A Time Series Analysis of Air Passenger Traffic

ISyE 6402 Time Series Analysis
Eugene Kang, Wonhee Lee, MacKenzie Ward
Team 3

Distribution of work:

- Eugene: ARIMA
- Wonhee: Unrestricted and restricted VAR full model
- MacKenzie: Unrestricted VAR subset model
- All equally contributed to final paper write up

TABLE OF CONTENTS

Summary

Our goal is to identify trends and seasonalities that exist in air passenger traffic data between 2002-2017. We explore the time series data from the Bureau of Transportation Statistics on a standalone basis using ARIMA and employ various implementations of the multivariate VAR model incorporating unemployment, interest rate, fuel price, and the ARCA airline index data in our analysis.

Ultimately, we found that ARIMA can be used to model international flights. ARIMA and an unrestricted VAR trained on a subset of our original factors model the data best. We will discuss these findings more in section 5.

Reason for study

With more open borders and advanced technology, as well as increased demand, air travel has dramatically increased over the past couple of decades. The purpose of this analysis is to examine such growth and identify a) trends and seasonalities that may exist in air travel and b) factors that help explain the phenomenon.

Expectation from study

On the univariate level, we expect to see an increasing trend of air travel accompanied by strong seasonalities; air travel tends to spike during vacations when children are on break from school and families take advantage of the timeframe to go on family trips.

Another general consensus on air traffic is that it is strongly correlated to disposable income, or the economy. Simply put, the more money people have to spend, the stronger the air traffic volume. Thus, we expect economic indicators such as unemployment rate and interest rate to have strong impacts on explaining air traffic.

Finally, we expect air traffic to lag changes in fuel prices, with the rationale being that the adjustment in airfare will most likely lag fuel prices by a certain period of time, in turn resulting in the decrease in demand. However, as fuel hedging strategies are different from airline to airline, the degree and quickness to which the change is observed may diverge.

Raw data

- The Bureau of Transportation Statistics records monthly air traffic data in the U.S. It is divided into domestic and international components; we will be using both for our analysis.
- We will be using the Bureau of Labor Statistics' unemployment data, also recorded monthly, as a measure of the health of the US economy.
- Based on various economic indicators, the Federal Reserve sets the base interest rate, which acts as a baseline for other financial institutions when determining rates. Generally, the Fed increases the base rate when the economy is booming(in an effort to curb inflation), and decreases the rate when the economy is struggling(to bolster spending, circulation of money). We will incorporate the 1-month treasury rate set by the Federal Reserve in our VAR model.

- Fuel prices are by far the biggest source of cost for airlines and often determine their performance. As we expect fuel prices to have significant explanatory power in estimating air travel, we will add per barrel U.S crude oil prices to our model also.

Statistical Analysis

The focus of this section is to see how we can model air travel – on both international and domestic flights. We attempt four modelling techniques to determine which can most accurately capture the data. The best model will be determined based on various error metrics of forecasted values.

1. Exploratory Data Analysis (Trend & Seasonality)

First, we plot the original time series to get a general understanding of the data. Figure 1.a shows the time plots of our six data sources. At first glance, it is obvious that all of them have trends. Domestic and international travel show upward trends. International travel has significant increasing trends in particular over the past 15 years. Other four data exhibit nonlinear trends which seem to be critically affected by 2008 financial crisis.

Furthermore, we observe clear seasonal patterns in domestic and international travel. However, this is not so clear in the other variables.

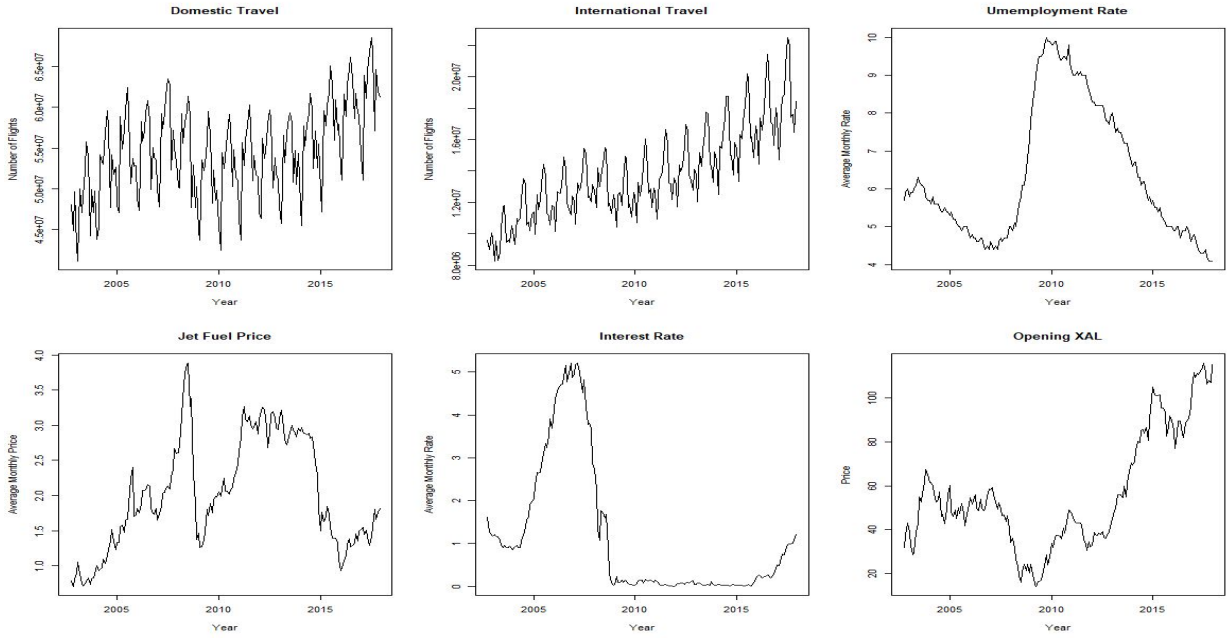
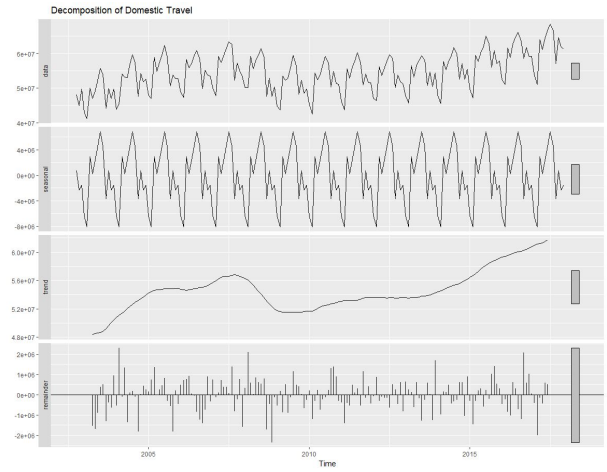


Figure 1.a: Time Plot of Original Data

Next, we decomposed the data into three parts: Y[t]=T[t]+S[t]+e[t], where Y[t] is the original data point at time t, T[t] is the trend component at time t, S[t] is the seasonal component at time t and e[t] is the random error component at time t. The seasonal patterns of both domestic and international travel show that there are more passengers during the summer season(June, July and August). Then, we observed ACF plots of residuals to check for

stationarity. Not only do all six ACF plots fail to decay rapidly, but they also illustrate patterns. Thus, we conclude that transformations must be performed in an attempt to achieve stationary for further analysis.



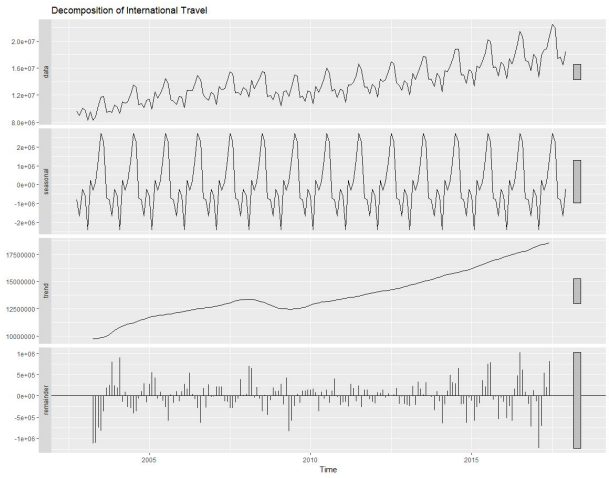


Figure 1.b: Decomposition of Domestic Travel *Figure 1.c:Decomposition of International Travel*

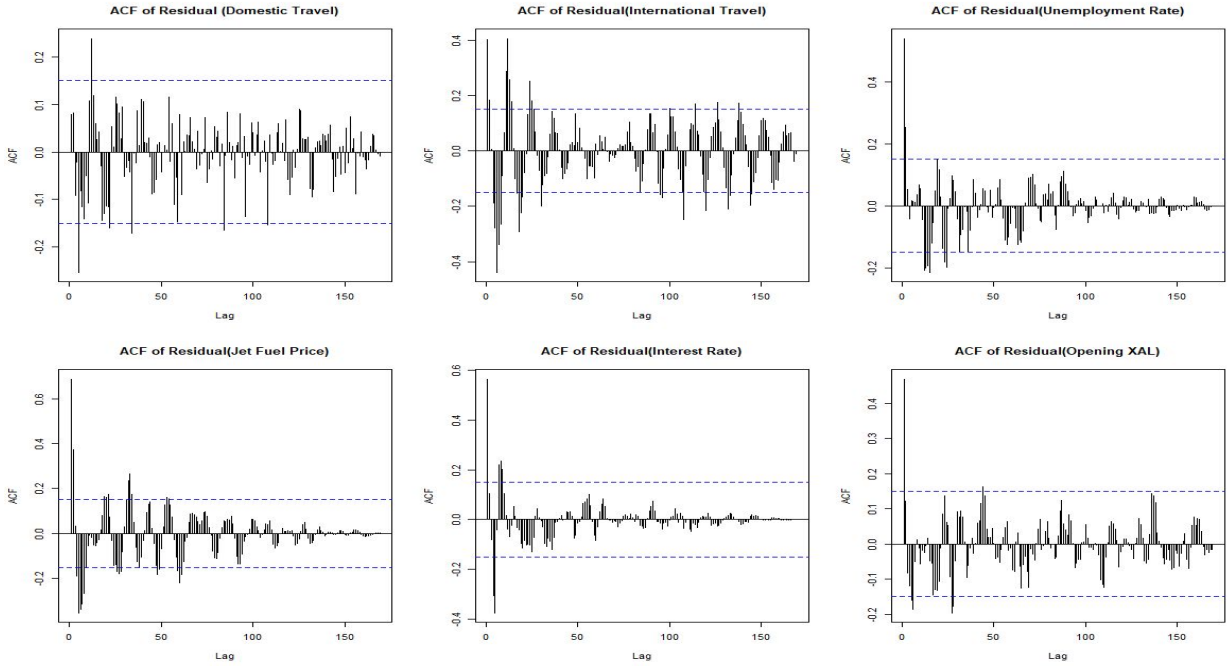


Figure 1.d: ACF of Residuals

2. Transformations

The original data for all six factors were not stationary. While various approaches involving differencing, de-seasonalizing with seasonal means, and taking the log were taken in an effort to transform factors to stationarity, most of the variables still remained in non-stationarity. We believe that several factors contributed to this; for instance, the financial crisis in 2008-2009 caused a severe downturn in the economy, resulting in unexpected distortions of unemployment and interest rate. After many attempts to remedy this issue, we were unable to create perfectly stationary time series.

However, we were able to generate time series closely resembling stationarity for our most important factors, domestic and international travel. In figures 2.b below, we can see from the ACF plots of the original, differenced, deseasonalized, and differenced & deseasonalized time series that by taking the difference and deseasonalizing with the seasonal means model, we are able to de-trend and remove the discrepancies that exist between months.

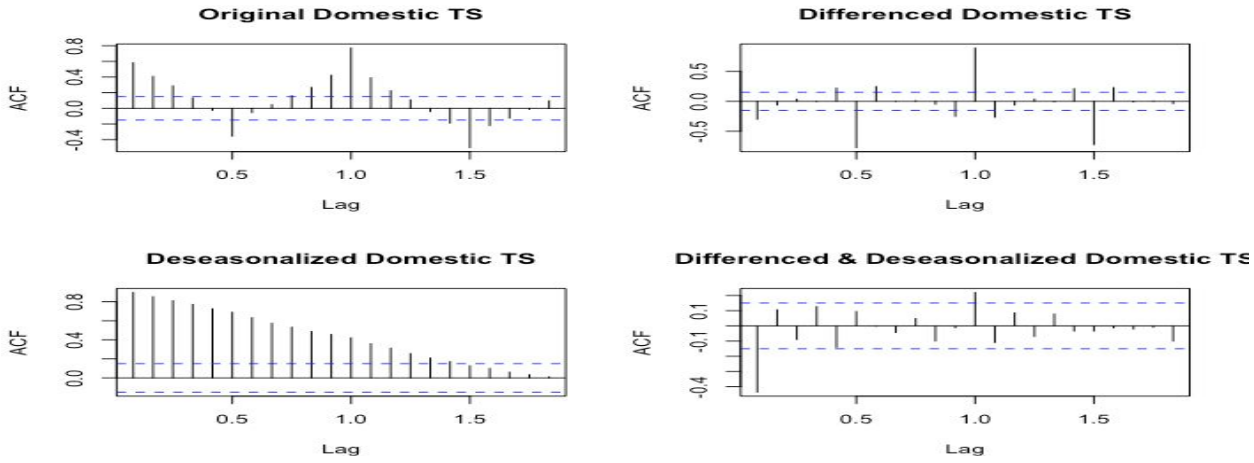


Figure 2.b: ACF plots of Original, Differenced and Deseasonalized Data for Domestic

Note that we still see a spike at the 1-year lag. Our efforts to remove this by differencing once again with the yearly lag was unsuccessful; we suspect this to be a result of not having enough data points (in terms of number of years) to be able to identify a long enough trend.

Similarly, we examined and performed identical steps to generate ACF plots for the original, differenced, deseasonalized, and differenced & deseasonalized time series of international travel.

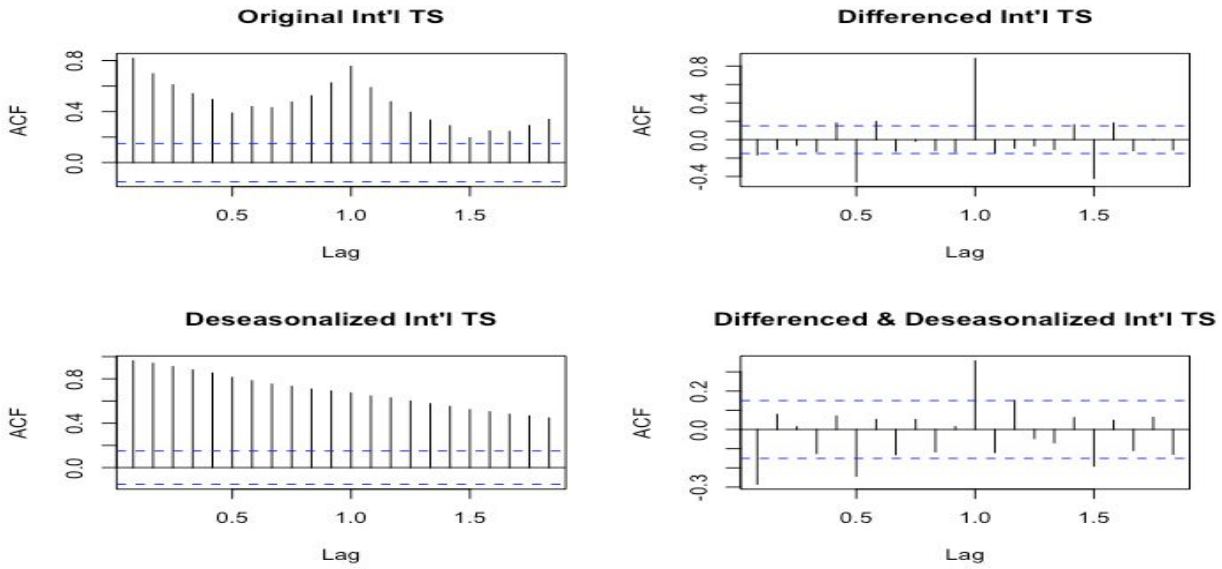


Figure 2.c: ACF plots of Original, Differenced and Deseasonalized Data for International

Again, the resulting time series is not perfectly stationary, but we can see that the differenced & deseasonalized time series more closely resembles a stationary ACF plot.

3. Vector Autoregression Model (VAR)

We fit six separate VAR models to determine the model with the best forecasting precision. The data was split to predict domestic flights and international flights. Each were modeled using an unrestricted VAR using all 5 factors, an unrestricted VAR using a subset of the factors, and a restricted VAR with all 5 factors. We split the data into training and test sets using the first 171 months for training and the remaining 12 months as testing. These last 12 points are used as test data to measure the accuracy of the forecasted values.

3.1 Unrestricted VAR with all factors

The unrestricted VAR model jointly examines movements in air traffic, unemployment rate, interest rate, jet fuel price, and the opening price of the Arca Airlines Index, a basket of airline industry stocks. We conduct our analysis in two separate parts, one for domestic air traffic and the other for international traffic. As mentioned above, we leave our last 12 data points as test data, so that we can train our model on the rest of the data points and forecast. The performance of the model is illustrated later in the forecasting section.

For both domestic and international traffic models, we arrive at a VAR(20) model. We realize that a VAR(20) model is quite complex; there are 20(lags)*5(factors) predictors that are involved in the model. Thus, we explore methods involving linear regression and restricted VAR to perform variable selection in the subsequent sections.

3.2 Unrestricted VAR with subset of factors (using regression)

To fit the subset VAR models, we first ran a linear regression using unemployment rate, jet fuel price, interest rate, and opening XAL price to predict both travel factors. From these models, we found that jet fuel price and opening XAL price significantly impacted domestic travel. Similarly, these two factors and interest rate impacted international travel. Moving forward in this section (3.2), when we refer to the domestic VAR model, this includes domestic travel, jet fuel price, and opening XAL price. When we refer to international VAR, this means international travel, jet fuel price, opening XAL price, and interest rates.

Using the three factors, the data was split into training and testing sets. The training included the first 173 months, and the test set was the remaining 10 months. A VAR(15) model was found to be the best fit on the training data. Using this model, we forecasted the next 10 months of domestic flights. Figure 4a shows the original data and the forecasted values. From the graph, the predictions seem quite good. Next, we used the same procedure to fit the international data. The VAR(19) model was used to forecast the test set (seen in Figure 4b).

After fitting the models, we performed a few diagnostic tests on the residuals. Although the model forecasted extremely well (as you will see in section 5), the residuals were not normally distributed, did not have constant variance, and were correlated. The domestic model is found to be stable, whereas the international model has 10 roots greater than 1.

3.3 Restricted VAR

The restricted VAR model selects lag-variable combinations that are significant in estimating the joint movements of the time series. Using the restrict() function in the VARS package, we identify restricted VAR models for the domestic and international travel models.

For both domestic and international models, the number of variables is reduced drastically to 18, resulting in a much simpler models. While predictions from this model may not be as accurate as the unrestricted model in certain circumstances, it can certainly be helpful in situations involving a large number of factors. The prediction errors for these models are discussed in the forecasting section.

4. Autoregressive Integrated Moving Average Model (ARIMA)

We began by splitting data into training and testing sets. In this section, our analysis will focus on domestic travel and international travel. The ARIMA model was trained on first 171 months, and we tested the model against the remaining 12 months. Using differenced data for domestic and international travel, we built two seasonal ARIMA models for each dataset using the auto.arima() function.

We found seasonal ARIMA(5,0,3)x(1,0,0) with period 12 for the domestic travel to be the best fit. To check the stationarity of the fitted model, we performed diagnostic tests of the residuals. ACF plot of residuals decay to zero after the first lag and all lags after that are within the confidence bands. Also, high p-value of Ljung-Box test indicates that there is no correlation among the residuals.

	AR1	AR2	AR3	AR4	AR5	MA1	MA2	MA3	SAR1
Coefficient	0.756	-0.541	0.428	0.158	-0.124	-1.193	0.755	-0.532	0.876
SE	0.370	0.397	0.161	0.156	0.101	0.377	0.568	0.229	0.035

Table 4.a: Result from Seasonal ARIMA Model For Domestic Travel

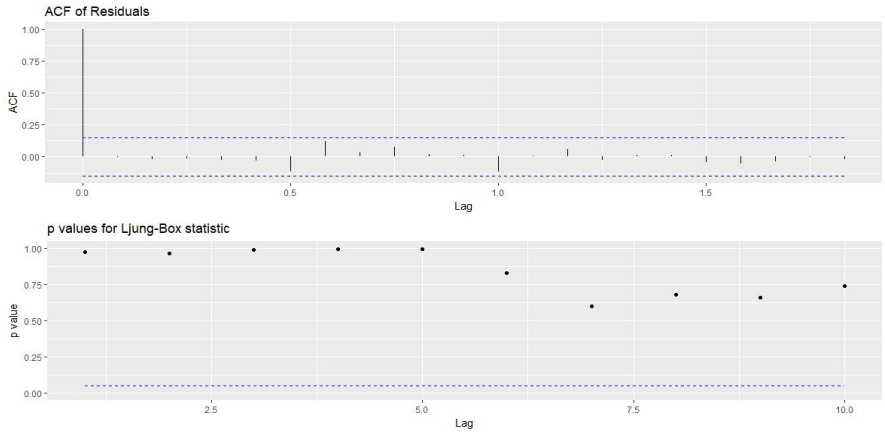


Figure 4.a: Diagnostic Domestic ARIMA

For international travel, we have seasonal ARIMA(1,0,1)x(0,1,1) with period 12. Once again, we tested residuals to check the stationarity of the model. The ACF plot of residuals and Ljung-Box test confirmed that the residuals are not correlated to each other.

	AR1	MA1	SMA1
Coefficient	0.562	-0.928	-0.119
SE	0.091	0.049	0.084

Table 4.b: Result from Seasonal ARIMA Model For Domestic Travel

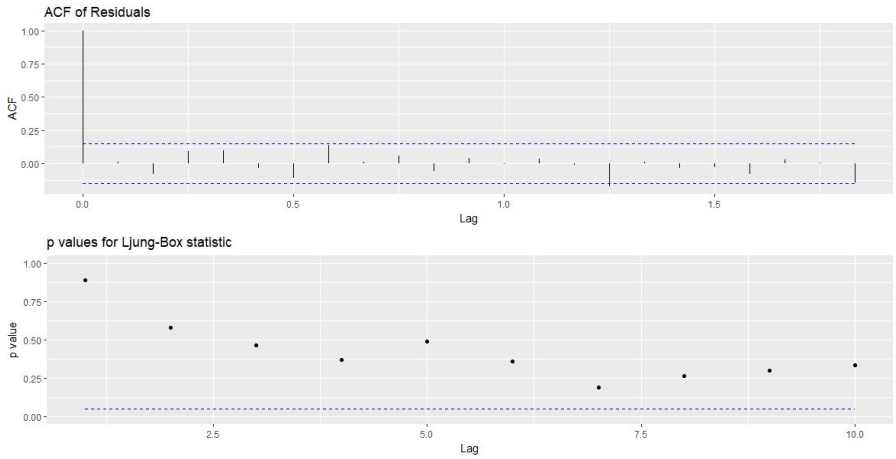


Figure 4. b: Diagnostic International ARIMA

5. Forecasting/Model comparison

After creating all the models, we used them to forecast 12 months out. We have provided the plots for these in figures 5a and 5b. Additionally, we used these forecasted values to measure accuracy. Note that because our model was trained on transformed time series, our predictions were converted back to original data. This was done for the log-differenced data by adding it back to the previous time point and taking the exponential of the resulting value. These accuracy measures are shown in Table 5.

Below are the plots generated by the test data and our forecasts for every model. In figures 5a and 5b, the black lines are from unrestricted sub VAR model, the red lines are from the unrestricted full VAR, the blue lines are from the restricted VAR, the green lines are from seasonal ARIMA model, and the dots are the true values.

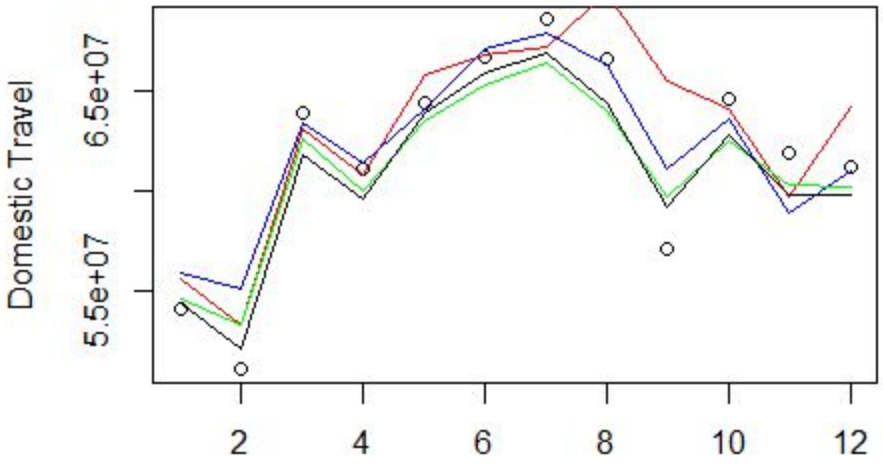


Figure 5.a: Forecasting of Domestic Travel

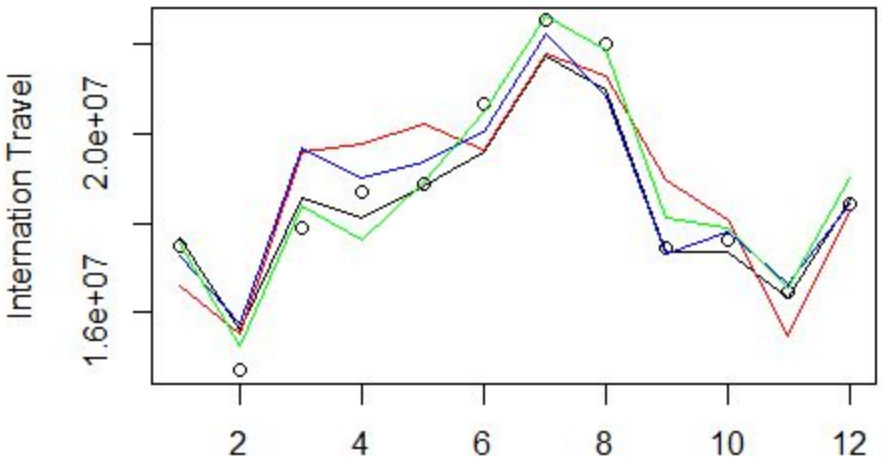


Figure 5. b: Forecasting of International Travel

Below is a chart measuring forecasting accuracy. The metrics we used to are Mean Squared Prediction Error, Mean Absolute Prediction Error, Mean Absolute Percentage Error, and Prediction Measure.

	MSPE	MAE	MAPE	PM
Unrestricted Domestic Full	8.866e+12	2087213	0.035	0.343
Restricted Domestic	3.869e+12	1380335	0.024	0.150
Unrestricted Domestic Sub	2.348e+12	1349611	0.022	0.096
Seasonal ARIMA Domestic	3.143e+12	1644211	0.026	0.121
Unrestricted International	1.078e+12	949852.9	0.052	0.235
Restricted International	5.477e+11	538715.7	0.029	0.119
Unrestricted International Sub	3.508e+11	445438	0.024	0.082
Seasonal ARIMA International	2.199e+11	353093.9	0.019	0.047

Table 5. Prediction Accuracy

Observations from the table above are:

- The subsetted Unrestricted VAR model performs the best across the board in predicting domestic travel.
- The seasonal ARIMA model performs the best across the board in predicting international travel.
- The full unrestricted model performs poorly in predicting both domestic and international travel; this may be due to unincluded external factors, insufficient number of data points, inadequate transformations of the data, etc.
- Models perform differently on domestic and international data.
- Note that the seasonal ARIMA model offers a simplicity advantage over VAR models. Given the small difference in prediction accuracy in estimating domestic travel, we deem that it may be sufficient to use seasonal ARIMA in this case.

6. CONCLUSION

The original purpose of this project was to effectively model air passenger traffic based on jet fuel price, XAL price, interest rates, and unemployment rates. By generating and comparing several different models, we were able to identify the following results. First, the unrestricted VAR model performed poorly in predicting both domestic and international travel. A variety of issues could have contributed to this, including but not limited to the exclusion of other external factors, insufficient number of data points, and non-stationarity of the data. Second, the univariate method we used, i.e. seasonal ARIMA, modeled traffic as well or better than any of our multivariate methods. This is particularly surprising as it is simpler model than VAR. While domestic travel was best estimated by the unrestricted subset VAR model, the difference in prediction accuracy between it and the seasonal ARIMA was negligible, especially given the advantage in simplicity ARIMA has over VAR. Based on our analysis, our recommendation is to use the seasonal ARIMA model to predict both domestic and air traffic.

SUBJECT MATTER IMPLICATIONS

From the decomposition of the domestic and international travel, we confirmed an increasing trend as well as seasonal boosts in traffic during the summer. We believe that this analysis can help airline traffic controllers better plan for future crowds. Additionally, our VAR models could be extended to include other external factors; insiders within the industry that have superior knowledge and data can expand our model by choosing factors that directly affect air passenger traffic based on their industry experience.

FURTHER INQUIRIES

We would like to see how the time series varies with more/different variables. As we are not experts on air passenger traffic, our best guess was to use factors that were closely related to the economy, based on the presumption that air traffic would be highly correlated to it. However, there may be other factors or metrics that insiders within the airline industry could include in the model, that we may not be aware of. We realize that the addition of such factors may improve the model drastically.

Moreover, in the future, we would like to investigate this issue over a longer period of time. One of the problems we faced was that our timeframe included the financial crisis, making it very difficult for us to transform factors such as unemployment and interest rate to stationary time series. Given a longer time frame, we believe that we may be able to capture longer term trends and mitigate the distorting effects of such shocks.

DATA SOURCES

- Bureau of Transportation Statistics: https://www.transtats.bts.gov/Data_Elements.aspx?Data=1
- Bureau of Labor Statistics: https://data.bls.gov/timeseries/LNS14000000
- U.S Energy Information Administration https://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=pet&s=f000000__3&f=a
- Federal Reserve https://fred.stlouisfed.org/series/DGS1MO