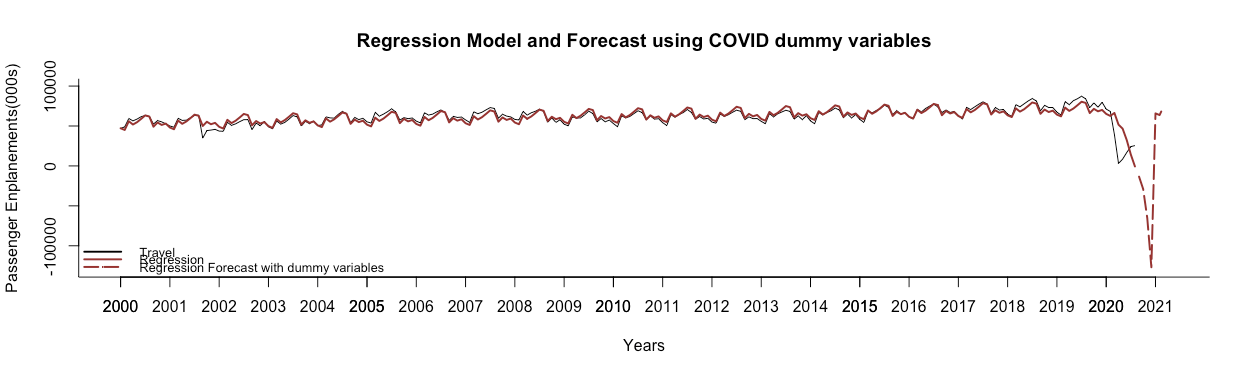


Figure-35: Regression Model forecast using dummy COVID variables

# Conclusion

Time series forecasting on COVID and travel data helped us understand the answers to our questions on the impact of COVID cases on travel, future scenarios for COVID and future of travel in the coming months. Owing to lack of data points for COVID we realized ARIMA would perform the best as it is a statistics-based method and does not require long historic data. However, for travel which had data points for almost 20 years, data driven models scored well as they could learn patterns from data. Also, inclusion of external variables like COVID cases helped predict travel using regression-based analysis. We should try to include multivariate analysis while forecasting as it provides a holistic view. Unfortunately, we could not use that model owing to a huge negative prediction which is logically incorrect when it comes to predicting the passenger count on the flight. Had it been an analysis on growth or share market trends etc. where negative output is acceptable then a regression based two-level model would have been the best model.

In our use case, we considered the next best models which was a combination of data driven and model based approach i.e. Holt’ Winter ZZZ (A, N,A) + AR (1) for residuals as the best model for forecasting travel with the dataset that we had. We could further assess an interesting approach and test how the travel would get back to normal. A possible explanation is that when COVID cases reduce may be because of existing lockdown conditions or post availability of vaccine by mid-next year, we would start to get immunity and hence national borders, offices will start opening up. This would allow people to travel both for work and leisure. It would take time, possibly a few years to get travel back to normal. As McKinsey and International Air Transport Association (IATA) forecast reports say travel would not be back to 2019 levels until 2024 globally.

# Bibliography

1. Travel Passenger Enplanement Data Source: <https://www.transtats.bts.gov/TRAFFIC/>
2. COVID 19 Daily Dataset: <https://github.com/owid/covid-19-data/tree/master/public/data>
3. Transportation Services Index (Reference Read): <https://data.bts.gov/stories/s/9czv-tjte#transportation-modal-data>
4. Travel Forecast: <https://www.ustravel.org/research/travel-forecasts>
5. Reimagining the $9 trillion tourism economy—what will it take?: <https://www.mckinsey.com/industries/travel-logistics-and-transport-infrastructure/our-insights/reimagining-the-9-trillion-tourism-economy-what-will-it-take>
6. Hyndman, R.J., & Athanasopoulos, G. (2018) Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2.

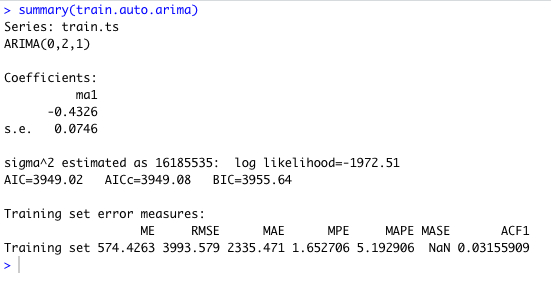
Accessed throughout the period of the course

1. Shmueli, G. and Lichtendahl Jr., K.C. Practical Time Series Forecasting with R, 2nd Edition, Axelrod Schnall Publishers, 2016. ISBN-13: 978-0-9978479-1-8.

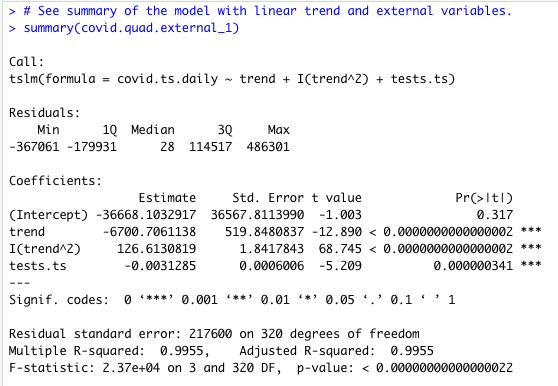
# Appendix

Summary for the models used in forecasting COVID and travel:

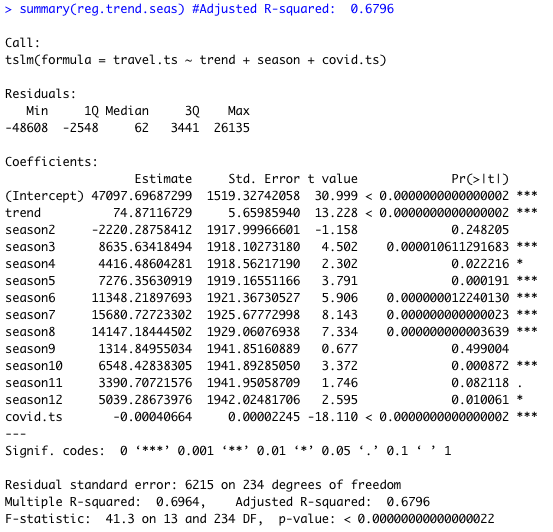
1. Auto Arima for COVID numbers forecasting:



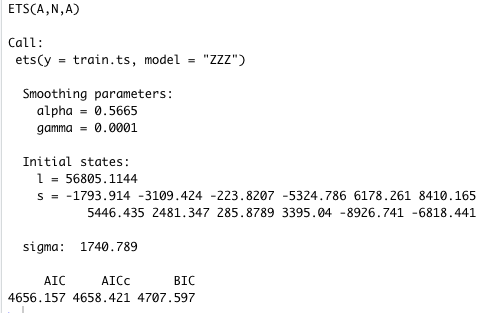
2. Quadratic trend with external variables i.e. number of tests for COVID



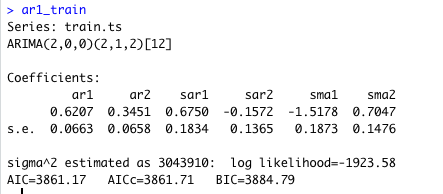
3. Forecasting travel with Regression (Linear Trend+ Season+COVID)



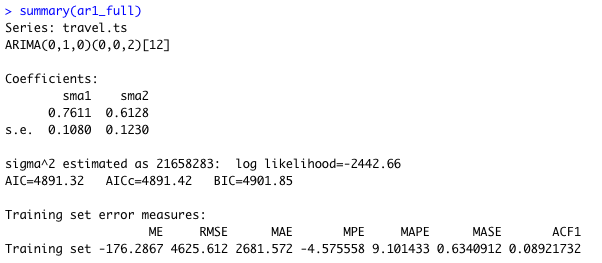
4. Holt- winter forecasting for travel - training data set model summary:



5. Auto-Arima training forecast summary for forecasting travel:



6. Auto-Arima full data set summary for forecasting travel:



7. AR(1) summary for Holt’s Winter’s residuals for forecasting travel:

