Flight Predictability Analysis of Los Angeles International Airport

Seung Woo Choi, Sai Gujja, Pedro Maia

1. Summary

The Los Angeles International (LAX) airport, one of the world's busiest, experienced a substantial surge in traffic during fiscal year 2022, necessitating a comprehensive analysis of its flight operations. With a focus on enhancing flight predictability, our research aims to identify key characteristics influencing time series patterns, explore trends between domestic and international flights, and assess external factors affecting flight operations. We considered internal factors including seasonality, trends, and arrivals and departures data and external factors such as weather conditions and major events. Our approach involves a time series analysis incorporating these internal and external factors to understand their effects on flight operations. We utilized various univariate and multivariate models, including ARIMA, ARMA-GARCH, ARIMAX, VAR, and VARX, to predict the time series for on-time arrivals and on-time departures. Our evaluation of the models showed that VARX had the best performance for both response variables. Furthermore, our analysis on the external factors demonstrated that precipitation was significant in making predictions for the ARIMAX and VARX models. By comparing trends across domestic and international flights, we found that domestic and international arrivals experienced different trends between 1990 and 2023.

2. Introduction

2.1 Motivation

The LAX airport is among the busiest airports in the world. During the fiscal year 2022, the airport had a total of 60.7 million passengers, with 48.5 million passengers (i.e., 80 percent) flying domestically and 12.2 million passengers (i.e., 20 percent) flying internationally.[6] With a 55.1 percent growth rate in departures and arrivals, LAX saw a huge surge in traffic, which resulted in the need to better understand and predict the performance of its flight operations.[6] As passengers continue to seek its flight services, the LAX airport needs to maintain and improve its flight operations such that passengers are satisfied with its service and return for additional flights. To improve its performance, LAX will need to analyze its flight operations to determine factors – both internal and external – that contribute to the predictability of its flight operations. To this end, we have conducted air travel analysis and would consult the LAX airport on the following three research questions: (1) what characteristics contribute most to the predictability of the time series, (2) what trends exist between flying domestic versus flying international, and (3) what external factors if any contributed to predicting capability of flight operations.

2.2 Research Overview

To achieve the three objectives outlined above, we reviewed the literature to better understand the features and trends required to complete our analysis.

2.2.1 Predictability of Flight Operations

Identifying the key characteristics of the time series can help us to understand the patterns and trends of flight operations. On that note, previous research has found that analysis of factors such as seasonality, trends, and cycles are crucial for predicting the number of flights, planning and optimizing resources for seasonal peak periods, and improving the efficient use of airport capacity and capital.[3]

2.2.2 Trends between Domestic vs. International Flights

To gain further insights into the dynamics of the airport's operations, we can analyze the relationship between domestic and international flights. The factors that are important in understanding this relationship include the volume of international versus domestic flights, the seasonality, and the impacts of external and internal events.[3]

2.2.3 External Factors

External factors that can affect flight operations include: (1) weather conditions, (2) economic conditions, (3) major events or natural disasters, and (4) airline industry or government policies.[4][5][8] Weather conditions can lead to delays and cancellations, even for flights not directly affected by the weather event, so historical weather patterns can help in predicting potential disruptions and improve flight scheduling and route adjustments. Economic conditions like recessions and economic growth rates can heavily influence passenger demand and flight volumes. Major events or natural disasters, such as the COVID-19 pandemic, can significantly affect flight volume and passenger count. Likewise, changes in the airline industry or government policies can also contribute to fluctuations in flight operations.

2.3 Prior Expectations

For our analysis, we used the provided LAX arrivals and departures data, LAX precipitation and temperature data from the National Weather Service, and COVID-19 dates from the CDC.

We expected that (1) features like flight type, on-time performance, and number of passengers would have the most weight in predicting flight operations, (2) domestic flights would exhibit different trends than international flights, and (3) external factors like temperature and precipitation would contribute to the predictability of flight operations.

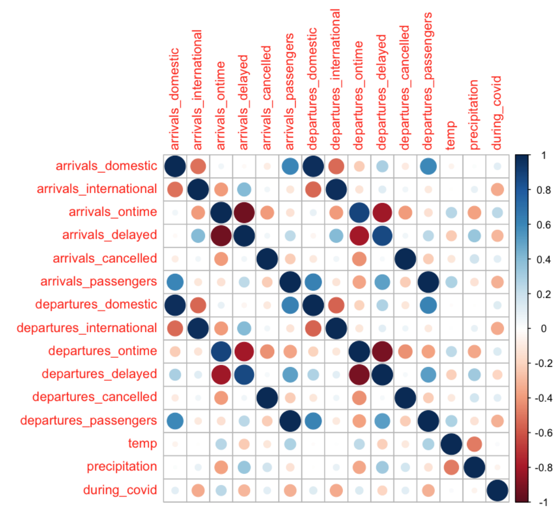
In terms of modeling, we expected that models that can capture seasonality and volatility would be able to forecast flight operations with a high degree of certainty. We planned to build univariate models, such as ARIMA, ARMA-GARCH, and ARIMAX, as well as multivariate models, like VAR and VARX, that could provide meaningful forecasts for on-time arrivals and departures. Moreover, we sought to incorporate exogenous variables, such as weather data and COVID-19 pandemic dates, into our time series analysis by applying models like ARIMAX and VARX to gain a better understanding of how external factors affect on-time arrivals and departures.

3. Analysis

Prior to tackling the three objectives outlined in the **Motivation** section, we conducted exploratory data analysis (EDA) to gain a better understanding of the data and relationships among all the time series.

To begin, we created a correlation matrix that illustrates the correlation coefficients among all the variables, as shown in **Figure 1**.

Figure 1. Correlation Coefficients

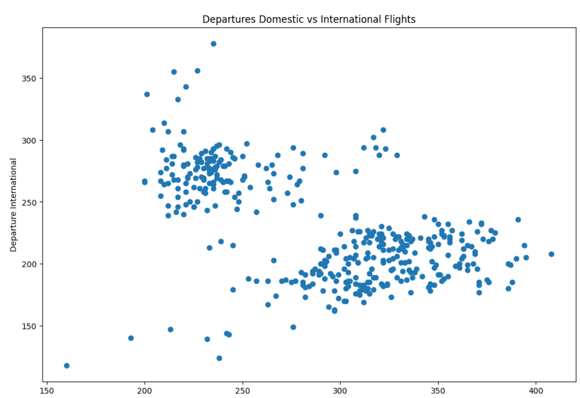


The correlation matrix in **Figure 1** shows that precipitation is moderately correlated with temperature and the on-time and delayed statuses of flights. Temperature, on the other hand, is only moderately correlated with precipitation, which may indicate that it has a lesser effect on other variables like the on-time status of flights. The on-time status of arrivals appears to be strongly correlated with arrivals_delayed, departures_ontime, and departures_delayed, whereas the on-time status of departures appears to be strongly correlated with arrivals_ontime, arrivals_delayed, and departures_delayed. These findings highlight the importance of considering exogenous factors and the flight status of other flights when evaluating on-time arrivals and departures.

The stationarity of all the individual time series was analyzed to determine if first order differencing is necessary for modeling. The ACF and PACF charts were plotted to visually inspect if the original time series required differencing and transformations to exhibit stationarity. Based on the plots, it appeared that first order differencing resulted in time series that were more stationary than the original time series. Furthermore, while logarithmic differencing helped with stabilizing the variances of the time series, it was not implemented due to limitations from two variables (i.e., precipitation and during_covid). The Augmented Dickey-Fuller (ADF) test was also implemented to test for stationarity and revealed that all of the original time series exhibit stationarity at a p-value of 0.10, while three time series remained non-stationary at a more restrictive p-value of 0.05. Notably, first-order differencing ensured stationarity for all time series at a p-value of 0.05.

EDA was also conducted on the domestic and international flights data. **Figure 2** below illustrates two clusters from plotting domestic departures against international departures.

Figure 2. Domestic vs. International Departures



Given the two distinct clusters of data points for domestic and international departures, the scatter plot in **Figure 2** suggests that LAX airport may have an operational limitation. These clusters are similar for arrivals and departures. The median ratio of domestic to international flights is 1.48 for arrivals and 1.46 for departures, suggesting that the LAX airport is limited in the total number of flights (i.e. arrivals or departures) that it can accommodate on any given date.

Lastly, EDA was completed on the exogenous factors to uncover if any interesting relationships exist.

Figure 3 below displays the average temperature and precipitation at the LAX airport from 1990 to 2023. The precipitation values have been multiplied by 500 for extrapolation to make the relationship between variables more visible. As the plot shows, there is a clear relationship between temperature and precipitation at LAX. In general, lower temperatures are associated with higher precipitation. This is because cold air can hold less moisture than warm air, so when cold air passes over LAX, it is more likely to condense and fall as precipitation.

Figure 3. Temperature vs. Precipitation

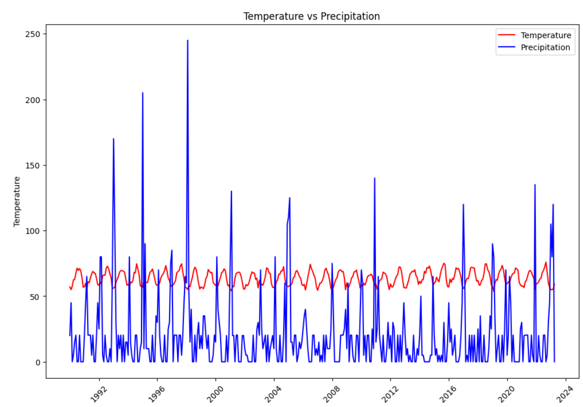
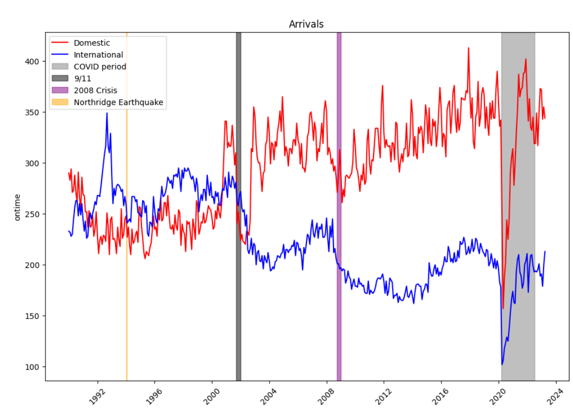


Figure 4 below illustrates the occurrence of major world events and how domestic and international on-time arrivals changed between 1990 and 2023.

Figure 4. Major Events Overlayed on Domestic and International On Time Arrivals



Major historic events significantly disrupted flight operations at LAX. The 1994 Northridge earthquake caused significant damage to the airport's runways and terminals, resulting in a closure for several days, explaining the sudden drop in arrivals around that time. The 9/11 attacks led to increased security measures at LAX, which resulted in longer passenger processing times and delays. This is likely why we see a decrease in arrivals around that time. The 2008 financial crisis led to a decrease in travel demand, which resulted in a decrease in the number of flights at LAX. The COVID-19 pandemic caused a steep decline in arrivals between 2020 and 2021. Fortunately, there has been a post-pandemic rebound, and a trajectory of growth appears likely in the future. The pandemic was the external factor with the largest impact on both domestic and international travel in history.

3.1 Predictability of the Time Series

The first objective of our analysis involved determining the characteristics that contribute the most to the predictability of the time series. We defined two time series to be our response variables - i.e. arrivals_ontime and departures_ontime - since these two variables best reflect the idea of "flight operations". Once the data sets were appropriately cleaned, prepared, and merged into a single data set, we conducted exploratory data analysis on the training data to investigate various qualities of the data and individual time series. Furthermore, we utilized various univariate and multivariate time series models to forecast the next six months of unseen data. The performance of these models were evaluated using three metrics: (1) mean squared error (MSE), (2) mean absolute percentage error (MAPE), and (3) precision measure (PM). The MSE and MAPE are loss functions that improve as the values approach 0, whereas the PM measures the variability of the predicted values and observed values and improves as it gets closer to 1. A comparison of the models utilizing these three evaluation metrics allowed us to determine the best model for predicting arrivals_ontime and departures_ontime for the LAX airport.

3.1.1 Methodology

Six time series models were built to forecast the next six months of unseen arrivals_ontime and departures_ontime data. These models include ARIMA, ARMA-GARCH, ARIMAX, VAR (i.e. unrestricted and restricted), and VARX. The methodology for each model will be discussed below. The time series data were split into training and testing data sets for each model.

For the univariate ARIMA model, an iterative method was applied to determine the optimal orders using BIC. The orders for p and q were limited to 8, and the order for d was limited to 1. The optimal models based on BIC values and model complexity were found to be ARIMA(2,0,1) for both arrivals_ontime and departures_ontime.

An iterative method was also applied to the univariate ARMA-GARCH model to find the best orders using BIC. The maximum orders were set to 8 for p and q and 2 for m and n. Once the iterative process was completed, it was clear that the optimal models were ARMA(3,1)-GARCH(1,0) for arrivals_ontime and ARMA(3,2)-GARCH(1,0) for departures_ontime.

The ARIMAX model was fitted for the time series variables, arrivals_ontime and departures_ontime, along with exogenous variables temp and precipitation. ARIMAX(1,1,1)(0,0,2)[12] was fitted for both arrivals_ontime and departures_ontime and includes autoregressive, moving averages, seasonal components along with temp and precipitation. Model performance metrics such as AIC, AICc, BIC and training set error measures were used to assess the goodness of fit.

For the multivariate VAR model, stepwise regression was applied for variable selection. The top three variables selected to predict arrivals_ontime were arrivals_ontime, arrivals_delayed, and arrivals_cancelled, whereas the top three variables chosen to predict departures_ontime were departures_ontime, departures_delayed, and departures_cancelled. The unrestricted VAR model was built first using the Hannan-Quinn (HQ) information criterion for order selection and the first order difference of the training data. The arrivals_ontime used an order of 4, whereas the departures_ontime used an order of 20. The restricted VAR model was built next using the unrestricted VAR model.

The multivariate VARX model was estimated for the time series dataset with endogenous variables including arrivals_ontime, temp, arrivals_delayed, arrivals_cancelled, and precipitation. The VAR model was constructed with a lag order (p) of 4, and the coefficients for each lag along with the constant term were estimated using OLS regression. The Wald test was run to estimate the coefficients' significance, and the Jarque-Bera (JB) test was run to to check the normality of the residuals.

3.1.2 Model Results

As mentioned before, the performance of the models was measured using three evaluation metrics: (1) MSE, (2) MAPE, and (3) PM. Below are two tables that show how each model performed on each metric. **Table 1** displays the results for arrivals_ontime, whereas **Table 2** displays the results for departures_ontime.

Table 1. Model Results for On Time Arrivals

Model	MAPE	MSE	PM
ARIMA	0.0760	0.00436	2.133
ARMA-GARCH	0.0885	0.00595	2.907
ARIMAX	0.101	0.00709	3.467
Unrestricted VAR	0.101	0.00722	3.528
Restricted VAR	0.187	0.0240	11.744
VARX	0.0508	0.00249	1.216

Table 2. Model Results for On Time Departures

Model	MAPE	MSE	PM
ARIMA	0.0556	0.00272	2.197
ARMA-GARCH	0.0701	0.00384	3.105
ARIMAX	0.0839	0.00520	4.201
Unrestricted VAR	0.0637	0.00348	2.811
Restricted VAR	0.0623	0.00334	2.700
VARX	0.0332	0.00118	0.954

Based on the literature, we expected models that included time series on flight type, on-time performance, number of passengers, and external factors like temperature and precipitation would perform the best. The results we achieved via experimentation supports the literature in that the best-performing model was VARX. VARX stands for exogenous VAR and is a multivariate model that performs well when trends and seasonality are present in the time series. For arrivals_ontime, the VARX model achieved a MAPE of 0.0508, an MSE of 0.00249, and a PM of 1.216. For departures_ontime, the VARX model achieved a MAPE of 0.0332, an MSE of 0.00118, and a PM of 0.954. In both cases, each score was the most optimal for each evaluation metric since MAPE and MSE are loss functions that improve as they approach 0 and PM reflects the variability of the predicted values and actual values and improves as it approaches 1.

Other models, such as ARIMA and ARMA-GARCH, also resulted in fairly accurate predictions with minor variability. ARIMAX and the unrestricted VAR had similar performances on the evaluation metrics but performed worse than the ARIMA and ARMA-GARCH models. Lastly, the restricted VAR model had the worst performance on all three evaluation metrics but provided valuable evidence that VARX may be a good model to fit the time series.

3.2 Trends Between Domestic vs. International Flights

The second objective of our analysis dealt with exploring trends between domestic and international flights. To do this, we analyzed the time series for domestic arrivals (i.e. arrivals_domestic) and international arrivals (i.e. arrivals_international). We applied four trend estimation approaches, including moving average, parametric quadratic polynomial, local polynomial, and splines regression, for both time series. **Figures 3 and 4** illustrate the fits for all four trend estimation methods.

Figure 3. Trend Estimation for Domestic Arrivals

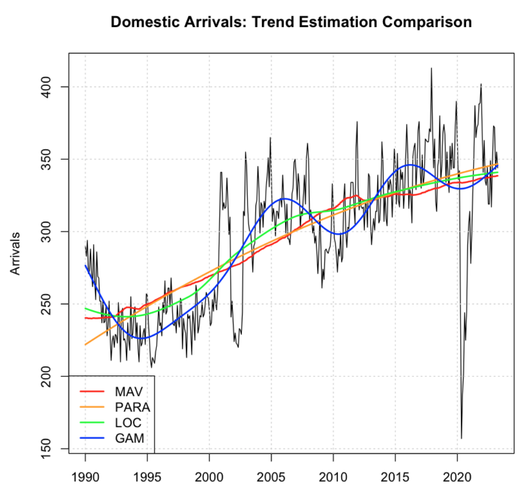
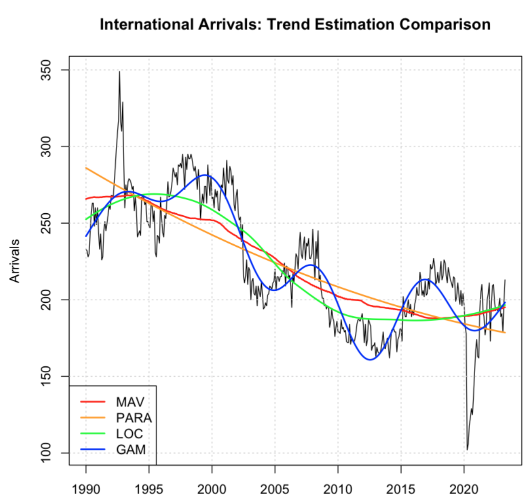


Figure 4. Trend Estimation for International Arrivals



For domestic and international arrivals, it appears that the splines regression model resulted in the best fit. For domestic and international arrivals, none of the models were able to incorporate the huge drop in arrivals that occurred when the COVID-19 pandemic hit. While there were fluctuations over time, domestic arrivals appear to have a clear upward trend between 1990 and 2023, whereas international arrivals experienced a clear downward trend between this time period, which aligns with our expected outcome. While trend analysis allows us to observe that domestic and international arrivals saw opposing trends between 1990 and 2023, it does not reveal what factors caused these trends to occur.

3.3 External Factors on Predicting Flight Operations

The third objective of our analysis involved determining the external factors that contributed to predicting flight operations. We added external factors like temperature, precipitation, and a binary variable called during_covid, to the analysis and constructed additional models to explore if exogenous factors played an important role in forecasting arrivals_ontime and departures_ontime. Specifically, we built the ARIMAX and VARX models on internal and external time series and found that precipitation was statistically significant in the ARIMAX and VARX models. Temperature was not found to be statistically significant in these models, and the during_covid time series was excluded since the binary nature of the time series was not particularly insightful. Through the magnitude of the coefficient, p-value for each lag, and Wald test, it was evident that the exogenous factor, precipitation, did help in predicting flight operations.

4. Conclusion

4.1 Implications on Flight Operations

Through our time series analysis of air travel at the LAX airport, we were able to achieve the three objectives of (1) determining the characteristics that contribute the most to the predictability of on-time arrivals and departures, (2) identifying the trends between domestic and international flights, and (3) understanding the effects of external factors on the predictability of flight operations.

The key characteristics that contributed to the predictability of on-time arrivals and departures include variable selection using stepwise regression, model orders informed by the Hannan-Quinn criterion, significant coefficient estimates indicating variable impact, stable roots of the characteristic polynomial, effective residual analysis with lower standard errors and higher R-squared values, global significance assessed through the Wald test, the JB test, and performance metrics (i.e. MAPE, MSE, PM). Additionally, for the ARIMA model on departures_ontime, the coefficients of autoregressive and moving average terms, along with seasonal moving average terms, played a crucial role in capturing temporal patterns.

Domestic and international flights trends diverge, with domestic flights experiencing an upward trajectory since the 1990s and international flights on decline since the 2000s. These trends were punctuated by fluctuations coinciding with major historical events such as the 9/11 attacks and the COVID-19 pandemic. Despite these fluctuations, the underlying trends remained consistent over this time period.

(3) The VARX model emerged as the most effective predictor of on-time arrivals and departures. This multivariate time series model, incorporating exogenous factors such as precipitation and temperature, demonstrated the best predictive capabilities. It was evident from the output that the precipitation had a higher significance compared to temp in predictive capabilities.

It is important to highlight that the factors mentioned in this project are interconnected and can influence each other, making it difficult to pinpoint a single reason for the observed trends. However, by considering these diverse factors, we can gain a more comprehensive understanding of the dynamics impacting flight patterns at LAX and predict future trends in the aviation industry.

4.2 Further Questions

In our analysis, we incorporated exogenous variables such as temperature, precipitation, and COVID-19 pandemic dates. In order to expand the scope of the analysis, additional factors like economic status (i.e. measured by the growth rate in gross domestic product) and flight prices can be included to gain a better understanding of flight operations.

The trend analysis conducted in this project showed that domestic and international arrivals experienced opposing trends between 1990 and 2023 but did not uncover the factors that were responsible for causing these trends. Future work that looks to expand on our analysis could dive into potential factors that caused the diverging trends to occur and determine whether the decline in international arrivals will continue.

References:

[1] “CDC Museum COVID-19 Timeline”. Centers for Disease Control and Prevention, https://www.cdc.gov/museum/timeline/covid19.html. Accessed 20 Oct. 2023.

[2] “Climate Data Online”. National Centers for Environmental Information, https://www.ncei.noaa.gov/cdo-web/. Accessed 6 Oct. 2023.

[3] Dobruszkes, F., Decroly, J.-M., & Suau-Sanchez, P. (2022). The monthly rhythms of Aviation: A Global Analysis of Passenger Air Service seasonality. *Transportation Research Interdisciplinary Perspectives*, 14, 100582. https://doi.org/10.1016/j.trip.2022.100582

[4] Etani, N. (2019). Development of a predictive model for on-time arrival flight of airliner by discovering correlation between flight and weather data. *Journal of Big Data*, 6(1). https://doi.org/10.1186/s40537-019-0251-y

[5] Hsu, C.-W., Liu, C., Liu, Z., & Mostafavi, A. (2023). Unraveling extreme weather impacts on air transportation and passenger delays using location-based data. arXiv.org. https://arxiv.org/abs/2304.12084

[6] “LAX Annual Financial Report-FY 2022”. Los Angeles World Airports, https://www.lawa.org/lawa-investor-relations. Accessed 6 Oct. 2023.

[7] “Los Angeles, Los Angeles International Airport”. National Weather Service, https://www.weather.gov/wrh/timeseries?site=KLAX. Accessed 6 Oct. 2023.

[8] Rodríguez-Sanz, Á., Cano, J., & Rubio Fernández, B. (2021). Impact of weather conditions on airport arrival delay and throughput. *IOP Conference Series: Materials Science and Engineering*, 1024(1), 012107. https://doi.org/10.1088/1757-899x/1024/1/012107