Project Proposal - Full Draft

Nick Coronato

Version of 2021-03-24 | Due 2021-03-24

SYS 5581: Time Series and Forecasting

Abstract

The concept of this this project is to analyze historical data from the past five years (2015-2020) in order to understand passenger trends in commercial aviation. Data is open source and provided by the US Bureau of Transportation Statistics (BTS). Revenue Passenger Miles, a key indicator of an airline's operational load, can be modeled as a time series. In this research, I aim to provide insight on the forecast for recovery amongst three leading US airline carriers in the post-COVID-19 environment.

Introduction

The COVID-19 pandemic brought leisure and business air travel to a near-screeching halt in the second quarter of calendar year 2020. Fear of the coronavirus's transmissibility and lethality kept many Americans from boarding a plane in non-emergency situations. Additionally, domestic and international policies basically enforced an immediate ban on unnecessary travel. As one would expect, this has severely impacted the entire travel industry. Experts propose that a full recovery could take upwards of 2.5 years. Smaller airlines may not have that much time to stay afloat; several regional airlines in the US and abroad have already folded. The US airline giants have felt a certain effect from flight cancellations, policy changes, and incurred costs for healthy safety measures.

This project aims to utilize time series analysis methods to answer a simple question about commercial aviation:

How should we expect the largest US passenger airlines to operate in the aftermath of COVID-19?

This question is one of many that will be asked in the coming months, as the demand for personal transportation is expected to eventually return to pre-pandemic levels or higher. The stakeholders in this industry are numerous: not just business travelers and vacationers or airline companies themselves, but practically anybody with a role or vested interest in moving people or things around the globe. The pace and magnitude of recovery for the air travel industry will have third order implications on fuel prices, employment rates, business logistics, the global economy, and so on. This particular question is important because it seeks to provide predictive insight to how the US airline giants will be expected to operate if the travel demand resumes in 2021 or 2022. If they can predict this, the airlines should be able to strategically ramp up and meet the demand. If they do not forecast appropriately, they may fail, this will affect nearly everybody in some way.

I will attempt to answer this question by generating an accurate forecasting model for three large US airlines. The major assumption, designed to reduce problem complexity, is that our largest airlines will be the most likely to survive the pandemic period, and that they are representative of the domestic air travel market at large. The airlines under study - Delta, American, and United - undoubtedly have some of the greatest resources at their disposal, and therefore were decently postured to take on the sudden drop in demand for passenger travel. Another assumption is that future demand for air travel will be roughly equal to the levels seen in the data prior to Quarter 1, Calendar Year 2020.

My model will use that historical data to identify Revenue Passenger Mile trends right up to the COVID-19 onset, and then forecast the expected output (RPM) of Delta, American, and United if normal travel resumes.

$$RevenuePassengerMiles(RPM) = NumberofPayingPassengers * DistanceTraveled$$

The Data

Data for this research is downloaded from the BTS website: https://www.transtats.bts.gov/Fields.asp?gnoyr_VQ=FIM or https://www.transtats.bts.gov/DL_SelectFields.asp?gnoyr_VQ=FIM.

This website allows the user to select key metrics of airline performance by market (domestic segment or international) and by year, along with several other filters. I extracted the raw data (2015-2020NOV) for the following features:

1. Departures Scheduled [integer]
2. Departures Performed [integer]
3. Payload (pounds gross weight) [integer]
4. Seats [integer]
5. Passengers [integer]
6. Distance (miles) [integer]
7. Air Time (minutes) [integer]
8. Unique Carrier (Airline identifier) [factor]
9. Origin Airport ID (five digit code) [factor]
10. Origin City Name [factor]
11. Origin State Name [factor]
12. Origin Country [factor]
13. Destination Airport ID (five digit code) [factor]
14. Destination City Name [factor]
15. Destination State Name [factor]
16. Aircraft Type [factor]
17. Year
18. Quarter
19. Month

The data I downloaded is structured as 1,031,938 rows of a .csv file, with headers that indicate each of the columns (features) identified above. Each row entry, or observation, is a flight segment. A domestic flight segment is defined as any route that terminated in the United States or its territories; it could have originated anywhere.

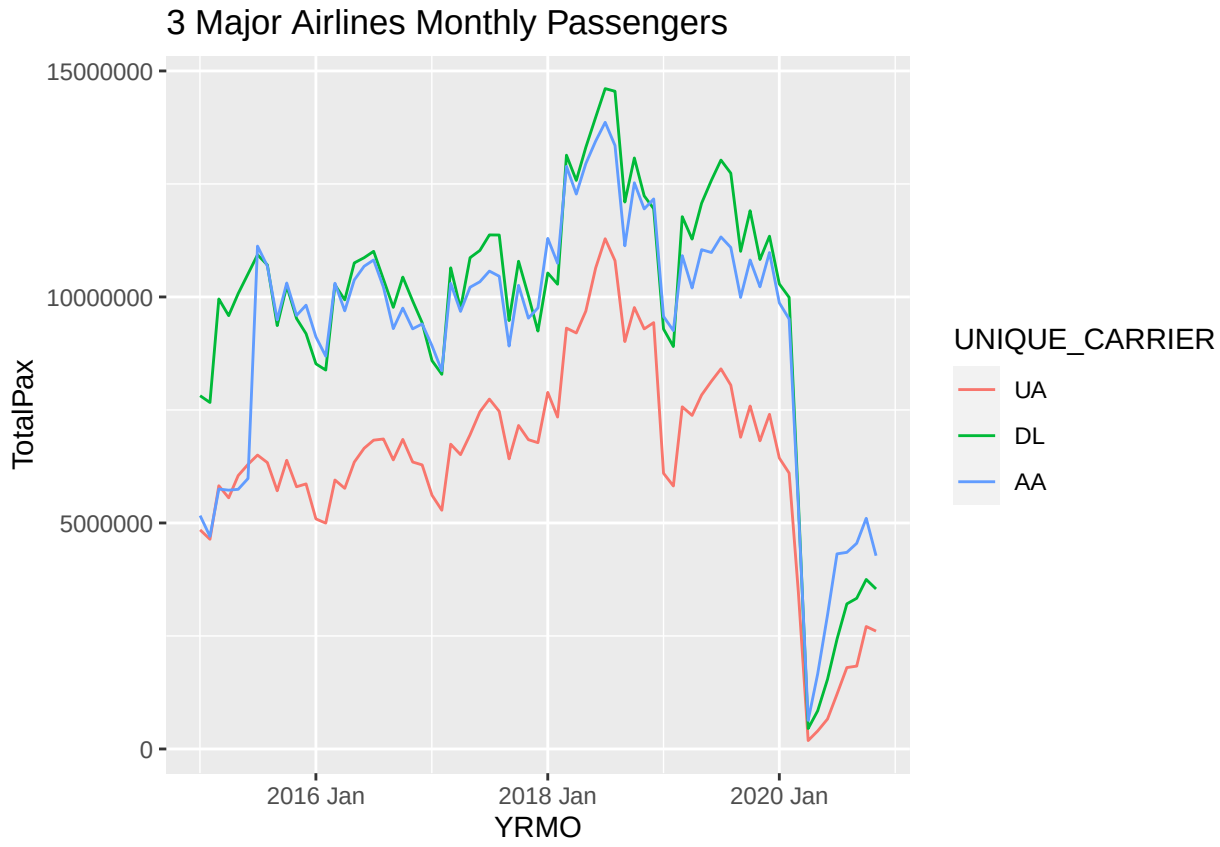
Most route segments were executed multiple times throughout the time period, as indicated by the "Departures Performed" column. If multiple departures of the same segment were performed, the data is already captured in such a way that the other variables were updated; in other words, if two departures of a 100-passenger segment were conducted, then the corresponding "PASSENGERS" column indicates "200" for that segment. I did not have to multiply $(2 * 100)$ to manually calculate for repeat segments.

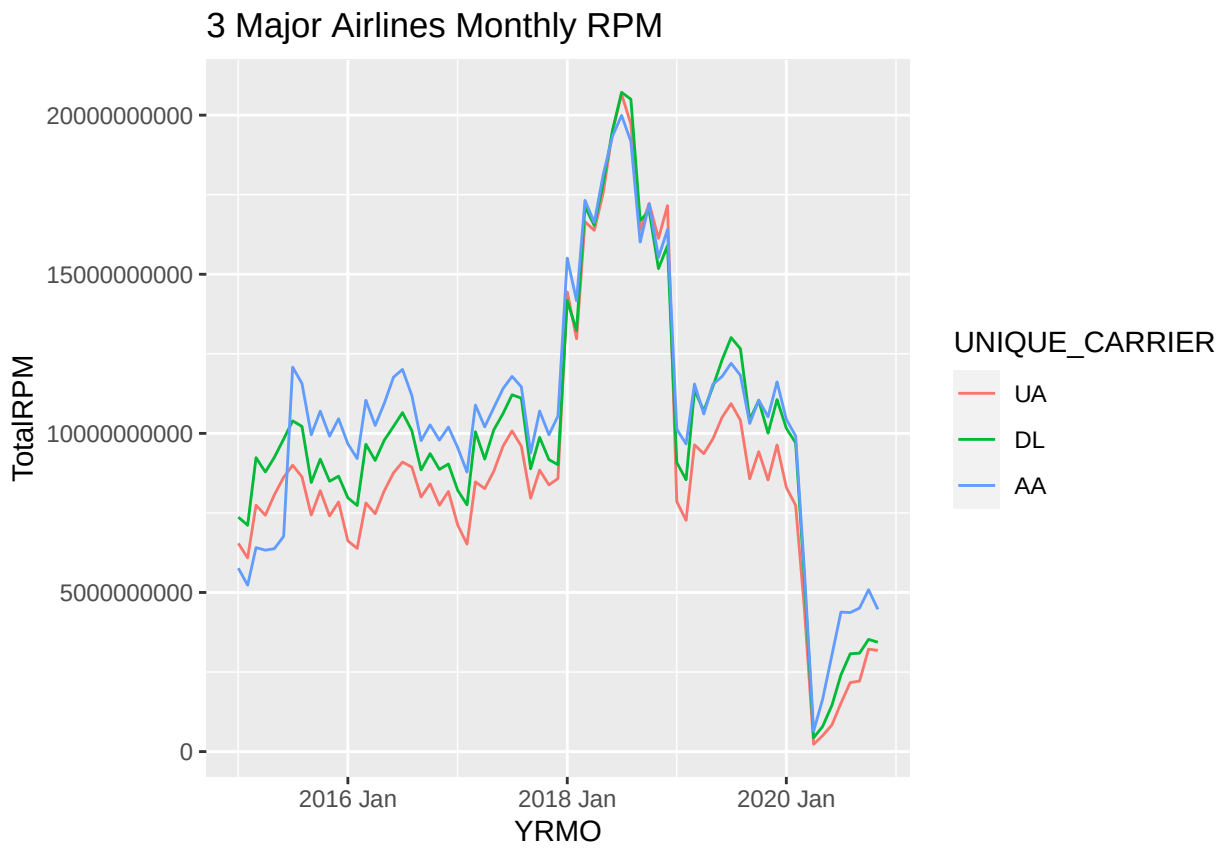
I processed the raw data to fit the requirements of a tidy time series set. This required an aggregation of metrics at a *monthly* frequency. For example, all of the United Airlines segments during January 2015 were summed to form new features: Total Passengers, Total Air Time, etc. The time period of interest was January 2015 to November 2020. This produced 71 monthly observations for each of the "Big 3" airlines, which is a total of 213 data points for each metric. A monthly interpretation of historical travel data seemed most relevant and stable, as it appropriately highlights the seasonal trends of the year without being too granular.

The table below is a sample of the data after initial processing.

```
## # A tsibble: 6 x 9 [1M]
## # Key:       UNIQUE_CARRIER [1]
## # Groups:    YEAR, MONTH @ YRMO [6]
##       YRMO  YEAR MONTH UNIQUE_CARRIER TotalPax TotalRPM TotalDistance TotalSeats
##      <mth> <int> <int> <fct>             <int>    <dbl>         <int>      <int>
## 1 2015 Jan  2015     1 UA              4847338   6.54e9       2446772    5956785
## 2 2015 Feb  2015     2 UA              4639362   6.08e9       2316983    5590926
## 3 2015 Mar  2015     3 UA              5823398   7.74e9       2474956    6831596
## 4 2015 Apr  2015     4 UA              5556445   7.43e9       2549360    6531908
## 5 2015 May  2015     5 UA              6048718   8.07e9       2585598    7051073
## 6 2015 Jun  2015     6 UA              6296079   8.62e9       2655756    7228780
## # ... with 1 more variable: TotalAIR_TIME <int>
```

Two time series plots below show the data for (1) Passengers transported and (2) Revenue Passenger Miles since January 2015. The dramatic drop in travel due to COVID-19 is quite obvious.





Formal Model of the Data Generating Process PRELIMINARY

As mentioned previously, the key to forecasting post-COVID travel numbers will be to consider only the historical data leading up to the 2020 rapid decrease. It is important to establish a "cut-off" point from which we will assume the travel industry experienced no impact from COVID-19. For this study, I consider that time to be December 31, 2019. The plots above appear to indicate normal seasonality in Passengers and RPM up until January 2020 and beyond.

I consider the best representation of the data-generating process for Revenue Passenger Miles to be a Seasonal Autoregressive Integrated Moving Average model (Seasonal ARIMA). According to "Forecasting Principles and Practice (3rd ed), ARIMA models are"capable of modelling a wide range of seasonal data," which is what we can see at a glance in the airline data. This data is not stationary, has moderately strong seasonality and possibly a trend, therefore a seasonal ARIMA may fit better than a basic ARIMA.

The seasonal ARIMA model is formed by including additional seasonal terms with the other non-seasonal terms of a standard ARIMA model. Where ARIMA requires (p, d, q), our airline forecast model must also include a seasonal component $(P, D, Q)_m$. Uppercase notation is used for the seasonal parts, which are multiplied by the non-seasonal terms.

Formal model will be revised after optimization

For example, an ARIMA$(1,1,1)(1,1,1)_4$ model (without a constant) is for quarterly data (m=4), and can be written as

$$(1-\phi_1 B)\ (1-\Phi_1 B^4)(1-B)(1-B^4)y_t = (1+\theta_1 B)\ (1+\Theta_1 B^4)\varepsilon_t$$

Discussion of the Statistical Model *Describe how the formal statistical model captures and aligns with the narrative of the data-generating process.*

Plan for Data Analysis

I plan to utilize this model to create a forecasting tool. While it may be possible to describe a *rate* of recovery for the airlines, I aim to simply forecast the future operational capacity of our "Big 3" airlines given the assumption of a return to normal demand. This forecast would accurately capture the recent data of the past 5 years to provide a baseline goal towards which Delta, United, and American Airlines can strive.

Through the analysis process, I may find out that there is a high variance, which may help the airlines assess their tolerance for risk in the coming months. I may see that there is a high probability of *not* meeting the goal (e.g. underperforming monthly or annually).

I expect to see some differences *between* the three major airlines; initial data exploration revealed that each airline may have a unique operational strategy that determines the length of their flight routes and number of times they fly each route. (The three airlines are not identically structured to begin with.) Full analysis may shed further light on these strategic choices. Again, the assumption is that the pandemic itself did not cause any of the Big 3 to change their corporate strategy in a dramatic way.

This information will be gained by leveraging R for computation, and I will specifically start with the *fable* package and some sub-functions including ARIMA().

I intend to decompose the historical data to see elements of interest, such as the ACF, for assessing the dynamic information in my time series. I will view the difference plots to help refine the model(s) and then check residuals. If necessary, I can use a Ljung-Box test to confirm that the residuals are similar to white noise. If there are multiple models, I will compare AICc or RMSE values to identify the best fit. Finally, I can evaluate the forecast with a test set. Ideally, I can then produce a plot that lays the forecast model onto the next 12 months to see what airlines could expect if travel demand picked up again.

References Forecasting: Principles and Practice, 3rd Edition, Rob J Hyndman and George Athanasopoulos, Monash University, Australia