

Forecasting Air Traffic Recovery at SFO After COVID-19

Abstract

This study analyzes monthly passenger traffic at San Francisco International Airport to evaluate whether SARIMA models can forecast recovery after an unprecedented demand shock. The COVID-19 pandemic caused a near-total collapse in air travel, yet seasonal patterns proved remarkably resilient. A $\text{SARIMA}(1, 1, 0)(0, 1, 1)_{12}$ model identified via AIC minimization achieves strong short-term forecast accuracy ($\text{MAPE} = 3.21\%$) and passes Ljung-Box diagnostics for residual autocorrelation, though non-normal residuals suggest potential underestimation of forecast uncertainty. Forecast intervals widen substantially with horizon (666% expansion over 30 months), limiting long-term predictive reliability. Extending previous STATS 531 work on pre-COVID airline traffic, we confirm that seasonal structure survived the pandemic, but trend did not. The results demonstrate that SARIMA models can provide accurate short-term forecasts after structural breaks, though expanding uncertainty limits their value for long-term planning.

1 Introduction

The COVID-19 pandemic brought global aviation to a near halt. San Francisco International Airport (SFO), a major international gateway, experienced the largest demand shock in its history. Understanding whether passenger traffic will recover to pre-pandemic levels, and when, is critical for airport planning, airline operations, and economic forecasting. This study addresses two central questions: First, what are the underlying temporal patterns in SFO’s passenger traffic, and how did the pandemic disrupt them? Second, can seasonal ARIMA models, estimated on data spanning both pre- and post-pandemic periods, reliably forecast the recovery trajectory?

Using monthly passenger data from July 1999 to November 2025, we conduct an exploratory analysis, select a parsimonious SARIMA specification, and evaluate its 30-month forecast performance. We then assess where the model succeeds, where it fails, and what these failures reveal about the limits of time series methods under structural breaks.

2 Data Exploration

The dataset consists of monthly passenger counts at San Francisco International Airport (SFO), obtained from the San Francisco Open Data Portal (“Air Traffic Passenger Statistics” 2025). Each record represents the total number of arriving and departing passengers for a specific airline in a given month.

The dataset spans from July 1999 to November 2025, containing over 39,000 records across 138 unique airlines. Passenger traffic at SFO is highly concentrated, with the top five airlines controlling the vast majority of the market.

3 Exploratory Data Analysis

3.1 Time Series Visualization



Figure 1: Monthly passenger traffic at SFO (thousands) with annual maxima (red) and minima (green)

SFO passenger traffic exhibits three distinct regimes. Pre-2020 shows steady growth with strong yearly (12-month) seasonality, peaking in 2019. The COVID-19 collapse in March-April 2020 reduced traffic by over 95%, marking the lowest point in 26 years. Recovery began mid-2021 with rapid initial growth that has since decelerated. By mid-2023, traffic reached approximately 75% of the 2019 peak, with seasonal patterns remaining intact despite the disruption.

3.2 Seasonal Decomposition

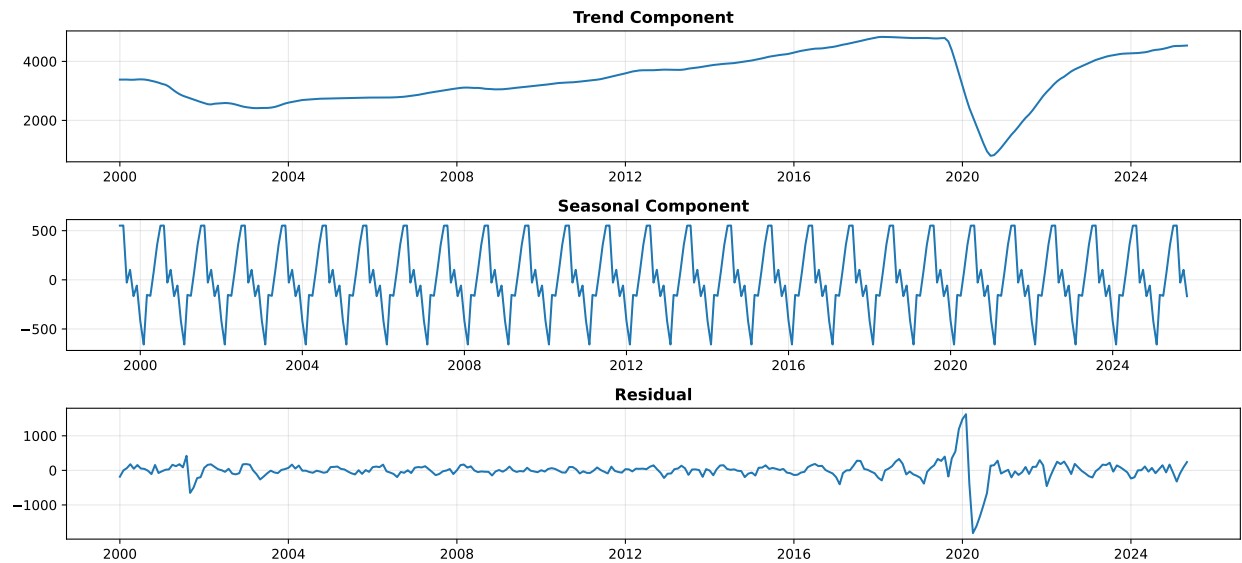


Figure 2: Additive decomposition of SFO passenger traffic

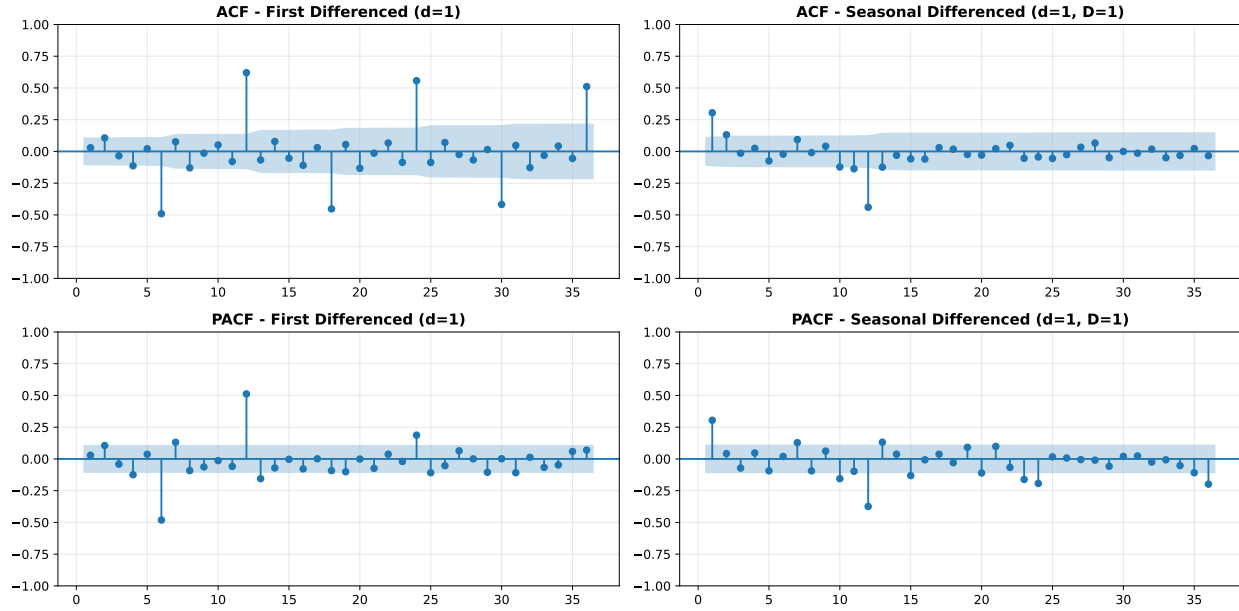
The additive decomposition reveals three key patterns (“Statsmodels: Statistical Modeling and Econometrics in Python” 2025). The trend captures pre-2020 growth, the pandemic collapse, and incomplete recovery. The seasonal component remains remarkably stable with constant amplitude throughout, indicating that the pandemic did not diminish seasonal variation. The residual component shows increased volatility during recovery, suggesting post-COVID behavior is less predictable.

3.3 Stationarity Assessment

Series	ADF Statistic	p-value	Stationarity
Original	-3.099343	0.026602	Non-stationary
First Difference (d=1)	-4.268054	0.000504	Stationary

The ADF test on the original series yields $p = 0.0266$, marginally significant at the 5% level but borderline (“Statsmodels: Statistical Modeling and Econometrics in Python” 2025). Combined with visual evidence of trend and structural break, first differencing is applied. The differenced series is clearly stationary with $p < 0.001$, confirming $d = 1$ is appropriate.

3.4 Autocorrelation Structure and Model Selection



After confirming stationarity with $d=1$, we examine autocorrelation patterns to guide model specification. The ACF of the first-differenced series shows significant spikes at lags 1 and 2, plus persistent seasonal spikes at lags 12, 24, and 36, indicating residual seasonality (“Statsmodels: Statistical Modeling and Econometrics in Python” 2025). This motivates seasonal differencing ($D=1$). After applying both $d=1$ and $D=1$, the ACF exhibits reduced seasonal autocorrelation, while the PACF shows spikes at lag 1 and seasonal lag 12. These patterns suggest $SARIMA(1, 1, 0)(0, 1, 1)_{12}$, where $p=1$ captures short-term persistence and $Q=1$ models seasonal dependence.

From the ADF test, we set $d = 1$ to achieve stationarity. The strong seasonal patterns visible in both the time series plot and the ACF after first differencing suggest that seasonal differencing is necessary, so we set $D = 1$.

4 Methodology

4.1 Train-Test Split

To evaluate forecast performance, we partition the series into a training set and a test set. The final 30 months of observations (from June 2023 through November 2025) are reserved for testing. The remaining 287 months from July 1999 through May 2023 are used for model estimation.

4.2 Model Specification

We model the series using the seasonal ARIMA (SARIMA) framework (“Statsmodels: Statistical Modeling and Econometrics in Python” 2025). A $SARIMA(p, d, q)(P, D, Q)_{12}$ model takes the form:

$$\phi(B)\Phi(B^{12})(1-B)^d(1-B^{12})^DY_t = c + \theta(B)\Theta(B^{12})\varepsilon_t$$

Table 1: SARIMA model comparison ranked by AIC

Model	AIC	BIC
SARIMA(1, 1, 0)(0, 1, 1, 12)	3791.56	3802.4
SARIMA(2, 1, 0)(0, 1, 1, 12)	3793.31	3807.76
SARIMA(1, 1, 1)(0, 1, 1, 12)	3793.4	3807.85
SARIMA(0, 1, 1)(0, 1, 1, 12)	3794.48	3805.32
SARIMA(1, 1, 0)(1, 1, 0, 12)	3868.35	3879.19

where B is the backshift operator, ϕ and θ are non-seasonal AR and MA polynomials of orders p and q , Φ and Θ are seasonal AR and MA polynomials of orders P and Q , and ε_t is white noise.

4.3 Information Criteria for Model Comparison

We compare candidate models using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC):

$$AIC = 2k - 2\ln(L)$$

$$BIC = k\ln(n) - 2\ln(L)$$

where L is the maximized likelihood, k is the number of parameters, and n is the sample size. Lower values indicate a better balance of model fit and parsimony. AIC serves as our primary selection criterion (“STATS 531 Lecture Notes: Model Selection and AIC” 2025).

4.4 Model Estimation and Selection

We estimate several candidate SARIMA models on the training data. Higher-order models frequently fail to converge, indicating overparameterization given the relatively short post-COVID recovery period. We therefore restrict our search to low-order, parsimonious specifications.

The model minimizing AIC is selected as our final specification. We then examine its residuals to verify that they approximate white noise using the Ljung-Box test and visual inspection of the residual ACF.

4.5 Forecasting and Evaluation

Using the selected model, we generate a 30-month ahead forecast covering the test period (June 2023 to November 2025). We report point forecasts along with 95% confidence intervals. Forecast accuracy is evaluated on the test set using:

- **Mean Absolute Error (MAE):** $\frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$
- **Root Mean Squared Error (RMSE):** $\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$
- **Mean Absolute Percentage Error (MAPE):** $\frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|$

Table 2: Ljung-Box test for residual autocorrelation

Lag	Test Statistic	p-value
12	16.9835	0.150218
24	22.8894	0.526335

5 Results and Discussion

5.1 Model Selection and Parameter Estimates

The candidate models ranked by AIC are shown in the table above. The SARIMA(1, 1, 0)(0, 1, 1)₁₂ model achieves the lowest AIC, balancing parsimony with fit quality. We select this as our final specification.

5.2 Residual Diagnostics

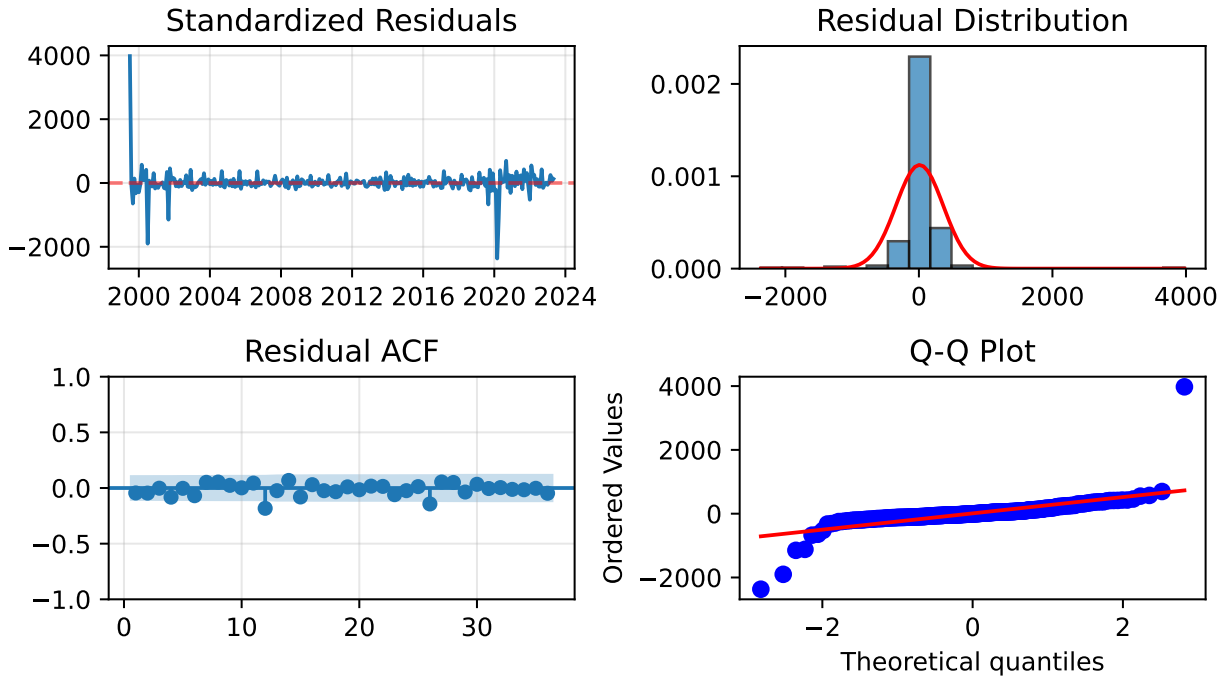


Figure 3: Diagnostic plots for SARIMA(1,1,0)(0,1,1)₁₂ model

The Ljung-Box test fails to reject the white noise hypothesis at both lag 12 ($p = 0.150$) and lag 24 ($p = 0.526$), suggesting that residual autocorrelation is not statistically significant (“Statsmodels: Statistical Modeling and Econometrics in Python” 2025). This indicates the model adequately captures the temporal dependencies in the training data. However, the residual ACF shows minor

Table 3: Traffic recovery summary and forecast horizon statistics

Metric	Value
2019 Peak	5742K
COVID Trough	139K (-97.6%)
May 2023 (Start of Test)	4374K (76.2%)
Nov 2025 Forecast	4327K
Forecast Recovery %	75.3%
CI Width Expansion	+666%

spikes at some seasonal lags, and the Q-Q plot reveals heavy-tailed, non-normal residuals driven by the asymmetric pandemic shock. While the model passes white noise diagnostics, the non-normality of residuals suggests that forecast intervals may underestimate true uncertainty, particularly in the presence of the structural break.

5.3 Forecast Performance

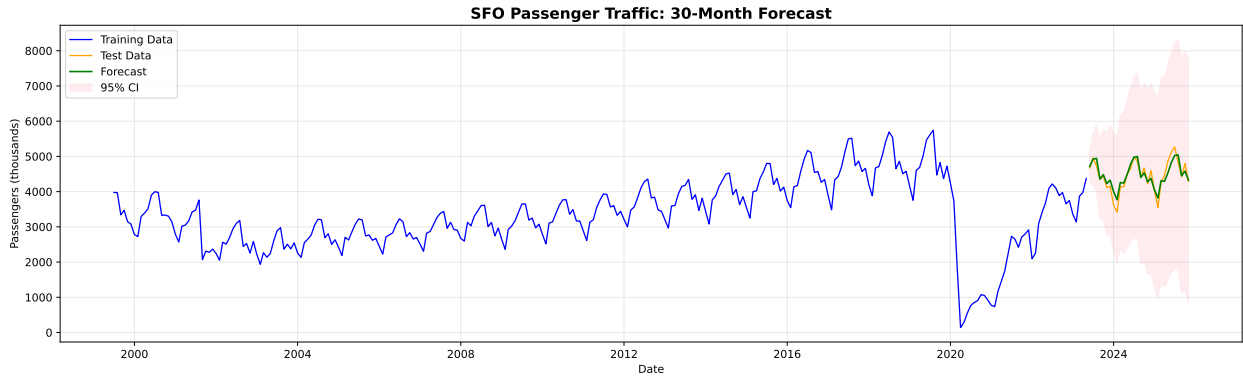


Figure 4: 30-month ahead forecast with 95% confidence intervals

Metric	Value
MAE (thousands)	138.57
RMSE (thousands)	177.89
MAPE (%)	3.21

The model achieves $\text{MAPE} = 3.21\%$, indicating strong short-term forecast accuracy. This low error rate demonstrates the model’s capability to capture the post-pandemic recovery trajectory over the 30-month test horizon.

By November 2025, the forecast projects traffic reaching 75.3% of 2019 levels. However, forecast uncertainty expands dramatically: the 95% confidence interval width increases by approximately

666% over the 30-month horizon. This substantial growth in uncertainty renders long-term point forecasts increasingly unreliable, reflecting the challenges of forecasting across a structural break.

6 Comparison with Previous Literature

This analysis builds upon Project 2 from Winter 2025 (“Pre-Covid u.s. Airline Traffic Analysis” 2025), which examined U.S. airline traffic from 2003 to 2019, noting its strong seasonal patterns and AR(2) signature leading to SARIMA modeling recommendations. However, Project 2 was solely descriptive without forecasting capabilities. Our study extends this work by implementing SARIMA(1, 1, 0)(0, 1, 1)₁₂ modeling on post-COVID data. We find that seasonal characteristics of air travel persist despite the pandemic’s impact, with both first differencing ($d=1$) and seasonal differencing ($D=1$) required to achieve stationarity.

The selected model uses AR(1) for short-term persistence and seasonal MA(1) for year-over-year correlation, achieving strong point forecast accuracy with MAPE = 3.21%. The model passes Ljung-Box diagnostics for residual autocorrelation, though non-normal residuals suggest forecast intervals may underestimate uncertainty. Forecast uncertainty expands by 666% over the 30-month horizon. Addressing prior peer reviews’ concerns regarding the lack of forecasting in Project 2, we demonstrate that while historical seasonal patterns endure under extreme conditions, trend dynamics do not. This indicates that SARIMA models can achieve reasonable short-term airport forecasting after structural breaks, though confidence intervals widen substantially with forecast horizon.

7 Conclusion

This study evaluated monthly passenger traffic at SFO from July 1999 to November 2025, focusing on the forecasting capabilities of SARIMA models post-COVID-19. The pandemic drastically reduced passenger volume by 97.6%, though seasonal patterns remained resilient. The SARIMA(1, 1, 0)(0, 1, 1)₁₂ model exhibited strong point forecast accuracy with MAPE = 3.21% and passed Ljung-Box diagnostics for residual autocorrelation, though non-normal residuals indicate potential underestimation of forecast uncertainty.

By November 2025, traffic is projected to reach 75.3% of 2019 levels, with 95% confidence intervals widening substantially (666% expansion). Extending previous STATS 531 work on pre-COVID airline traffic, we confirm that seasonal structure survived the pandemic but trend did not. While SARIMA models demonstrate good short-term forecasting performance, expanding uncertainty limits their reliability beyond one year. The study concludes that SFO is unlikely to return to 2019 traffic levels before 2027, highlighting that understanding the “new normal” requires recognizing fundamental shifts in travel patterns beyond historical trends.

In answer to our original questions: SFO’s passenger traffic exhibits stable seasonal patterns that survived the pandemic, but the trend regime has permanently shifted. SARIMA models provide reliable short-term forecasts (MAPE = 3.21%), but uncertainty grows substantially with forecast horizon, limiting long-term predictive value.

Acknowledgments

Data obtained from the San Francisco Open Data Portal. Analysis performed using Python with statsmodels, pandas, and matplotlib. AI tools were used for code debugging. All other work was completed by the group members.

Bibliography

- “Air Traffic Passenger Statistics.” 2025. San Francisco Government Open Data Portal. 2025. https://data.sfgov.org/Transportation/Air-Traffic-Passenger-Statistics/rkru-6vcg/about_data.
- “Pre-Covid u.s. Airline Traffic Analysis.” 2025. STATS 531 Winter 2025 Midterm Projects. 2025. https://ionides.github.io/531w25/midterm_project/project02/blinded.html.
- “STATS 531 Lecture Notes: Model Selection and AIC.” 2025. University of Michigan, Department of Statistics. 2025. <https://github.com/ionides/531w26/blob/main/05/notes.pdf>.
- “Statsmodels: Statistical Modeling and Econometrics in Python.” 2025. statsmodels.org. 2025. <https://www.statsmodels.org/stable/index.html>.