Proceedings of CONF-BPS 2026 Symposium: GenAI, Labour Markets, and the Economics of Human and Financial Capital
DOI: 10.54254/2754-1169/2026.LD31678

# *Comparative Forecasting of International Air Passenger Demand: Exponential Smoothing vs. ARIMA Models*

**Yuhuen Fei**

*Sage Hill School, Newport Beach, USA*
*fionafeiqwq@gmail.com*

***Abstract.*** Airlines and airports use estimates of the future volume of international flight activity to plan their capacity, prices, routes, and future investments. While there is broad consensus that these forecasts are important, the challenge arises due to the difficulty in developing reliable forecast estimates. Passenger loads exhibit great seasonality and volatility from month to month making long-term forecasts much more complicated than initially perceived.In light of these issues, this research compared two traditional forecasting methodologies (exponential smoothing and ARIMA) utilizing data derived from the U.S. Bureau of Transportation Statistics T-100 International Market Data set (2010-2019) and was able to jointly analyze long-term growth and multiplicative seasonal factors. The forecast was validated using rolling-origin tests at 1, 3, and 6-months ahead as well as a 2-year holdout for evaluating the long-horizon forecast. The criteria used to assess forecast accuracy included the following measures: RMSE, MAE, MAPE, and MASE. The findings from this research are mixed, at 1, 3 and 6 month ahead forecasts both ARIMA and exponential smoothing produced very similar forecasts with neither method exhibiting superiority over the other when the forecasts are updated frequently. However, for longer-range forecasts this gap widens significantly as ARIMA forecasts drift and accumulate forecast error while exponential smoothing is more stable and achieves higher forecast accuracy with the 2-year holdout. The general conclusion drawn from this research is that what model to use when developing forecasts is driven by what level of forecasting is required. Operational forecasting at shorter time frames can be conducted with either method while longer-term forecasting is better accomplished with exponential smoothing method. This may seem intuitive; however, the evidence presented here supports the claim that the model selection based on the forecast requirements is of a greater significance than is typically reflected in most forecasting research.

***Keywords:*** International Air Passenger Demand, Exponential Smoothing, ARIMA Models, Forecast Horizon, Model Comparison

DOI: 10.54254/2754-1169/2026.LD31678

# 1. Introduction

## 1.1. Research context and industry significance

Civil aviation has evolved over the years from a supporting role to becoming one of the main aspects to move about the world for mobility and trade and aid development. The pandemic has accelerated many of the changes have been seen in the society today. For example, compared to 2023, worldwide passenger revenue kilometres are expected to increase by 10.4% in 2024 and surpass pre-COVID-19 levels by 3.8%. At the same time, the global load factor is projected to reach an all-time high of 83.5% [1]. While those figures may seem abstract, they highlight that both airlines and airports are not just returning to their previous status but have begun to expand into new areas. In essence, the overall size of the new demand represents a significant increase in the overall macroeconomic weight of the industry. However, because the rates of growth and the distribution will not be equal across every area of the world, the increases in both passenger traffic and overall capacity will steer true to macroeconomic theory and imply increases in all areas of the economy.

Long before the pandemic hit, economists were already discussing how increased access to air travel could fundamentally alter the spatial pattern of economic activity [2]. This was accomplished by creating new, more efficient, and more effective routes of connectivity, which then raised the level of productivity associated with that travel. By way of example, take a for instance, a small to medium-size coastal city gaining a new set of reliable international routes. With those new connections, that city would likely experience an influx of business conferences, foreign direct investment, and possibly even niche tourism markets. Although the causal relationship between these interrelated phenomena is rarely linear, the resulting spillover effects often lead to growth within a community. For example, as the utility of a new international air travel route became apparent, airline demand for that route would increase, thereby leading airlines to add additional capacity to that market. Now, for airline demand managers (the individuals in charge of forecasting future demand), that means they can utilize reliable and validated information to forecast airline demand. Airline managers also use this same information indirectly, through government and regulatory channels, when formulating future policies related to airline service delivery and other government regulation [3].

While it is true this demand information will provide substantial access for future aviation planning needs, there are still several issues concerning the actual data required for those forecasts (via the Air Traffic Analysis) and, in turn, how to interpret and distribute that data for future aviation use. To start, the demand for air travel services fluctuates during the course of the calendar year for a multitude of reasons that are only indirectly related to economy. Tourists may visit a tourist destination more frequently when there is an active tourist season, while other tourist peaks exist during traditional academic semesters (fall and spring) and institutional holidays. Worldwide airport analyses show pronounced capacity and flow seasonal patterns among many airports [4]. The semi-continuous short-term fluctuations characterizing air travel demand will exist superimposed on longer-term trends such as changing route networks and more extensive structural adjustment processes, meaning that airline demand patterns can at times appear erratic [3].

In general terms, forecasting airline demand is complicated. Airline demand managers must consider a host of factors: macroeconomic variables, changes in airline route accessibility and competition, and the influence of seasonal patterns and time-varying demand patterns, just to name a few. Company and government personnel who receive the airline demand information are expected to utilize this information in conjunction with existing regulations to create appropriate demand-based responses to changes in demand and supply. Thus, both regulatory agencies and airlines must

DOI: 10.54254/2754-1169/2026.LD31678

continue to use various forecasting methods and improve their own methods in order to remain competitive and efficient.

The simultaneous presence of strong demand growth, seasonal patterns, and structural adjustments creates an almost insurmountable puzzle for accurate forecasting. However, they create simulation possibilities as well, so it is relevant to continue exploring whether the classic statistical methodology for forecasting (e.g., exponential smoothing, autoregressive integrated moving average [ARIMA], etc.) is still adequate for industrial and governmental clients. While comparisons of specific methodologies such as exponential smoothing and ARIMA continue to show validity to both industry practitioners and policymakers, primarily due to their transparency and operationally sound nature, differences in the performance of these methodologies under different forecasting scenarios further support their continued usage. Based on these circumstances, a review of the current time-series forecasting methodologies employed for aviation demand is appropriate and will allow for better-informed decisions by both the aviation industry and government as these policies continue to evolve.

### 1.2. Literature review: forecasting methods and applications

Forecasting data that combine seasonality and trends has placed the Holt-Winters or exponential smoothing models as index of reliance by forecasters on statistical products. Exponential smoothing is used by forecasters to represent many different products such as; electric load, grocery stores, and transportation [5]. The primary value of the exponential smoothing model is in its ability to take a time series and break it into level, trend, and seasonal components. Each of these components can be updated using their unique smoothing parameter. Therefore, a forecaster can determine whether seasonality is represented as an additive constant or as a multiplicative factor which scales in proportion to the level in the time series [6]. The Gardner review has given exponential smoothing a solid statistical footing through a single error state-space approach to understanding exponential smoothing within the framework of providing probabilistic forecasts and a basis for making inferences based on information criteria [5]. Holt-Winters methods give more than simple recipes to produce daily use forecasts, they provide with coherent frameworks from which to expand the description of seasonal variations in time series. The Gardner framework allows for a more complex catalogue of trend related components such as; damped trends, and multiplicative trends which cannot easily fit into the linear ARIMA framework. Empirically, Wiguna et al. have studied the multiple monthly rainfall records which show a strong cyclical seasonal component, and, through their study of the multiplicative variant versus the additive variant of Holt-Winters, have demonstrated that the multiplicative variant produces better forecasts than the additive variant [6]. Their research is relatively modest but an excellent example of the continued relevance of Holt-Winters as a forecast option for highly seasonal time series models while more complex options develop.

A second stream of literature focuses on autoregressive integrated moving average (ARIMA) models and the Box–Jenkins methodology. Makridakis and Hibon review how Box–Jenkins popularised ARIMA by providing a systematic workflow—achieving stationarity via differencing and transformations, identifying model orders from ACF/PACF, estimating parameters, and checking white-noise residuals—which made it possible to represent the autocorrelation structure of many real-world series in a compact linear form [7]. However, their empirical analysis also shows that out-of-sample forecast accuracy is highly sensitive to how the "I" component is handled: inappropriate differencing and seasonal specification can make Box–Jenkins models no better, or even worse, than much simpler benchmarks [7]. Dickey’s work on stationarity formalises this

DOI: 10.54254/2754-1169/2026.LD31678

concern by defining stationarity in terms of constant mean, variance, and autocorrelation, and by introducing unit-root tests to decide whether differencing is required [8]. He further demonstrates that mis-differencing (under- or over-differencing) distorts the ACF and PACF and leads to misspecified ARIMA models with unreliable parameters and forecasts [8].

On the applied side, seasonal ARIMA (SARIMA) models have been widely used for strongly seasonal monthly or daily data. For example, Ebhuoma et al. fit a SARIMA model to monthly malaria incidence in KwaZulu-Natal and show that, after appropriate seasonal differencing, a parsimonious SARIMA(0,1,1)(0,1,1)12_{12}12 specification yields white-noise residuals and forecasts that closely track out-of-sample counts, with the lowest AIC/BIC among competing models [9]. Chuwang and Chen apply Box–Jenkins and SARIMA models to daily and weekly urban rail passenger flows, explicitly combining unit-root tests, information criteria (AIC, BIC), and error metrics (MAE, RMSE) to select the best specification [10]. Their results confirm that ARIMA-type models perform particularly well when demand exhibits inertia, stable autocorrelation, and clear seasonal patterns, while more complex machine-learning or hybrid models are only needed when nonlinearities and irregular behaviours dominate [10]. Taken together, this literature positions ARIMA/SARIMA as a natural baseline for this study of international air passenger volumes, especially given their strong trend, seasonality, and serial dependence.

Recent comparative studies further suggest that the relative forecasting performance of ARIMA and exponential smoothing depends strongly on the length of historical data and the intended forecasting horizon. Using COVID-19 case data, Bahuguna et al. compare ARIMA and exponential smoothing models across datasets of different temporal spans and report a clear horizon-dependent pattern: ARIMA achieves lower forecast errors when fitted to recent short-term data, whereas the Holt–Winters exponential smoothing model performs more accurately when longer historical series are used [11]. This empirical evidence indicates that ARIMA may be better suited for short-term forecasting driven by recent dynamics, while exponential smoothing benefits from its explicit representation of trend and seasonality over extended periods.

Related findings are reported in applied economic forecasting. Mujiyanto et al. show that although hybrid ARIMA–LSTM models can outperform classical methods in highly volatile commodity price data, ARIMA and exponential smoothing remain competitive baselines when simplicity, interpretability, and computational efficiency are prioritised [12]. Together, these studies highlight that no single model dominates universally and motivate a focused empirical comparison of ARIMA and exponential smoothing in structured, strongly seasonal demand data, such as international air passenger volumes [11,12]. Despite this extensive literature on forecasting methods, several critical gaps remain in the understanding of their relative performance, specifically for international air passenger demand forecasting.

### 1.3. Research gap and objectives

Researchers have performed extensive analytical work comparing classical methods for conducting time-series forecasting, including Exponential Smoothing and Autoregressive Integrated Moving Average (ARIMA), over a variety of settings. The methods studied span the spectrum from epidemic surveillance to the pricing of commodities to the flow of traffic to the admissions of patients into hospitals for public health planning [9-12]. The consistent finding from these studies is that neither method has proven to be superior to the other. Instead, the performance of either method is dependent on several factors related to the characteristics of the data, including seasonality, volatility, and forecast duration. Models that have short forecast durations tend to favour predictive results from models that use an updating mechanism, whereas long forecast durations reveal

DOI: 10.54254/2754-1169/2026.LD31678

weaknesses in the way trends and seasonal components are factored into the models. However, the majority of these studies have focused on measuring impacts of infectious disease epidemiology and/or regional economic indicators; therefore, it raises an inherent issue of geographic scope. It should be noted that air transportation is another logical choice for similar analyses considering that air transportation is specifically linked to global patterns in human mobility and economic activity; nevertheless, there is a surprisingly limited body of literature on systematic forecasting of international passenger volume transactions employing such classical forecasting techniques. The air travel industry is an unusual combination of two economic trends that typically exhibit a degree of combination—long-term economic growth coupled with a high seasonality (as opposed to financial markets or historic epidemic events). These five characteristics provide an excellent opportunity to test how well the ARIMA and Exponential Smoothing forecasting models perform under both the rigors of an organized system and the constraints of the demand for air travel. There are only a few studies available in the literature that longitudinally compare the performance of ARIMA and Exponential Smoothing regarding the forecasting of international air travel demand, specifically those studies that incorporate explicitly seasonal time-series data. The purpose of this study is to fill this void and analyse both forecasting methodologies to better determine which method is more appropriate for forecasting international air travelling demand.

### 1.4. Analytical framework and methodology

The Analysis uses an analytical forecasting framework that is relatively structured in comparison to the two methods (Exponential Smoothing Methodology and ARIMA) used to model international air passenger demand. The analysis starts with a simple overview of the international air passenger demand data and focuses on how the data appears as time progresses through a long-term trend, seasonal ups and downs, and other structural changes like gradual shifts in levels and/or amplitudes.

This step sets the foundation for all of the modelling decisions made later in the analysis based upon what the data indicates, rather than what the individual modeller prefers based on theory. In doing so, the modeller is able to identify what models are appropriate for estimating international air passenger demand based on the data characteristics, rather than simply by the modeller's own beliefs.

In following this exploratory look at the data, the next step is to implement a seasonal and trend component using the Exponential Smoothing Method (ETS). In doing so, the modeller identifies a model to fit the data that has a multiplicative seasonal structure, rather than a constant seasonal structure. The ETS model framework allows for both a seasonal and trend structure to be included in a single equation, therefore allowing for greater flexibility, the inclusion of a multiplicative seasonal structure, and better predictive capabilities than either method could produce separately. Additionally, in parallel, the ARIMA approach is developed to model serial dependency in the data. The modelling cycle follows the standard identification, estimation, and diagnostic checks.

Finally, all models were run through out-of-sample forecasting. Subsequently, the different models were evaluated against widely used error metrics to provide a relatively straightforward basis for the comparison of prediction performances across methodology types. Collectively, the analysis enables assessment for the capability of both ETS and ARIMA methods for forecasting international air passenger demand, as well as providing solid empirical evidence supporting the gaps identified in previous research.

DOI: 10.54254/2754-1169/2026.LD31678

## 2. Method

### 2.1. Secondary data source and variable construction

Primary and secondary data will be used to create an initial dataset of International airline passengers. The quantitative information needed for the analysis will be obtained through the United States Bureau of Transportation Statistics. This analysis will rely upon the T-100 database provided by the U.S. Bureau of Transportation Statistics. In addition, the T-100 International Market (All Carriers) table contains electronically supplied records for international air travel by carrier and that record is further disaggregated by international flight segment. The airline passenger count data from these two sources will be aggregated together on a monthly basis to create a monthly series that contains the total number of International airline passengers. Given that the data provided in these databases are supplied monthly, the resulting dependent variable contains a seasonal frequency of twelve. A total number of 120 monthly records will be generated by using a time period from January 2010 to December 2019 (10 years of data). A ten-year time period is sufficient for the results of this analysis to capture the reoccurring seasonal patterns and gradual changes in trend, but not so long as to be unhampered by structural breaks as a result of the COVID-19 pandemic. Additionally, by only relying on the data prior to the COVID-19 pandemic, the results of this analysis should be easier to interpret; hence the number of airline passengers will likely have been influenced more by changes in airline policies resulting from the COVID-19 pandemic rather than the underlying demand for passengers.

### 2.2. Data preparation and exploratory analysis

Data from Trans Stats are most often downloaded in yearly sections and, therefore, an important part of the analysis will include concatenating the sections extracted from Trans Stats prior to the final step of aggregating the records. While concatenating and finalising the aggregation of all records, a variety of data cleaning processes will be conducted. These processes include implementing a standard Year-Month index for all records' Date field: checking for duplicate primary keys caused by concatenating each year's records, ensuring the identification of all months are represented within the final time period, identifying missing months within the resulting dataset, and inserting a record to represent all missing months, if any.

After assessing the results of the data preparation, fundamental assessment of the time-series data will also be performed. Specifically, graphical representations of the time-series will be used to assess the seasonal and trend structures of the series. Additional statistical assessments of the time-series include examination of the autoregressive behaviour of each time-series record. One of the principal issues will be whether there is a positive effect of the magnitude of the seasonal peaks/troughs versus the overall level of the series. Therefore, the findings will indicate multiplicative seasonality; therefore, when designing the models and determining which forecast modelling approach to use, allowance may be made for the multiplicative nature of the seasonal behaviour, as indicated by the exploratory findings.

### 2.3. Forecasting models

The research will implement two types of classical forecasting models. Exponential Smoothing (ETS) will be used to formulate the model from the results of the exploratory analysis. ETS will be utilised to allow both trend and seasonal components and multiplicative seasonality where indicated

Proceedings of CONF-BPS 2026 Symposium: GenAI, Labour Markets, and the Economics of Human and Financial Capital
DOI: 10.54254/2754-1169/2026.LD31678

by the exploratory findings. In addition to the baseline ETS models, Box-Jenkins modelling techniques (ARIMA and SARIMA) will also be utilised in the research study. All time-series records will be transformed with log transformations and differencing where necessary to achieve approximate stationarity. Model orders for both the ETS and ARIMA models (and their trend components) will be determined by applying standard methodologies and utilizing various forms of information to create the orders using the established methods.

### 2.4. Evaluation design and error metrics

As part of the evaluation design, all forecasting models will be assessed based on an out-of-sample approach using a Rolling Origin forecasting design. The out-of-sample evaluation will reflect the methodologies used by operational organisations in practice to forecast demand for their products. The data will be divided into an initial training window for estimating models for a limited time frame followed by a timeframe of holding out observations. Forecasting errors will be evaluated based upon various different evaluation designs — Rolling One-Step-Ahead, Rolling-Multi-Step (h=3 & 6), and a Single Long-Horizon Evaluation (h=24). Additionally, a variety of summary measures will be applied (e.g., RMSE, MAE, MAPE, and MASE) to support a comprehensive evaluation of forecasting performance.

### 2.5. Software and reproducibility

Finally, all analyses will be developed using RStudio and the methodologies employed in this analysis will rely on existing time-series R packages (e.g., forecast, fable, and tsibble). The entire workflow, from importing and aggregating the data to model fitting and rolling evaluation, will be provided as fully reproducible scripts. The utilisation of reproducible scripts will allow researchers to maintain consistency between the descriptions of the methodologies used, the reported findings, and the code used to produce the findings. It also allows the potential for other researchers to replicate or further enhance the work done in this analysis.

## 3. Empirical results: forecasting performance and comparative analysis

### 3.1. Forecasting performance across horizons

#### 3.1.1. Descriptive analysis of international passenger demand

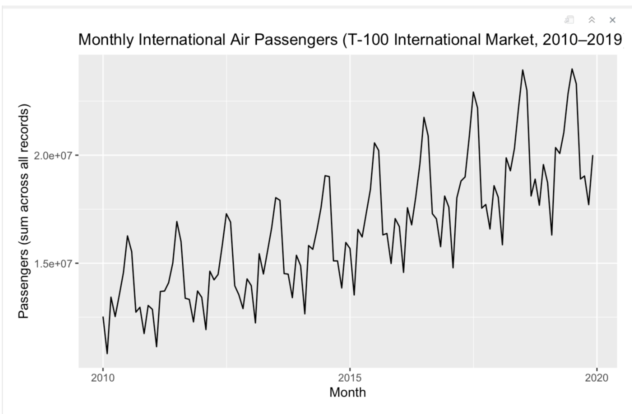


Figure 1. Monthly air passengers (Photo credit: Origin)

DOI: 10.54254/2754-1169/2026.LD31678

Figure 1 displays international passenger volumes on a month-by-month basis between 2010 and 2019 using the T-100 International Market dataset. From first appearing to show consistent growth over time, the graph exhibits an overall increase in the level of international passenger travel. The volumes recorded at the end of the decade were significantly greater than those recorded at the start of the decade; however, this was achieved through steady increases, rather than sudden jumps or spikes.

Superimposed over this gradual upward trajectory, there exists a feature that reflects seasonality, which occurs on a recurring annual cycle with similar peaks and valleys. For example, the months of June through August typically represented the highest number of passengers traveling internationally while the months of January through February and late December were typically the months with the lowest number of international passengers. Those familiar with international travel can relate to this typical seasonal trend, as airports and international terminals fill during the summer vacations and then empty out again when children return to school. Additionally, the seasonal effect is greater in absolute terms as the overall level of international passenger volume increases, but relative to the overall volume, the seasonality does not change drastically.

Decomposing the international passenger volume series into its various components has produced evidence supporting this perception of orderly movement, not random fluctuations. Approximately 34.9 percent of the total variance in the international passenger volume series is due to seasonality; furthermore, there were no significant changes in levels or breaks in the series during the sample period. In summary, the data are demonstrating a surprisingly high degree of regularity and predictability for an actual transport passenger series. Therefore, forecasting methods such as Exponential Smoothing (ETS) and Seasonal ARIMA (SARIMA) will likely yield accurate results for international passenger volume in this context.

Table 1. Forecast accuracy of ETS and ARIMA/SARIMA models under alternative evaluation designs

| Design | Model | RMSE | MAE | MAPE | MASE |
|---|---|---|---|---|---|
| Single holdout (24-step once) | ETS | 382459.8 | 323993.8 | 1.601994 | 0.4191809 |
| Single holdout (24-step once) | ARIMA/SARIMA | 620118.3 | 472441.4 | 2.357460 | 0.6112414 |
| Rolling-origin (1-step) | ETS | 396714.2 | 323946.7 | 1.609830 | 0.4191200 |
| Rolling-origin (1-step) | ARIMA/SARIMA | 420945.3 | 331876.4 | 1.663999 | 0.4293794 |
| Rolling-origin (3-step) | ETS | 431575.5 | 360156.6 | 1.786715 | 0.4659680 |
| Rolling-origin (3-step) | ARIMA/SARIMA | 475231.1 | 366995.6 | 1.822362 | 0.4748163 |
| Rolling-origin (6-step) | ETS | 470817.4 | 387829.5 | 1.917332 | 0.5017711 |
| Rolling-origin (6-step) | ARIMA/SARIMA | 506083.9 | 404124.0 | 2.019353 | 0.5228527 |

DOI: 10.54254/2754-1169/2026.LD31678

### 3.1.2. Short-horizon forecasting performance (h=1)

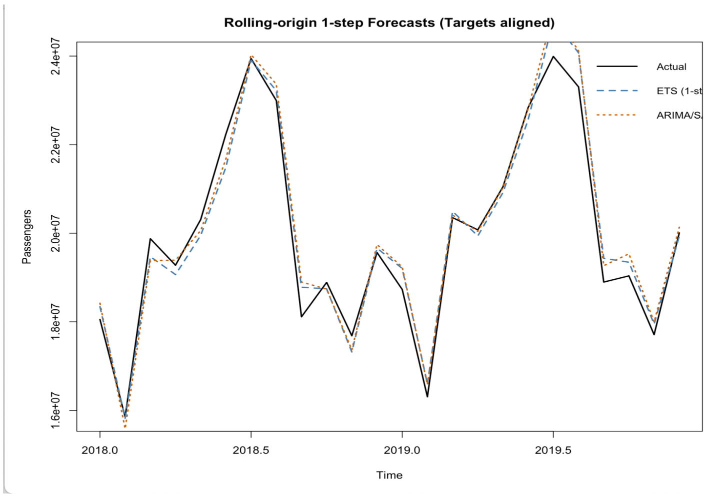


Figure 2. Rolling-origin one-step-ahead forecasting (Photo credit: Origin)

Figure 2 shows that forecasts produced by ETS and ARIMA/SARIMA (horizon = 1 step), plotted against the actual number of passengers for the holdout period of 2018 to 2019, are very similar. The figures are comparable for 12-month forecast intervals and demonstrate the two forecasting approaches producing a similar level of accuracy and performance (Table 1). At this time frame, it appears that both forecasting methods have similar patterns of results relative to observed data. Both approaches provide accurate estimates of the monthly peaks and valleys of international passenger traffic. It is difficult to identify any point in time where one method performs much worse than the other, and comparisons of forecast errors (where available) reveal the number of significant errors is small compared to the overall volume of international passenger traffic.

The information in Table 1 consolidates what was seen visually above. The ETS method produced slightly lower RMSE and MAE values than ARIMA/SARIMA, and again, the scale- and model-free accuracy measures (MAPE and MASE) produced similar results. While differences between models do exist, the magnitude of those differences is not great enough to support the identification of one method as being superior to the other. Therefore, when viewed from a practical perspective, both methods are capable of providing accurate one-step forecasts for international passenger traffic when applied to rolling-design data sets.

As a result, when considering a forecasting time frame of less than a year and a predictive model that is periodically updated, both ETS and ARIMA/SARIMA are expected to provide similar predictive accuracy. When comparing the performance of the two models, ETS was found to be marginally better across most of the metrics presented. Based on the evaluation of short-term operational international passenger traffic forecasting through these two methods, either could work well based on this type of evaluation.

DOI: 10.54254/2754-1169/2026.LD31678

### 3.1.3. Medium-horizon forecasting performance (h=3,6)

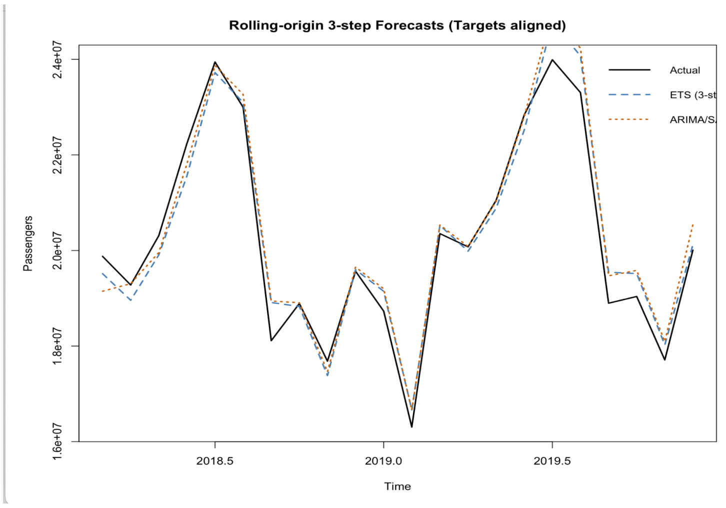


Figure 3. Rolling-origin 3-step forecasts (Photo credit: Origin)

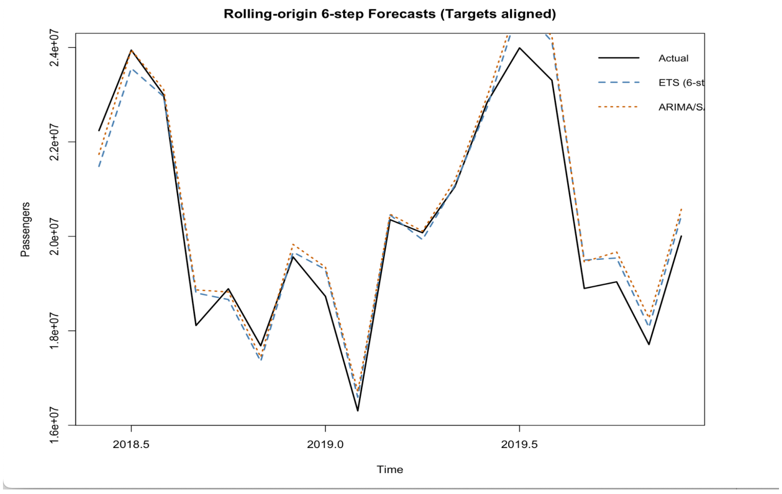


Figure 4. Rolling-origin 6-step forecasts (Photo credit: Origin)

In Figures 3 and 4, it can be seen how the three-step-ahead rolling and six-step-ahead rolling forecasts were performed for both ETS and ARIMA/SARIMA with respect to each other. This information is supported by the accuracy measures in Table 1. All measures of forecast error (RMSE, MAE, MAPE, and MASE) increase as the forecast length increases; this is true for both forecasting methods.

With respect to broad seasonal patterns (i.e., peaks falling on schedule and troughs occurring when expected), the three-step forecasts, shown in Figure 3, provide reasonably good estimates of when to expect both peak and low points throughout the year. Although three-step forecasts are generally closer than one-step forecasts to the actual values, the alignment of three-step forecasts with the real data are looser than that of one-step forecasts. Short-term fluctuations are frequently smoothed out, and there is more difficulty aligning turning points with the actual series. ARIMA/SARIMA does appear to drift around the peak and trough values of the seasonal highs and lows, suggesting that there may be some degree of horizon sensitivity to how ARIMA/SARIMA handles level adjustments.

DOI: 10.54254/2754-1169/2026.LD31678

The differences between ETS and ARIMA/SARIMA are magnified even further in the six-step forecasting scenario (Figure 4). The six-step forecasts made by ETS remain fairly close to the actual series, despite the longitudinal aspect of six months of extrapolation. Overall, ARIMA/SARIMA shows greater variability in the differences compared to ETS during rapid seasonal transitions when the series is changing direction. This requires a greater effort to create forecast estimates. The greater difference in RMSE and MAE between the ETS and ARIMA/SARIMA for the three-step forecast and the six-step forecast is also represented consistently in Table 1, where ETS provides lower RMSE and MAE than ARIMA/SARIMA for both levels of forecast.

The overall results of the rolling origin/multi-step forecasts provide supporting evidence that both the ETS and ARIMA/SARIMA forecasts are less accurate when forecasted further into the future. However, the accuracy level of ETS gradually decreases, while the performance of the ARIMA/SARIMA model sharpens as the forecast horizon lengthens beyond the very shortest time periods.

### 3.1.4. Long-horizon forecasting

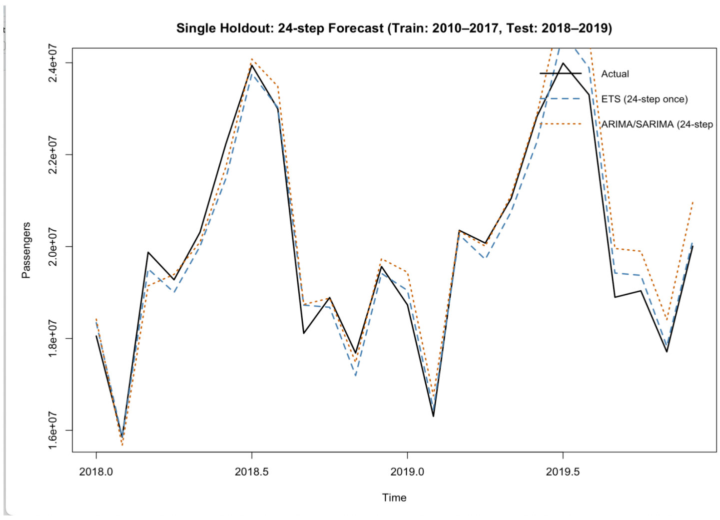


Figure 5. Single holdout long-horizon forecasting (24-step) (Photo credit: Origin)

A single holdout experiment is used to evaluate the long-term performance of both models. The models were estimated once on the 2010-2017 data set and asked to forecast the entire 2018-2019 period, a two-year forecast of 24 months. The forecast accuracy for each model is reported in Table 1, and a graphical comparison of the two forecasts is provided in Figure 5.

The picture differs considerably from what was shown in the rolling-origin performance results. The results of the holdout experiment indicate that the ETS model consistently outperformed the ARIMA/SARIMA model in every measure of forecast accuracy reported in Table 1. The RMSE and MAE values for the ETS model were significantly lower than those for the ARIMA/SARIMA model; the MAPE and MASE scale-free metric values also produce the same conclusion, but the magnitude of the differences is staggering when compared to the one-step, three-step, or six-step rolling-origin performance results, where both models exhibited similar performance levels.

The 24-month forecast trajectory of the ETS model remains close to the actual passenger level and retains much of its seasonal rhythm throughout the forecast horizon, as shown in Figure 5. Although at the end of the forecast period the seasonal amplitude appears to be slightly dampened, the timing of seasonal peaks and troughs in the forecast is sensible, and the overall direction of

DOI: 10.54254/2754-1169/2026.LD31678

travel in the forecast is consistent with historical data for international passengers. In contrast, the forecast trajectory produced by the ARIMA/SARIMA model shows larger and longer-lasting departures from the actual passenger data as the forecast horizon increases. In addition, seasonal patterns of demand produced by the ARIMA/SARIMA model exhibit particularly pronounced deviations from historical seasonal patterns of demand in months of higher demand, indicating that the ARIMA/SARIMA model is unable to extrapolate seasonal patterns of demand reasonably well without regular reestimation.

Based on the comparison between the two models using the holdout method and the comparison of forecast accuracy, the uniqueness of the holdout method for estimating the long-term performance of the two models is sufficient to warrant some level of caution when making short-term forecasts. However, while the performance levels of the ETS and ARIMA/SARIMA models are comparable when forecasts are updated frequently, the difference in performance between the two models when forecasting two years ahead demonstrates that the ETS model is expected to produce the most reliable forecasts for international passenger demand.

## 3.2. Discussion and implications

### 3.2.1. Horizon-dependent performance deterioration of ARIMA

In Section 3.1, it is noticed that ARIMA/SARIMA's predictive accuracy relies heavily on the length of the forecasting horizon. The one-step rolling origin forecast was quite similar in accuracy to ETS, while the accuracy was reduced progressively over the three- and six-step horizons (Table 1 displays this trend, as do the RMSE, MAE, MAPE, and MASE metrics).

In fact, the contrast is largest in the 24-step single holdout test. The number of errors for ARIMA/SARIMA was well above those for ETS and the widening gap was much larger than anything experienced in the rolling origin forecasts. Figure 5 clearly shows the progression of the deviations of the ARIMA/SARIMA forecasts from the actual values during 2018–2019, and particularly large divergences occurred at seasonal peaks where there is a rapid increase in demand for travel.

These data demonstrate that ARIMA/SARIMA appears to be particularly sensitive to forecasting horizon length. One possible cause for this result is that the process of differencing, which is present in ARIMA models, tends to increase the variance of the forecasts as the forecasting horizon increases. For the short-range forecasts, the variance is generally less problematic because of the frequent re-estimation of the model. But when an ARIMA model forecasts 24 months into the future without any intermediate updates, the amount of compounding uncertainty over that period may be overwhelming. With a highly seasonal nature of international air travel data, the question must be raised that while ARIMA/SARIMA models have been found to be adequate for making short-term operational forecasts, they may not be a wise choice for long-term demand forecasting.

### 3.2.2. Evaluation design sensitivity and practical implications

Another aspect that became evident in these results was how sensitive the relative performance of ETS and ARIMA/SARIMA was to the design of the evaluation. Rolling-origin one step ahead forecasts showed both models to be almost equal, with ETS only slightly ahead (Section 3.1.2; Table 1). Three-step and six-step horizons also had a moderate gap, with both methods managing to adequately track the observed time series (Figures 3 and 4).

DOI: 10.54254/2754-1169/2026.LD31678

The situation was different when the evaluation was conducted using a holdout sample from one single long-horizon forecast. The models were evaluated at once when forecasted off the same long-horizon sample (i.e., 24 steps ahead, the 2018–2019 period), and ETS was markedly ahead of ARIMA/SARIMA on all metrics (Table 1; Section 3.1.4). This disparity makes it more difficult to designate a single optimal model, as the type of evaluation design will dictate the results.

It is also important to recognize that the design of the evaluation framework can have an implication for the practical application of the results. The way in which forecasting operations are structured in real-world situations differ greatly with regard to both the frequency of re-estimation and how far ahead decisions must be made. Therefore, a method that produced good results within a rolling-horizon framework might not do as well with regards to a longer-term plan when a model is not frequently re-estimated. Using one evaluation framework to arrive at conclusions may lead to clean results on the surface, but be inappropriate for the real-world contexts in which forecasting results are delivered.

## 4. Discussion

### 4.1. Core interpretative insights

#### 4.1.1. Synthesis and interpretation

The results show that there is an exponential smoothing (ETS) and ARIMA-based agreement in that ETS and ARIMA are very similar when forecasting international air passenger demand at different time periods. ETS and ARIMA exhibit very little difference in the very near future, particularly for rolling one-step-ahead evaluations where the RMSE and MAE between both models are generally less than 5%. This indicates that it is possible to use both ETS and ARIMA for performing operational forecasting on a regular basis, using updated information as it becomes available. As expected, because the most recent movements of passenger demand are most important for determining future patterns versus historical demand data, ARIMA and ETS are both able to generate competing forecasts without difficulty.

Once began to stretch out the forecast duration beyond immediate levels, things begin to change. For example, under a rolling multi-step evaluation of the accuracy of ETS and ARIMA, ETS continues to remain stable as the forecasts for both ETS and ARIMA increase. The ARIMA model produces a much larger error accumulation than the ETS model (the difference becomes obvious when looked at the results of the single 24-step holding experiment). In the holding experiment, ETS consistently outperformed ARIMA as indicated by all of the calculated metrics, causing the rankings of models to change based on the forecast duration. While ARIMA remains a viable model for short-term forecasting, ETS is clearly more advantageous for forecasting patterns of seasonality in aviation demand.

#### 4.1.2. Contextualization within existing literature

The patterns identified here agree with much of the growing literature on the comparative evaluation of classical time-series models, particularly those related to forecasting passenger traffic on air transportation networks. For example, while ARIMA tends to provide superior forecasting results when fitted using the most recent information over a short period of time, Holt-Winters-type models have been shown to provide superior forecasting results when fitted using longer periods of time [11]. The results of the current study support the conclusion that Holt-Winters-type forecasting

DOI: 10.54254/2754-1169/2026.LD31678

models perform better with longer historical datasets. However, they also illustrate that each environment has different time-based characteristics that can result in different comparative performance ranking outcomes, even though the methodology for developing the forecasts was similar.

There are also a number of other studies that show slightly different conclusions. For example, the work by Chuwang and Chen illustrates that ARIMA-type models provide superior performance when used to forecast future transit demand [10]. The results of this study could be viewed as contradictory to those of Chuwang and Chen's study due to the difference in the forecasting design. The findings from Chuwang and Chen's study were based on short-term forecasts that regularly receive updates, while the work presented here compared a rolling-origin evaluation to a single long-horizon holdout evaluation. The results from the analysis presented here indicate that the same methods of comparing two forecasting models can yield different rankings. This suggests that the comparative conclusions ultimately depend on how the forecasting task was described and what information was included in the planning of the task.

The study makes a valuable contribution to the aviation forecasting literature by comparing ARIMA and ETS models across numerous time horizons and evaluation methods. There are few studies available that have specifically investigated the performance characteristics of classical time-series forecasting models over long-horizon international air passenger demand data containing significant seasonal variations.

#### 4.1.3. Theoretical interpretation

Differentiating between ETS and ARIMA at long-range forecasts largely depends on how the trend and seasonal components are modeled in each framework. Both methodologies build components related to the level, trend, and seasonal trend within their model and then update them iteratively so the model can extrapolate both long-run trends and regular repeating seasonal patterns once it knows the pattern. When the seasonality of the time series data is stable within its data, the ETS framework has a clear advantage in carrying it forward through forecasting.

In contrast, ARIMA develops a path to reach long-range forecasts and differentiates the overall level of the time series, i.e., converts the smooth long-run trend into many short-run 'shock' or changes. The autoregressive and moving average components of the ARIMA model explain how shocks or changes are related to one another temporality. Therefore, while ARIMA performs well in forecasting near-term observations based on the most recent changes, long-range forecasts using the ARIMA model generate forecasts from the summation of multiple small shocks or changes, resulting in an accumulation of variance over the long horizon that may obscure the seasonal pattern of the time series.

Based on this interpretation of ETS and ARIMA as two tools to accomplish different forecasting goals, ETS provides a complementary tool for long-range forecasting in highly seasonal time series environments. Conversely, the ARIMA model has a natural fit in the environment of continuous short-range or near-term observations due to frequent re-estimations of the model based on current conditions.

DOI: 10.54254/2754-1169/2026.LD31678

## 4.2. Practical implications

### 4.2.1. Practical implications for industry practice

The differences between the two approaches can also be seen in practice when making aviation-related operational decisions (e.g., planning fleet swaps, scheduling) in the near term (1-3 months). In this context, both ETS and ARIMA perform similarly (i.e. when assessed via rolling one-step evaluations) and so one does not need to worry about selecting one over the other based on relative performance. One will be able to choose based on what is easiest and/or most convenient for his/her needs.

Longer-term forecasts—those relating to strategic planning (e.g., planning for future airline growth through increased capacity, developing new routes or expanding infrastructure)—have a different set of needs than short-term planning. For these longer-term strategic forecasts, ETS would probably be a stronger and more reliable model based on the fact that it continues to produce consistent levels of accuracy as the forecast horizon increases, whereas ARIMA produces less consistent levels of accuracy as the forecast horizon becomes longer. Finally, there are differences in terms of data maintenance between ETS and ARIMA. To maintain the best level of accuracy, ARIMA requires more frequent re-estimation of model parameters than ETS. Thus, frequent refresh cycles for ARIMA will typically produce greater levels of expense in terms of analytical maintenance costs compared to ETS.

Conversely, ETS models can tolerate longer gaps in between updates than ARIMA without causing a significant loss of accuracy; thus, different planning horizons will affect the choice between the two methods, as will the frequency at which the model parameters will need to be re-estimated and the available resources to maintain the forecasting operation. So, in addition to looking at error rates to make a choice between either the ETS or ARIMA approaches, it should also be decided on how often it is expected to refresh the forecasting models and what types of resources are available to maintain them.

### 4.2.2. Importance of evaluation design

How performance is assessed will be the final lesson to learn from this series. The fact that rolling-origin versus single hold-out methods highlight the importance of framing the evaluation when deciding on the superior models. A model that performs well when repeatedly calibrated on a monthly basis will likely underperform when being asked for a 2-year prediction without updates. Almost all of the experimental studies published have only one evaluation design and are often disconnected from the actual operational context in which they were originally planned. This disconnection will create a clean but erroneous conclusion.

Practitioners and researchers could benefit from ensuring that their evaluation design is aligned with the purpose of the forecasting task. If it is expected to re-evaluate and update frequently, a rolling-origin framework will be appropriate; however, if it is planned to have long-range strategic plans and forecast, then a single hold-out design will be more suitable. In an ideal world, both would be presented together, creating a scenario similar to this. Both approaches present a side-by-side comparison making it obvious which models have the sensitivity to the desired horizon.

DOI: 10.54254/2754-1169/2026.LD31678

## 5. Conclusion

### 5.1. Concluding synthesis

Taken as a complete entity the side-by-side comparison of ETS versus ARIMA gives a conditional view as opposed to a simple ranking. In the short term forecast environment where models are regularly updated, the performance of both methods is essentially similar. When the task requires longer period forecasts which do not include intermediate re-estimating, the stability and accuracy of forecasts using the ETS method far exceeds that of the ARIMA method. Therefore, to state that there is one model that is superior is inaccurate. Which model is best will depend on how far in advance it is prepared for the forecast and how it will be utilized, which is the main purpose of this comparison.

### 5.2. Contributions and implications

There are a number of implications for both practitioners and researchers. The results of the evaluation across the various testing formats should help practitioners understand when to use classical methods to forecast demand for international aviation. The findings are particularly applicable to airlines, airports, and aviation authorities, since medium and long-term planning decisions (fleet size, network development, or capital investment) tend to be more concerned with accuracy than stability. For that level of planning, ETS would be more appropriate than ARIMA. At the operational level, where decisions are made on a monthly or even weekly basis, the fact that ARIMA and ETS are nearly equivalent under rolling one-step forecasts indicates that both methods can support operational activities (e.g., scheduling or tactical capacity adjustments).

On a larger scale, more reliable demand forecasts will decrease the likelihood of excessive overcapacity and/or congestion, as well as allow for the more effective utilisation of transportation infrastructure. In addition, the improved calibration of forecasting tools may lead to more successfully developed aviation networks and their associated tourism sectors. The evaluation methodology outlined in this paper is intentionally clear and transparent and could be readily transferred to other forecasting environments where both seasonal and planning horizons impact demand. The methodology illustrates how much the overall conclusions depend on the structure of the evaluation design and time horizon, moving the field of forecasting closer towards practical applications instead of traditional academic methodology.

### 5.3. Limitations and future research

There are limits to what it can be learnt from the current model. The only source of the historical measures used for comparison between the models is passenger volume; therefore, this keeps the data clean but also restricts the forecasts' ability to anticipate upcoming trends. Exogenous shocks (e.g., negative macroeconomic impacts due to policy changes) will not be apparent in the data until the next year after they occur. Similarly, the modeling technique is naturally narrow because it only includes univariate Exponential Smoothing (ETS) and ARIMA methods to maintain the integrity and comparability. While using univariate methods provides fair comparisons, it limits the ability to predict structural or sudden shifts in passenger traffic; that explains partially why the forecast accuracy declines over long periods.

Future research could take several pathways. First, it would be reasonable to develop hybrid or multivariate extensions of ETS that continue to use the same rigorous methodology as ETS but also

DOI: 10.54254/2754-1169/2026.LD31678

incorporate factors external to the transportation industry, such as income data, governmental policy, and fuel prices; these changes should improve performance over an extended forecast period. Second, combining quantitative forecasts with qualitative insight on the industry gained through interviews and surveys could provide a framework for better understanding the assumptions built into the modeling; this could lead to helping align the modelling with the decision-making processes for which the models were designed to support.

## References

[1] International Air Transport Association. (2025). Global Air Passenger Demand Reaches Record High in 2024. Available online: https: //www.iata.org (accessed 9 December 2025).
[2] Lenaerts, B., Allroggen, F., & Malina, R. (2021). The economic impact of aviation: A review on the role of market access. Journal of Air Transport Management, 91, 102000.
[3] Lundaeva, K. A., Saranin, Z. A., Pospelov, K. N., & Gintciak, A. M. (2024). Demand Forecasting Model for Airline Flights Based on Historical Passenger Flow Data. Applied Sciences, 14(23), 11413.
[4] 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.
[5] Gardner Jr, E. S. (2006). Exponential smoothing: The state of the art—Part II. International journal of forecasting, 22(4), 637-666.
[6] Wiguna, I. K. A. G., Utami, N. L. P. A. C., Parwita, W. G. S., Udayana, I. P. A. E. D., & Sudipa, I. G. I. (2023). Rainfall forecasting using the Holt-Winters exponential smoothing method. Jurnal Info Sains: Informatika dan Sains, 13(01), 15-23.
[7] Makridakis, S., & Hibon, M. (1997). ARMA models and the Box–Jenkins methodology. Journal of forecasting, 16(3), 147-163.
[8] Dickey, D. A. (2025). Stationarity issues in time series models. SAS Users Group International, 30, 19.
[9] Ebhuoma, O., Gebreslasie, M., & Magubane, L. (2018). A seasonal autoregressive integrated moving average (SARIMA) forecasting model to predict monthly malaria cases in KwaZulu-Natal, South Africa. South African medical journal, 108(7).
[10] Chuwang, D. D., & Chen, W. (2022). Forecasting daily and weekly passenger demand for urban rail transit stations based on a time series model approach. Forecasting, 4(4), 904-924.
[11] Bahuguna, A., Uniyal, A., Sharma, N., & Semwal, J. (2023). Comparison of exponential smoothing and ARIMA time series models for forecasting COVID-19 cases: a secondary data analysis. International Journal of Research in Medical Sciences, 11(5), 1727-1734.
[12] Mujiyanto, M., Nurindahsari, S., & Izza, R. N. (2024). Comparison of Linear Regression, ARIMA, Simple Exponential Smoothing, Hybrid ARIMA-LSTM, and EWMA in Forecasting Commodity Prices. Telematika, 17(2), 112-127.