Air Transportation Direct Share Time Series Forecasting: A Hybrid Model

Xufang Zheng * and Peng Wei†
Iowa State University, Ames, Iowa, 50010, US

In modern air transportation, the direct share is the ratio of direct passengers to total passengers on a directional origin and destination (O&D) pair. The forecasting of direct share time series on the O&D level, as part of the detailed demand forecasting, plays a fundamental role in air transportation planning and development. An accurate forecasting of the O&D direct share time series can benefit the air transportation planners, airlines, and airports in multiple ways. Based on the previous analysis, the direct share time series is O&D specific. This research focuses on developing accurate direct share time series forecasting models on O&D markets with different characteristics. Both classical time series models and supervised learning regression models are investigated carefully. A novel hybrid model that combines time series concept and machine learning modeling techniques is proposed, which can provide more accurate forecasting performance and valuable insights into the O&D markets. To automatically select the forecasting model for each O&D pair, we proposed a general modeling framework for direct share time series forecasting. Based on the forecasting performance comparison, the modeling framework can provide promising direct share time series forecasting, which is a reliable replacement for the model used in the Federal Aviation Administration Terminal Area Forecast.

Nomenclature

$directShare_{A\rightarrow B,t}$	=	direct share on O&D market $A \rightarrow B$ in quarter t
$directShare_{A\rightarrow B}$	=	quarterly direct share time series on O&D market $A \rightarrow B$
M_{TAF}	=	the model used for O&D direct share time series forecasting in the FAA TAF
M_{hybrid}	=	the hybrid model for O&D direct share time series forecasting proposed in this research
M_{hybrid_MLR}	=	the hybrid model employing Multiple Linear Regression
M_{hybrid_RF}	=	the hybrid model employing Random Forest

*Ph.D candidate, Department of Aerospace Engineering, Iowa State University, Ames, Iowa, 50010, US, AIAA student member.
†Assistant Professor, Department of Aerospace Engineering, Iowa State University, Ames, Iowa, 50010, US, AIAA senior member

I. Introduction

In modern air transportation, the directional origin and destination airport pair, on which the airlines provide direct and non-direct flight services, is known as O&D pair or O&D market. Comparing to the non-direct flight services, the direct flight services are generally more convenient and time-saving, but more costly at the same time. Passengers make reservations based on their traveling purpose and preference. On a certain O&D, the passengers taking direct flights are direct passengers, otherwise, they are non-direct passengers. O&D direct share is the ratio of direct passengers to total passengers on an O&D pair. The series of direct share observations on a certain O&D pair indexed in the time order is the O&D direct share time series. As an essential component of the detailed demand, it is of great importance for air transportation planners, airlines, and airports to have a better understanding and more accurate forecasting of O&D direct share time series.

A. Definition

Illustrated in Fig.1 are the itineraries on O&D pair BOS (Boston Logan International Airport, Boston, MA) → DFW (Dallas/Fort Worth International Airport, Fort Worth, TX). The passengers flying directly from BOS to DFW (n_1) and the passengers taking the only connect at ORD (O'Hare International Airport, Chicago, IL) without flight change (n_2) are direct passengers. The passengers taking one connect at ORD with flight change (n_3) and the passengers taking more than one connects (n_4) are non-direct passengers. Direct share is the ratio of direct passengers to total passengers on a directional O&D during a period of time. Assuming all the existing itineraries from BOS to DFW are shown in Fig.1, the direct share on BOS → DFW in quarter t can be formulated as Eq. (1), in which $n_{D,t}$ is the number of direct passengers and $n_{T,t}$ is the number of the total passengers in quarter t.

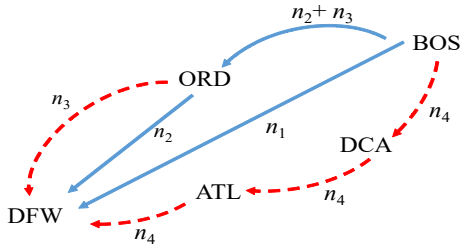

Fig. 1 Illustration of itineraries on O&D pair BOS → DFW

$$directShare_{\mathrm{BOS}\to\mathrm{DFW},t} = \frac{n_{D,t}}{n_{T,t}} = \frac{n_{1,t}+n_{2,t}}{n_{1,t}+n_{2,t}+n_{3,t}+n_{4,t}} \tag{1}$$

We denote the observation of direct share on O&D pair A → B at quarter t as $directShare_{\mathrm{A}\to\mathrm{B},t}$ ($t \in T$). The chronological sequence of $directShare_{\mathrm{A}\to\mathrm{B},t}$ during T is the direct share time series on O&D pair A → B, which can be denoted as $directShare_{\mathrm{A}\to\mathrm{B}}$. Shown in Fig.2 are examples of direct share time series on four directional O&D pairs ANC (Ted Stevens Anchorage International Airport, Anchorage, AK) → SFO (San Francisco International

Airport, San Francisco, CA), ATL (Hartsfield-Jackson Atlanta International Airport, Atlanta, GA) → PHX (Phoenix Sky Harbor International Airport, Phoenix, AZ), AUS (Austin-Bergstrom International Airport, Austin, TX) → BWI (Baltimore/Washington International Thurgood Marshall Airport, Baltimore, MD), and CLT (Charlotte Douglas International Airport, Charlotte, NC) → DAY (Dayton International Airport, Dayton, OH).

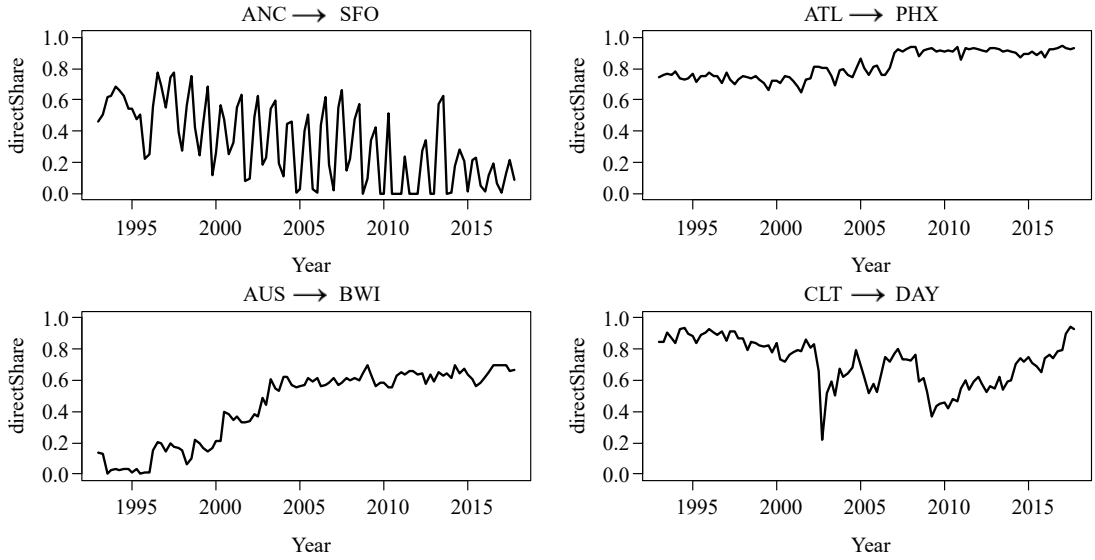

Fig. 2 Direct share time series on different O&D pairs

In Fig.2, the characteristics of the four direct share time series vary significantly from each other. On O&D pair ANC → SFO, there is a distinct seasonality in the direct share time series with an underlying decreasing trend. Comparing to O&D pair ANC → SFO, the direct share time series on ATL → PHX is relatively more stable. The fluctuations of direct share are within a smaller range. On O&D pair AUS → BWI, there are three segments in the direct share time series. Before 1995, the direct share is relatively low and stable. From 1995 to 2005, there is a clear increase in the direct share time series until it reached a higher level in 2005. Comparing to the other three examples, there is more randomness in the direct share time series on CLT → DAY. We can not eyeball any distinct trends or seasonality from the plot.

The direct share time series is O&D specific, which means for different O&Ds the characteristics of direct share time series may vary significantly from each other. For some O&D pairs, there is strong seasonality in the direct share time series. The underlying trend in the direct share time series may differ significantly for different O&D pairs. There may be segments in a direct share time series with different seasonality or trends. The O&D specific characteristic of the direct share time series makes it necessary to develop different models for different O&Ds, which can mostly capture the characteristics of different direct share time series.

B. Motivations and objectives

With the rapid development of commercial air transportation, detailed demand forecasting has been more and more important for air transportation planning and development. As an essential component of the detailed demand forecasting, the forecasting of direct share time series is fundamental for transportation planners' decision making on network planning, airport development, and investment. Direct share forecasting is an significant component in the Federal Aviation Administration (FAA) Terminal Area Forecast (TAF). The FAA TAF is the official forecasting for enplanements, airport operations, and terminal radar approach control facilities operations by the FAA, which is employed in various applications [1]. It provides major guidance for budgeting billions of US dollars in airport planning, runway and facility maintenance, and hiring and training plans for air traffic controllers, etc. The forecasting of direct share is directly used for airport investment decision making on both federal and state levels.

Passengers' itinerary preference is one of the top concerns for airlines. When a passenger making a reservation, several important factors may affect the passenger's decision, which include flight service type, departure and arrival time, and airfare, etc. On a certain O&D market, there are usually both direct and non-direct flight services provided, especially between busy hubs. The passengers may prefer direct flight services because of their convenience, or prefer the non-direct flight services because of the distinct pricing difference between the two types of services. Direct share shows the passengers' general preference for direct flight services on a certain O&D market, which is an essential factor to be considered for airline market strategy making [2]. For O&D pairs on which the passengers have a higher preference for direct flight services, the airlines tend to provide more direct flight services to compete for the passengers at the origin. The strategy has further impacts on airlines' routing plans and network structure. Additional to the existing operation network, airlines also invest and develop new direct markets to improve revenue. It is significant for airlines to have an in-depth insight into the evolution of direct share on the O&D market, which decide the decision making about investment or withdrawal for certain O&D market.

Direct and non-direct passengers require different services at the airports. With one or more connections in an itinerary, there are risks of connection missing and baggage delay. For airport operation, the frequency of direct services to the chosen destination can have an impact on airport utility [3] and a better understanding of the air travelers' preference to direct flight services is crucial for airport capacity planning [4]. More accurate forecasting of direct share can benefit the airports in labor supply scheduling and service optimization, which helps airports in improving service quality and becoming more competitive in multiple airports metropolis regions.

In this research, we carefully study the characteristics of direct share time series on different O&D pairs, and aim to develop accurate direct share time series forecasting models for 1295 different O&Ds across the U.S.

C. Literature review

Demand forecasting has always been fundamental for the air transportation industry. Accurate forecasting of passenger demand is of great importance for air traffic administrations, airlines, and airports. Models were developed to forecast the passenger demand on airport level [5], city pair level [6], and network level [7–9]. Both classical economic methods [10–12] and time series approaches have been explored with various innovations and combinations [5, 13, 14]. With the rapid development in the air transportation industry, detailed passenger demand forecasting is playing an increasingly important role in air traffic planning, airlines and airport operations. The itinerary demand forecasting is the initial exploration of detailed demand forecasting [15–17]. O&D direct share is a significant component in the detailed demand forecasting. However, in most of the previous research and practical applications, the O&D direct share is generally assumed as a constant over time [15–18]. This method is easy to comprehend and implement. However, based on the examples illustrated in Fig.2, the assumption of constant direct share is not held for real air transportation practice, especially for long-term forecasting. To the best of our knowledge, this is the very first in-depth study of O&D direct share time series based on data mining and machine learning techniques. To fully exploit the modeling capabilities of different models, both classical time series models and machine learning models are investigated in this research.

In classical time series forecasting models, there are additive or multiplicative components to model the signal and noise in the time series. Signal indicates any pattern caused by the intrinsic dynamics of the process from which the data is recorded [19]. Two of the most important characteristics of the signals are the trend and seasonality (or periodicity). The noise components may contain the noise from the current process or process far from the present. Exponential Smoothing models and Autoregressive Integration Moving Average (ARIMA) models are the most widely used classical time series forecasting models [20]. Exponential Smoothing models are a class of window-function based models, which can separate the signals and the noise as much as possible in a relatively simple form [21–23]. Holt-Winters Filter is one type of the Exponential Smoothing models, which is specially developed for time series with seasonality or periodicity [24, 25]. ARIMA models are a wide family of models with a large number of variations [26] and applications [27–30]. The variation of ARIMA model with seasonal components is the SARIMA model, which can capture the characteristics of time series with or without seasonality [31, 32].

Classical time series models forecast the target at the current step based on the observations. Comparing to the classical time series models, the supervised learning regression models forecast the target at the current step by the matching feature set.

The supervised learning regression models can automatically extract knowledge of the relation between the response and the features from the data and predict the response in the future [33–35]. Based on whether there are predetermined formulations and a fixed number of parameters for the underlying model, the regression models can be categorized into parametric and non-parametric models [36, 37]. The structure of parametric models is more straightforward compared to the non-parametric models, which makes the parametric models more interpretable. The most widely used parametric

regression model is the Multiple Linear Regression model, which models the relationship between the response and the features in a linear manner [38, 39]. For non-parametric models, there is no predetermined model formulation of the underlying model. Non-parametric models allow a more flexible regression modeling of the response that combines the features in a nonparametric manner [40]. Decision trees is one type of non-parametric model which involves stratifying or segmenting the feature space into several simple regions [41]. The models based on decision trees are tree-based models, which are widely used in regression problems. The wide range of applications shows the promising prediction and forecasting performance of tree-based models [42–45]. Random Forest is a tree-based model, which combines a large number of decision trees to yield a single consensus prediction of the response.

To exploit the modeling capability of different models and develop accurate direct share forecasting models for different O&D pairs, both classical time series models and machine learning models are investigated carefully in this research. The remainder of this paper is organized as follows. In Sec. II, we introduce the data source and data processing in this research. The model development is introduced in detail in Sec. III, which includes the classical time series models and the newly proposed hybrid model. In Sec. IV, we introduce the general modeling framework for direct share time series forecasting model development on 1295 O&Ds across the U.S. In Sec. V, we conclude this research, and the future work is discussed in Sec. IV.

II. Data and Data Processing

To generate the quarterly O&D direct share time series, the publicly available database Airline Origin and Destination Survey (DB1B) is used in this research. The DB1B database is a 10% sample of quarterly airline tickets data reported by air carriers' to the Bureau of Transportation Statistics (BTS), which is available from 1993 [46]. The data are organized on three detail levels by three data tables. The DB1BTicket data table records the data on the round trip level, which includes the total distance flown, total airfare, and the number of coupons of a round trip. The DB1BMarket data table records the data on the directional market level, which includes the distance flown, airfare, and other information about a single trip on an O&D market. The connection information on an itinerary is included in the DB1BMarket data table as well. The DB1BMarket data table is used in this research to generate the O&D direct share time series and other features for machine learning modeling. The B1B1Coupon data records the data on a further detailed level, which is on the coupon level. The B1B1Coupon data is over detailed for this research.

This research focuses on the O&D markets which connect the major commercial hubs across the U.S. Based on the FAA airport category [47], 223 primary hubs which have more than 10,000 passenger boardings each year are included. To guarantee there is enough historical data for modeling, we select the O&D pairs on which direct flight services have been provided since 1993, which are 1295 O&D pairs across the U.S.

In each O&D direct share time series, there are 100 observations from the first quarter in 1993 (1993 Q1) to the fourth quarter in 2017 (2017 Q4). In the modeling process, each time series is split into three subsets. The training set

(1993 Q1 - 2007 Q4, the first 60% of the observations) is used for the coefficient estimation. The second 20% of the observations (2008 Q1 - 2012 Q4) is the validation set, which is used for model hyperparameter tuning and model selection. The testing set contains the last 20% of the observations (2013 Q1 - 2017 Q4), which is used for model forecasting performance measurement. The model training, validation, and testing performance are measured separately on the three data sets.

III. Model Development

As a typical time series forecasting problem, the classical time series models are explored in this research. Both Holt-Winters Filter and SARIMA models are developed for different O&D pairs. To improve the forecasting performance, extra information about the O&D pair is included in modeling. A novel hybrid model that combines the time series concept and machine learning techniques is proposed.

A. Logit and logistic transformation

Since $directShare_{A\to B,t}$ is a random variable between 0 and 1, logit and logistic transformations are necessary for the modeling process to guarantee the forecasting boundaries.

Illustrated in Fig.3 is the modeling process with logit and logistic transformations. Denote an actual observation as y, which is between 0 and 1. y' is the logit transformation of y based on Eq. (2), which is between positive and negative infinity. y' is used as the target in model training to estimate the coefficients. The prediction based on the developed model is denoted as $\hat{y}'$, which is the between positive and negative infinity as well. Logistic transformation is applied to transform the $\hat{y}'$ to $\hat{y}$, which is a ratio between 0 and 1. The equation of logistic transformation is as Eq. (3). The $\hat{y}$ is further compared with y to measure the modeling performance.

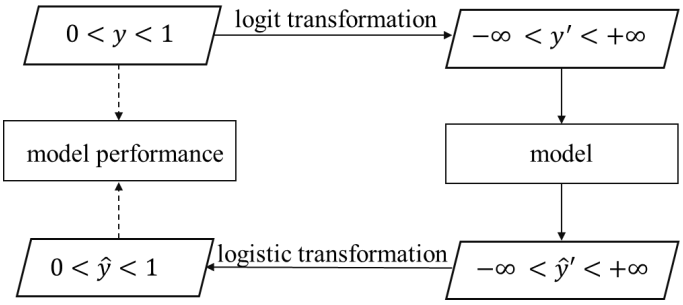

Fig. 3 Modeling process with logit and logistic transformations

$$y' = log(\frac{y}{1-y}) \tag{2}$$

$$\hat{y} = \frac{e^{\hat{y}'}}{1 + e^{\hat{y}'}} \tag{3}$$

In this research, prediction specifically refers to one-step forward prediction (direct share prediction for next quarter), while forecasting specifically refers to twenty-step forward iterative forecasting (direct share iterative forecasting for next 20 quarters). To keep the model performance comparison as concise as possible for model selection, we consistently applied one measurement of model accuracy. The Root of Mean Square Error (RMSE) is a commonly used measurement of model accuracy, especially for numeric regression problems. In this research, the RMSE is used to measure the modeling performance, which is formulated as Eq. (4).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{4}$$

B. Hyperparameter tuning

Hyperparameters are the parameters that determine the architecture of the model. Hyperparameter tuning is the process that searches for the hyperparameters which can optimize the modeling performance. In this research, hyperparameter tuning is employed for the SARIMA model and the hybrid model development.

The most classical and straightforward hyperparameter tuning method is grid searching. All possible combinations of hyperparameters are tried exhaustively, and the best combination is selected based on certain modeling performance metrics. This exhaustive searching method suffers from computational efficiency. Because the size of the tuning grid can increase rapidly with the increase in the number of hyperparameters.

Bayesian optimization (BO) is a reliable and practical alternative for hyperparameter tuning. The most notable advantage of BO lies in its capability of hyperparameter optimizing for black-box functions [48]. The modeling performance is modeled as samples from a Gaussian process in BO, which induces tractable posterior distribution. The information obtained at the current step enables the optimal choices of hyperparameters to try for the next step [49]. In this research, BO is applied for hyperparameter tuning, in which the validation RMSE is used as the modeling performance measurement.

C. Classical time series models

Classical time series models usually consist of addable or multiplicative components of the trend, seasonality, and noise of a time series, which are relatively more interpretable comparing with nonparametric models. In this research, Holt-Winters Filter, which is the Exponential Smoothing model specifically for seasonal time series modeling, and the most widely used SARIMA models are explored for direct share time series forecasting.

1. Holt-Winters Filter

Holt-Winters filter is the Exponential Smoothing model specifically developed for the time series with seasonality or periodicity [50]. Denote the observation at t as y_t, the Holt-Winters Filter with addable components can be formulated as Eq. (5). L_t, T_t, and S_t are the level, trend, and seasonal components respectively [51]. In Eq. (5), m is the term of periodicity, and $\hat{y}_{t+h}$ is the h-steps forward prediction of y at t. α, β, and γ are the weights of the relevant components, which are all between 0 and 1. Greater α, β, and γ indicate the related component at t depends more on recent observations.

$$\begin{aligned} \hat{y}_{t+h} &= L_t + hT_t + S_{t-m+h_m^+} \\ L_t &= \alpha(y_t - S_{t-m}) + (1-\alpha)(L_{t-1} + T_{t-1}) \\ T_t &= \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1} \\ S_t &= \gamma(y_t - L_{t-1} - T_{t-1}) + (1-\gamma)S_{t-m} \end{aligned} \tag{5}$$

To analyze the modeling capability, the Holt-Winters Filters developed for the direct share time series on ANC → SFO and DEN (Denver International Airport, Denver, CO) → TUS (Tucson International Airport, Tucson, AZ) are shown as examples. The coefficient estimations and modeling performance are shown in Table 1. To visualize the modeling performance, the fitted values on the training set, the predictions on the validation set, and the forecasting result on the testing set are plotted together with the actual direct share time series in Fig.4.

Table 1 Model details and modeling performance of Holt-Winters Filter

O&D	α	β	γ	Training RMSE	Validation RMSE	Forecasting RMSE
ANC → SFO	0.0001	0.0001	0.3448	0.1319	0.1077	0.1182
DEN → TUS	0.5015	0.0014	0.0001	0.0669	0.0360	0.0336

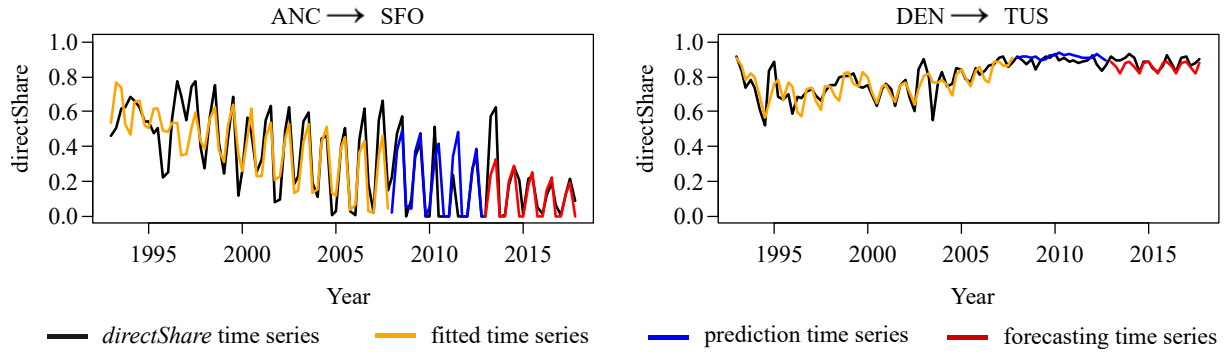

Fig. 4 Holt-Winters Filter modeling performance

In the two direct share time series, there is seasonality with different intensity. Based on the modeling performance,

the Holt-Winters Filter can capture and forecast the seasonality in the time series well. We can see the clear seasonality in the forecast of $directShare_{\text{DEN}\rightarrow\text{TUS}}$, even though the seasonality in the training set is relatively weak. The Holt-Winters Filter model the trend in a linear manner. In the $directShare_{\text{ANC}\rightarrow\text{SFO}}$, there is a consistent decreasing trend over time, which is modeled appropriately by the Holt-Winters Filter.

2. SARIMA model

Seasonal Autoregressive Integration Moving Average (SARIMA) model is a variant of the ARIMA model with seasonal components. The SARIMA model is based on the back-shift operator (B), which makes it capable of modeling the trend of the time series more dynamically. Shown as Eq. (6) is the SARIMA model. y_t is the observation at t, and w_t is the white noise at t. There are six components in a SARIMA model. $\phi(B)$ is the AR(p) component, which is related to the one-step or multiple-steps lagged observations in the time series. The MA(q) component models the impact of current and previous white noise errors on the current observation, which is denoted as $\theta(B)$. ∇^d is the integration component I(d), which is related to the differencing effect on the time series. There are seasonal AR(P), MA(Q) and I(D) in the SARIMA model, which are denoted as $\Phi_P(B^s)$, $\Theta_Q(B^s)$, and ∇_s^D respectively.

$$\Phi_P(B^s)\phi(B)\nabla_s^D\nabla^d y_t = \delta + \Theta_Q(B^s)\theta(B)w_t \tag{6}$$

Seven hyperparameters decide the formation of a SARIMA model together, which are the regular orders (p, d, q), the seasonal orders (P, D, Q), and the period s. Hyperparameter tuning is an essential part of SARIMA model development. The most widely used method is manually selecting the hyperparameters based on the visual observation of the time series plot and the diagnosis graphs. This method is relatively subjective and time-consuming when dealing with a large number of different time series. To automatically select the hyperparameters more efficiently, BO is applied for SARIMA model hyperparameter tuning.

To illustrate and analyze the modeling capability of the SARIMA model, SARIMA models developed for three direct share time series, $directShare_{\text{ATL}\rightarrow\text{PHX}}$, $directShare_{\text{ABQ}\rightarrow\text{MDW}}$*, and $directShare_{\text{IND}\rightarrow\text{CLT}}$† are illustrated as examples. The details about the SARIMA models and the modeling performance are shown in Table 2. There are seasonal components in the SARIMA models for $directShare_{\text{ATL}\rightarrow\text{PHX}}$ and $directShare_{\text{ABQ}\rightarrow\text{MDW}}$. The SARIMA model for $directShare_{\text{IND}\rightarrow\text{CLT}}$ is relatively simpler.

Holt-Winters Filters are developed for the same direct share time series for performance comparison. Shown in Fig.5 is the comparison between the Holt-Winters Filters (on the left) and the SARIMA models (on the right). The SARIMA model is capable of modeling and forecasting both seasonal and non-seasonal direct share time series. The testing RMSEs of the Holt-Winters Filters are 0.0312 (ATL → PHX), 0.0435 (ABQ → MDW), and 0.1227 (IND → CLT)

*ABQ: Albuquerque International Sunport, Albuquerque, NM; MDW: Chicago Midway International Airport, Chicago, IL

†IND: Indianapolis International Airport, Indianapolis, IN

Table 2 Model details and modeling performance of SARIMA model

O&D	SARIMA	Training RMSE	Validation RMSE	Forecasting RMSE
ATL → PHX	$(1, 0, 0)(2, 1, 1)_4$	0.0308	0.0275	0.0240
ABQ → MDW	$(1, 0, 2)(3, 2, 1)_4$	0.0506	0.0424	0.0423
IND → CLT	$(1, 0, 3)(0, 0, 0)$	0.0460	0.0463	0.0404

respectively. Comparing to the Holt-Winters Filters, the SARIMA models can provide better forecasting performance for the direct share forecasting on the three O&D pairs.

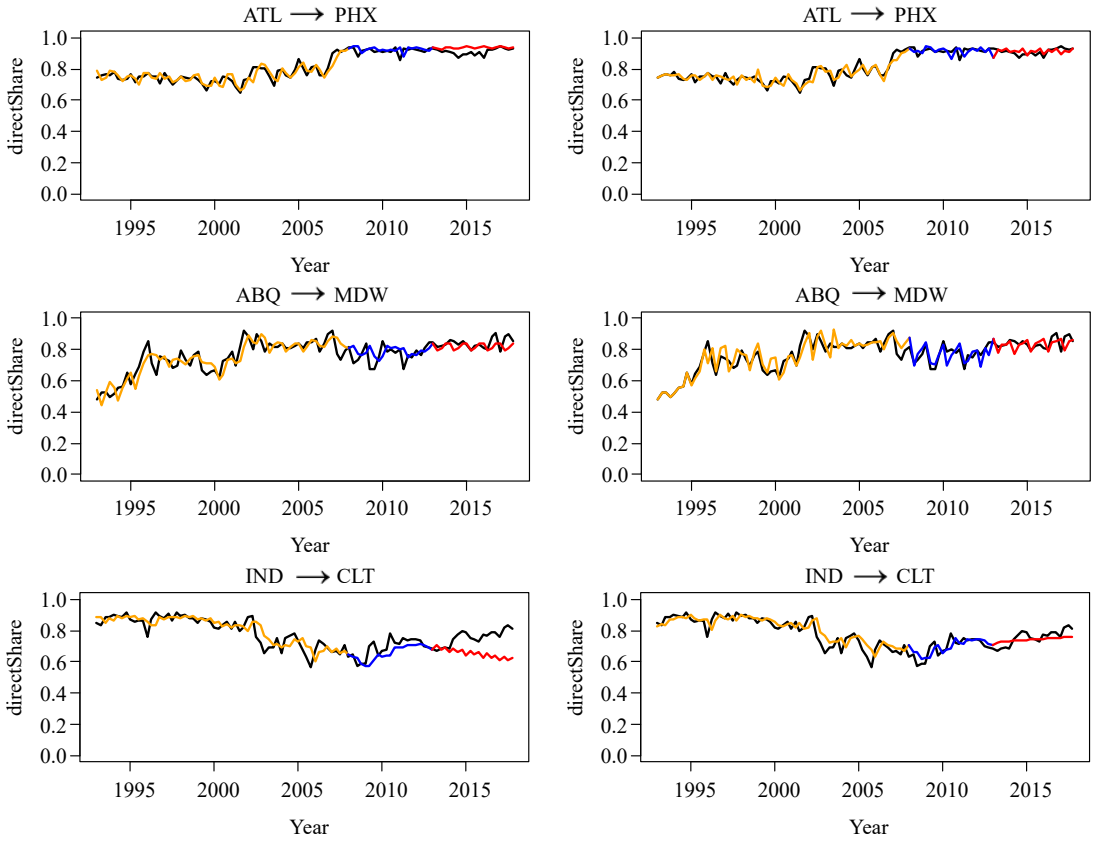

Fig. 5 Modeling performance comparison of Holt-Winters Filter and SARIMA model. (Holt-Winters Filters on the left and SARIMA models on the right)

For the three examples in Fig.5, there is no distinct overall trend in the direct share time series. SARIMA model can provide better forecasting by employing a more dynamic modeling of the trend. In the forecasting of $directShare_{\text{IND}\to\text{CLT}}$ by Holt-Winters Filter, there is a distinct decreasing trend, which is captured in the training set by the model. In the forecasting based on the SARIMA model, because the trend is modeled by the back-shift operation, there is no decreasing trend in the forecasting, which is corresponded to the reality. The SARIMA model is capable of forecasting both seasonal and non-seasonal direct share time series, by which the forecasting depends more on the

recent observations instead of the observations from far from the present.

D. Hybrid model

For some O&D pairs, the direct share time series can be modeled properly by only using the historical observations. However, for other O&D pairs, the direct flight market is highly impacted by certain factors, which makes it difficult to forecast the direct share only based on the historical observations. In this case, we can take advantage of the supervised learning techniques which describe the response using features. In this research, we propose a novel hybrid model that combines the time series concept and machine learning techniques. The proposed model is denoted as M_{hybrid}. Both parametric and non-parametric machine learning techniques are explored.

1. Features and data

Based on the previous research work, there are five features which have significant impacts on O&D direct share, which are the historical $directShare$, $RelativeFare$, $LegacyShare$, $OriginPax$, and $DestPax$ [52]. The historical $directShare$ shows the distribution of direct passengers on a direct flight market, which is the most important feature for direct share forecasting *RelativeFare* is the feature that reflects the pricing difference between the direct and non-direct flight services on an O&D market, which is based on Eq. (7). $LegacyShare$ is a special feature for direct share analysis. It indicates the market share by the legacy carriers, which is as Eq. (8). On an O&D pair, there are usually legacy carriers (e.g. American Airlines, Delta Airlines, and United Airlines, etc.) and low-cost carries (e.g. Southwest Airlines, Frontier Airlines, and JetBlue Airlines, etc.) competing for the customers. O&D market dominated by the low-cost carriers tends to have more direct flight services because of the low-cost carriers' point-to-point operation style. $OriginPax$ is the quarterly total departure passengers from an origin airport, and $DestPax$ is the quarterly total arrival passengers at a destination airport. The two features show the throughput and popularity of the origin and destination airports.

$$RelativeFare_{\mathrm{A}\rightarrow\mathrm{B}} = \frac{mean(Airfare_{\mathrm{A}\rightarrow\mathrm{B},direct})}{mean(Airfare_{\mathrm{A}\rightarrow\mathrm{B},non-direct})} \tag{7}$$

$$LegacyShare_{\mathrm{A}\rightarrow\mathrm{B}} = \frac{\sum Pax_{\mathrm{A}\rightarrow\mathrm{B},\text{carried by legacy airlines}}}{\sum Pax_{\mathrm{A}\rightarrow\mathrm{B},\text{total}}} \tag{8}$$

Air Carrier Statistics (T100) database contains domestic and international airline market and segment data [53]. T100 data bank is used to generate the *LegacyShare*. For other features, the information can be obtained from the DB1B database. Shown in Table 3 is the summary of the extra features.

Table 3 Summary of the extra features

Feature Name	Minimum	Mean	Maximum
$RelativeFare$	0.00	1.01	74.60
$LegacyShare$	0.00	0.33	1.00
$OriginPax$ (thousand)	0.68	141.20	587.78
$DestPax$ (thousand)	0.68	144.17	585.16

2. Hybrid model development

We newly propose a hybrid model (M_{hybrid}) for direct share time series forecasting in this research. The M_{hybrid} combines the time series concept and machine learning techniques by employing a time-ordered feature set in supervised learning regression. Shown in Fig.6 is the methodology of the M_{hybrid} model. For each feature element at t, x_t, there are five features, which are $directShare_t$, $RelativeFare_t$, $LegacyShare_t$, $OriginPax_t$, and $DestPax_t$. The lag length defines the number steps backward the feature elements will be included in modeling. Assuming the lag length is four, the feature elements four steps backward from t are selected in time order as a feature set X_t, which can be formulated as Eq. (9). The supervised learning regression model can be a parametric or non-parametric model, which can describe the direct share at t by the feature set X_t. In this research, the parametric supervised learning model (Multiple Linear Regression) and non-parametric supervised learning model (Random Forest) are explored.

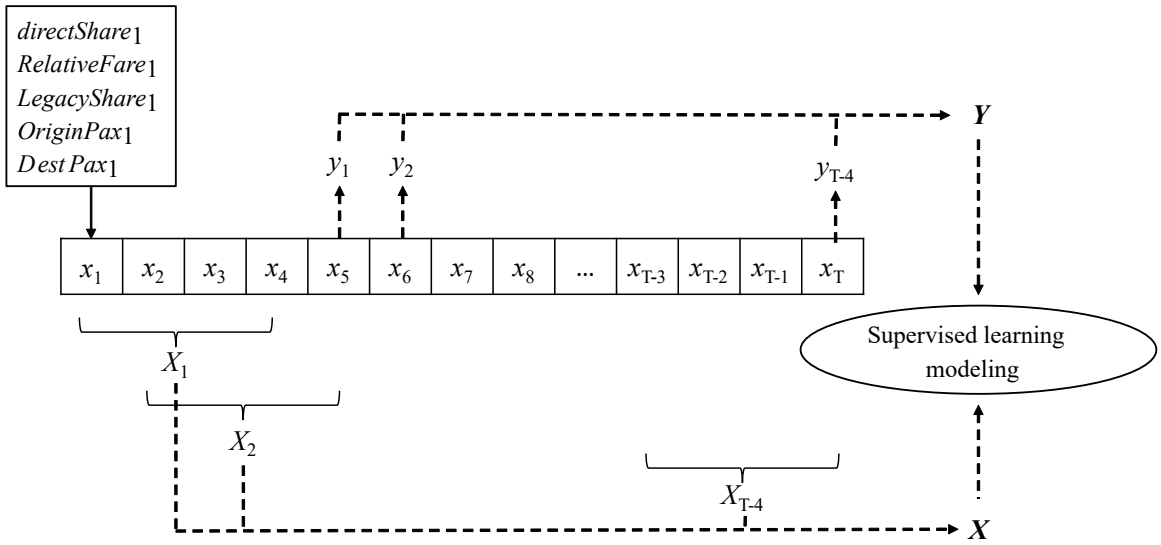

Fig. 6 Methodology of the M_{hybrid} model

$$
\begin{aligned}
X_t &= [x_{t-4}, x_{t-3}, x_{t-2}, x_{t-1}] \\
&= [directShare_{t-4}, RelativeFare_{t-4}, LegacyShare_{t-4}, OriginPax_{t-4}, DestPax_{t-4}, \\
&\quad ..., \\
&\quad directShare_{t-1}, RelativeFare_{t-1}, LegacyShare_{t-1}, OriginPax_{t-1}, DestPax_{t-1}]
\end{aligned} \tag{9}
$$

3. M_{hybrid_MLR} development

The M_{hybrid} model employing Multiple Linear Regression is denoted as M_{hybrid_MLR}. In Multiple Linear Regression, the relation between the response and multiple features is modeled in a linear manner [38, 39]. Comparing to the non-parametric models, the Multiple Linear Regression has a relatively simpler formation and more interpretable. With different lag lengths, different amounts of information will be included in the model. Therefore, the lag length is a hyperparameter to tune for the M_{hybrid} models. Because of data size limitation, the tuning range used in this research is between 1 to 16. The lag length tuning of the M_{hybrid_MLR} for $directShare_{\text{AUS}\rightarrow\text{BWI}}$, $directShare_{\text{CLT}\rightarrow\text{DAY}}$[‡], and $directShare_{\text{DFW}\rightarrow\text{PSP}}$[§] is shown in Fig.7. The lag length is selected when validation RMSE reaches the minimum.

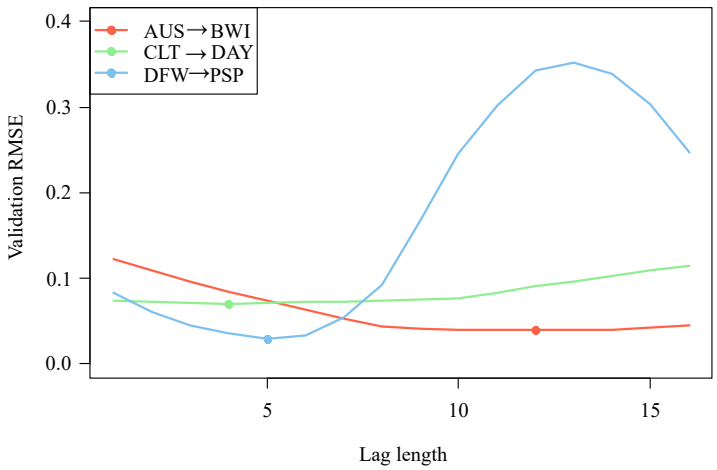

Fig. 7 Lag length tuning for M_{hybrid_MLR} models

Twelve quarters of lagged features are selected in the M_{hybrid_MLR} model for $directShare_{\text{AUS}\rightarrow\text{BWI}}$, comparing to which the features selected for $directShare_{\text{CLT}\rightarrow\text{DAY}}$ (4 quarters of lagged features) and $directShare_{\text{DFW}\rightarrow\text{PSP}}$ (5 quarters of lagged features) are much closer to the present. Shown in Table 4 are the details about the M_{hybrid_MLR} models. Even though the selected lag lengths for $directShare_{\text{CLT}\rightarrow\text{DAY}}$ and $directShare_{\text{DFW}\rightarrow\text{PSP}}$ are similar, the number of non-zero coefficients in the two models differs greatly. It is shown that, for different direct share time series, the amount of information needed is different.

[‡]DAY: Dayton International Airport, Dayton, OH
[§]PSP: Palm Springs International Airport, Palm Springs, CA

Table 4 Modeling details and performance of M_{hybrid_MLR} models

O&D	Lag length	Number of non-zero coefficient	Training RMSE	Validation RMSE	Forecasting RMSE
AUS → BWI	12	16	0.0499	0.0368	0.0404
DFW → PSP	5	13	0.0475	0.0459	0.0690
CLT → DAY	4	5	0.0860	0.0698	0.1047

Shown in Fig.8 is the modeling performance of the M_{hybrid_MLR}. The M_{hybrid_MLR} is capable of forecasting for both seasonal and non-seasonal direct share time series.

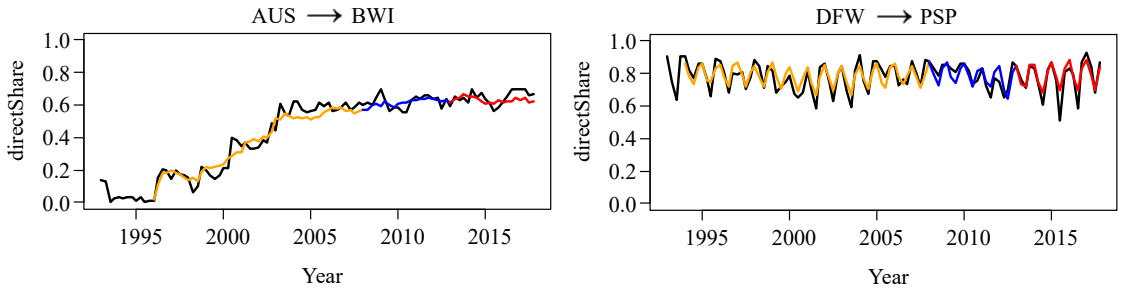

Fig. 8 M_{hybrid_MLR} modeling performance

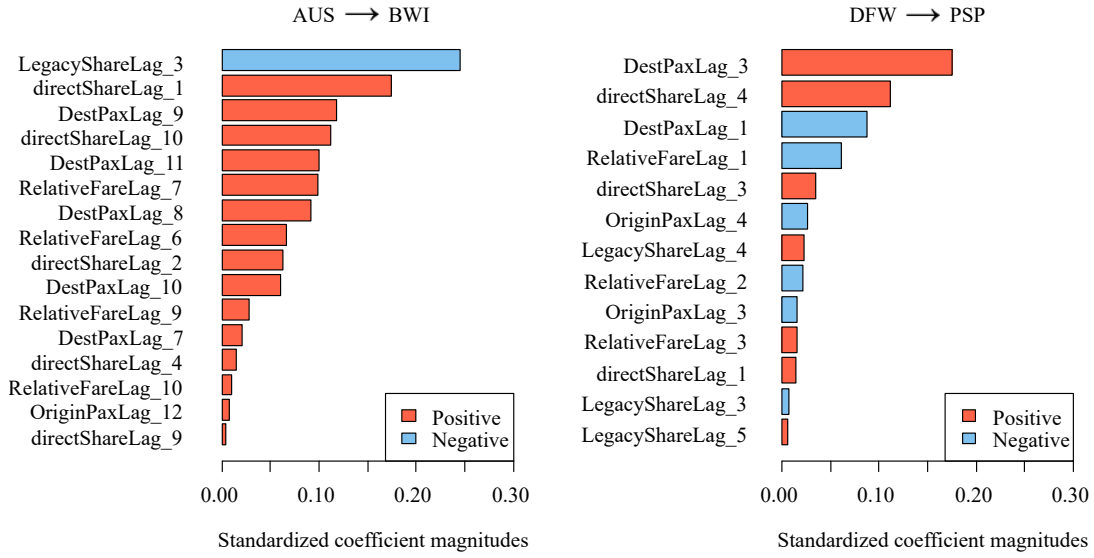

Fig. 9 Feature importance of M_{hybrid_MLR} models

To find the features that have significant impacts on direct share time series forecasting, we created the feature importance plot of the M_{hybrid_MLR}, which are shown as Fig.9. 'Lag_1' refers to the lag of last quarter (quarterly lag),

and 'Lag_4' refers to the lag of four quarters ago (yearly lag).

For $directShare_{\text{AUS}\rightarrow\text{BWI}}$, the most important features include $LegacyShareLag$, $directShareLag$, and $DestPaxLag$, which means the historical legacy share, direct share and the arrival passengers at the destination airport (BWI) are factors that have significant impacts on direct share on AUS → BWI. BWI is one of the three major airports in the Washington D.C. metropolitan area. As shown in Fig.10, after a short period of shrinkage after 2001 (industry fluctuation after 911), BWI has become more and more popular with arrival passengers. With the increasing popularity of BWI, the carriers would prefer to provide more direct flight services to the destination airport (BWI) to compete for the passengers at the origin airport (AUS). Before 1995, American Airlines and US Airways carried most of the passengers on the O&D market AUS → BWI, which provides very few direct flight services. In 1995, the Southwest Airlines joined this O&D market and started providing more and more pricing competitive direct flight services, which brings the direct share on this O&D market to a higher level and became the major carrier on this O&D. The example shows how the important features, such as $LegacyShare$ and $DestPax$, can be used to describe and forecast the direct share time series based on the M_{hybrid_MLR}. M_{hybrid} model can take advantage of feature engineering to reveal the driven factors for O&D direct share.

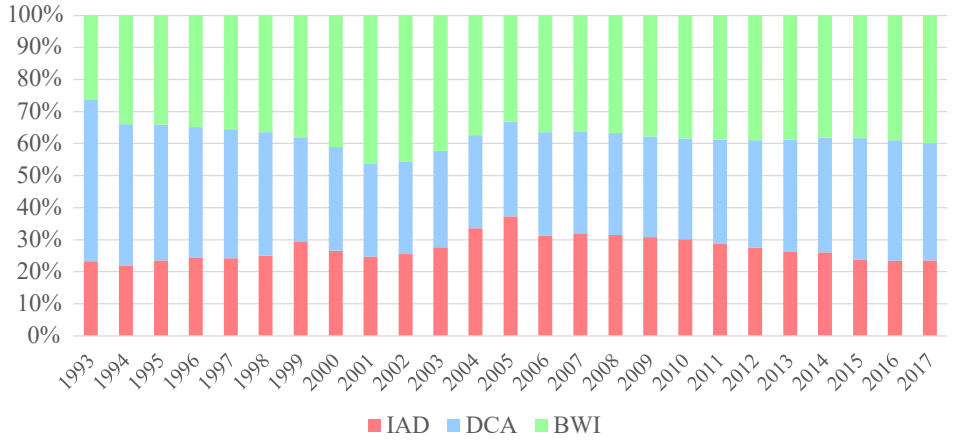

Fig. 10 Passenger share among the three major airports in Washington D.C. metropolitan area

Additional to identify the driven factors to direct share, the M_{hybrid_MLR} can also provide promising forecasting performance for certain O&Ds. For comparison, Holt-Winters Filter and SARIMA model are developed for $directShare_{\text{CLT}\rightarrow\text{DAY}}$ as well. Shown in Table 5 and Fig.11 are the model details and modeling performance comparison of the three models. The Holt-Winters Filter and SARIMA model failed to provide proper forecasting for $directShare_{\text{CLT}\rightarrow\text{DAY}}$. The fundamental reason is that the trend modeling in classical time series models is based on historical observations instead of extra information. However, for the O&D pairs as CLT → DAY, the factors of the direct flight market play significant roles in the direct share change. The M_{hybrid} takes advantage of the feature engineering in machine learning techniques, which uses the related features to describe the O&D direct share.

Table 5 Model details and modeling performance of models developed for $directShare_{\mathbf{CLT\rightarrow DAY}}$

Model name	Parameters	Training RMSE	Validation RMSE	Forecasting RMSE
Holt-Winters filter	α = 0.5514 β = 0.0001 γ = 0.0001	0.0928	0.0735	0.3097
SARIMA	(2,2,2)(0,0,0)	0.0887	0.0635	0.2377
$M_{feature\_MLR}$	lag length = 5 number of features = 6	0.0860	0.0697	0.1031

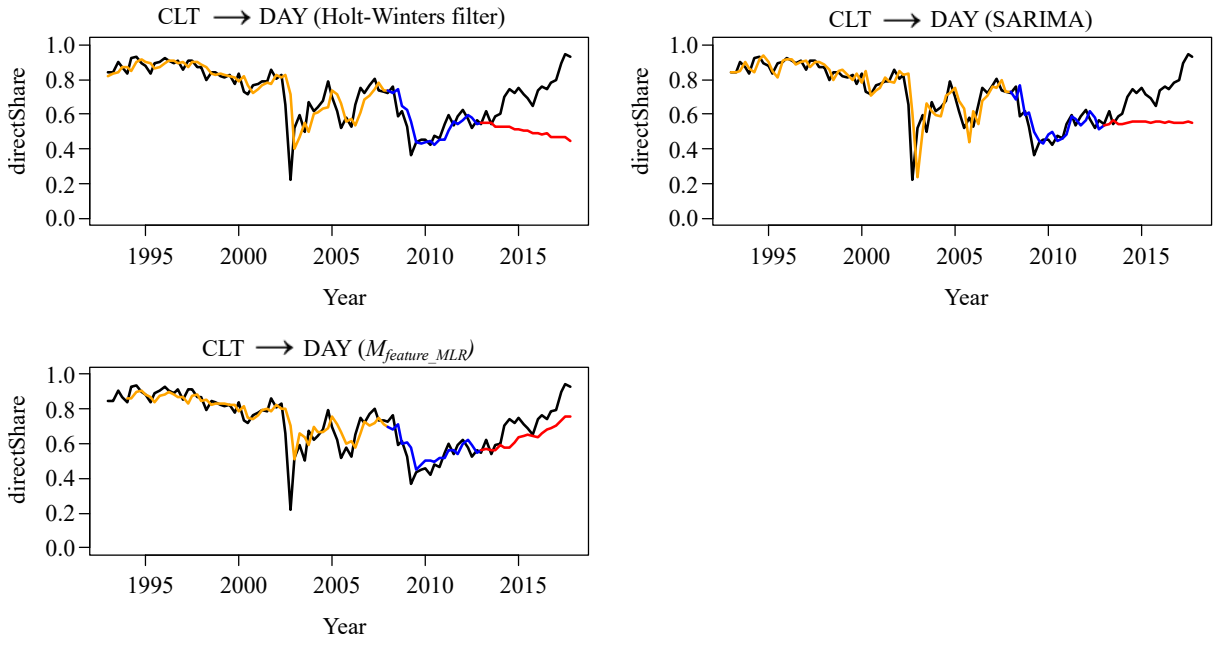

Fig. 11 Modeling performance comparison for $directShare_{\mathbf{CLT\rightarrow DAY}}$

4. M_{hybrid_RF} development

M_{hybrid_RF} is the M_{hybrid} model applying Random Forest as the supervised learning model. Random Forest is a tree-based model, which is widely used in regression and classification problems [54, 55]. Random Forest constructs a multitude of decision trees to yield a single consensus prediction by taking the average. The architecture of a Random Forest model is determined by the hyperparameters, which plays a significant role in modeling performance as well. There are three major hyperparameters for a Random Forest model, which are the *Mtry*, *MaxDepth*, and *Ntrees* [56]. *Mtry* determines the number of features randomly picked at each split when growing the trees. *MaxDepth* determines how deep each tree can grow. *Ntrees* is the number of decision trees grown in a Random Forest model.

Hyperparameter tuning is critical to the Random Forest model development. However, when dealing with a relatively small dataset, tuning hyperparameter excessively may cause a big risk of over-fitting. In this research, the

three hyperparameters are set as *Mtry* = (feature number)/3, *MaxDepth* = 20, and *Ntrees* = 50 as common defaults [57]. Only the lag length which determines the structure of inputs is tuned based on the validation RMSE.

Developed M_{hybrid_RF} models for $directShare_{\text{IAH}\rightarrow\text{SAN}}$[¶] and $directShare_{\text{DTW}\rightarrow\text{PBI}}$[‖] are shown as examples for modeling capability analysis. Shown in Table 6 are model details and modeling performance of the developed M_{hybrid_RF} models. The tuned lag lengths for $directShare_{\text{IAH}\rightarrow\text{SAN}}$ and $directShare_{\text{DTW}\rightarrow\text{PBI}}$ are 16 (16 quarters of lagged features) and 4 (4 quarters of lagged features) respectively.

Table 6 Model details and modeling performance of M_{hybrid_RF} models

O&D	Lag length	Important Features	Training RMSE	Forecasting RMSE
IAH → SAN	16	$directShareLag\_1$ $OriginPaxLag\_2$ $OriginPaxLag\_1$	0.0112	0.0529
DTW → PBI	4	$DestPaxLag\_4$ $directShareLag\_4$ $LegacyShareLag\_4$	0.0162	0.0826

To compare the modeling capability among different models, the Holt-Winters Filters, SARIMA models, and M_{hybrid_MLR} models are developed for the same direct share time series as well. The modeling performance is compared in Fig.12.

For $directShare_{\text{IAH}\rightarrow\text{SAN}}$, there is no distinct overall trend or seasonality that can be eyeballed from the time series plot. The direct share fluctuates in a relatively random manner. The classical time series models failed to forecast $directShare_{\text{IAH}\rightarrow\text{SAN}}$ probably, especially the trend in the testing set. In the developed $M_{feature\_MLR}$, the tuned lag length equals 8, which means 8 quarters of lagged features are included in the developed model. The most important feature in the developed $M_{feature\_MLR}$ is $directShare\_1$, which is significantly more important than other features. The forecasting based on $M_{feature\_MLR}$ mostly depends on the observation or prediction of direct share from the previous quarter, which results in constant forecasting of direct share time series. Moreover, the error at one previous step may be passed on to the following forecasting. The developed $M_{feature\_RF}$ model includes features from the recent 16 quarters, among which the most important features are $OriginPaxLag\_1$, $OriginPaxLag\_2$, and $directShare\_1$. In the developed $M_{feature\_RF}$ model, the departure passenger volume at the origin airport (IAH) has significant impacts on $directShare_{\text{IAH}\rightarrow\text{SAN}}$.

For $directShare_{\text{DTW}\rightarrow\text{PBI}}$, there is a distinct seasonality in the direct share time series. Comparing to $directShare_{\text{ANC}\rightarrow\text{SFO}}$ in Fig.2, there is more randomness in $directShare_{\text{DTW}\rightarrow\text{PBI}}$. The important features in the developed M_{hybrid_RF} model are $DestPaxLag\_4$, $directShareLag\_4$, and $LegacyShareLag\_4$, which are yearly lags of $DestPax$, $directShare$, and $LegacyShare$. In Fig.12, it is clearly shown that the M_{hybrid_RF} model is capable of forecasting the seasonal direct

[¶]IAH: George Bush Intercontinental Airport, Houston, Texas

[‖]DTW: Detroit Metropolitan Wayne County Airport, Detroit, MI; PBI: Palm Beach International Airport, West Palm Beach, FL

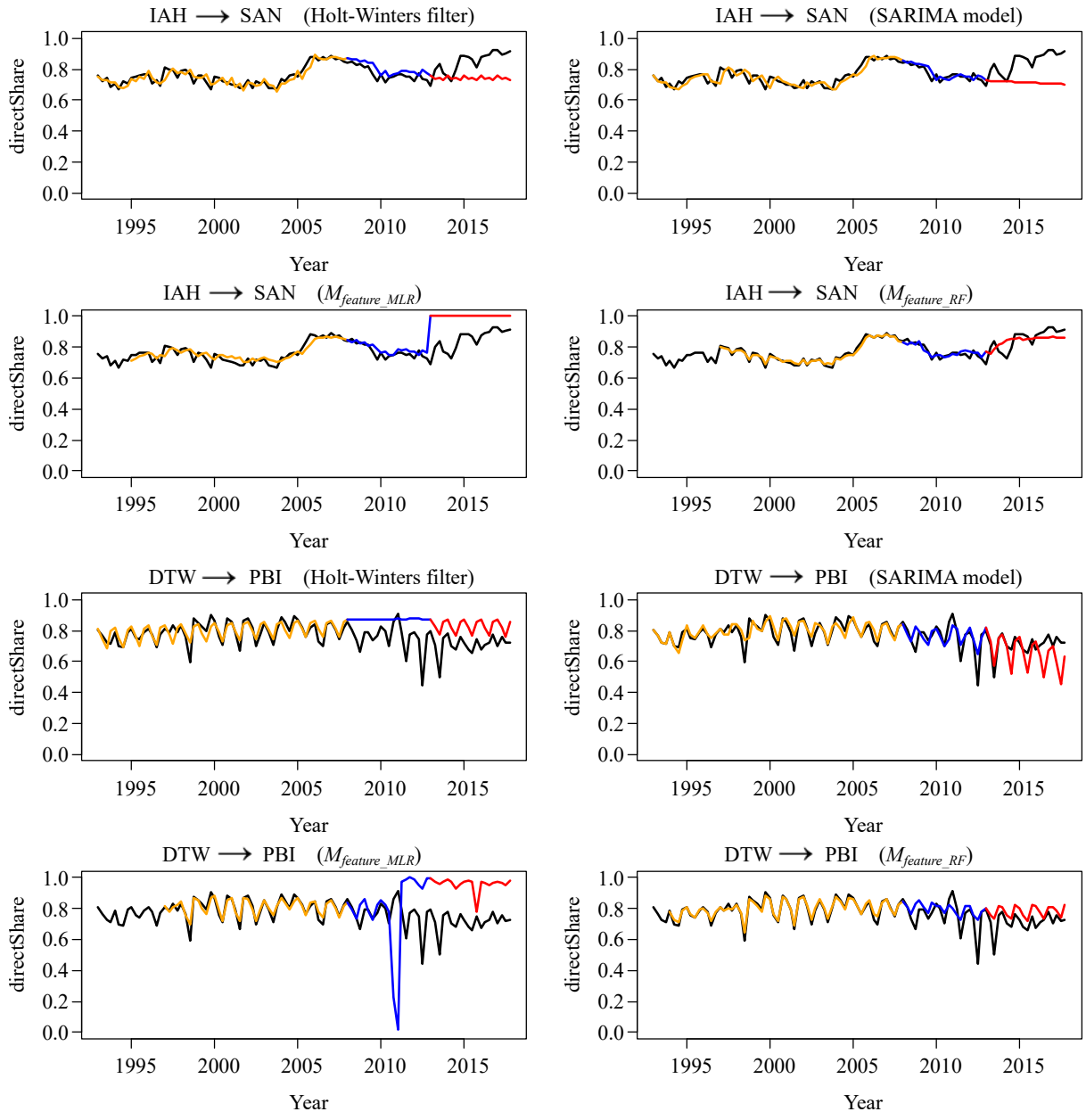

Fig. 12 Modeling performance comparison among Holt-Winters Filter, SARIMA, $M_{feature\_MLR}$, and $M_{feature\_RF}$

share time series. Comparing to the other models developed, the M_{hybrid_RF} is capable of modeling the relationship between response and features in a nonlinear manner. For some direct share time series, the impacts from features are not linear, especially when there are multiple important features in the model. The non-parametric machine learning models can handle the nonlinearity and randomness more properly compared to the linear models. Therefore, the M_{hybrid_RF} is more capable of modeling direct share time series with nonlinearity and randomness compared to the

M_{hybrid_MLR} model.

IV. Direct Share Modeling Framework and Modeling Performance Analysis

O&D direct share time series is O&D specific, which means for different O&D pairs, the characteristics of the trend, seasonality, and noise may vary significantly from each other. Based on the analysis and modeling comparison previous sections, we can safely conclude that for different direct share time series, the model which can provide the best forecasting performance may vary greatly as well. The Holt-Winters Filter can provide proper forecasting of direct share time series with strong seasonality and linear overall trend. The SARIMA model is capable of forecasting both seasonal and nonseasonal time series, by which the trend is modeled more dynamically compared to the Holt-Winters Filter. The newly proposed M_{hybrid} model takes advantage of additional information by feature engineering. The M_{hybrid} model can identify the driven factors to the direct share change on a certain O&D market. When dealing with direct share time series which is greatly impacted by the factors of the O&D market, the M_{hybrid} model can provide more promising forecasting compared to the classical time series models.

In this research, we aim to develop the accurate direct share time series forecasting models for 1295 O&D pairs across the U.S. It is extremely time consuming if we select the best model for each O&D pair manually. It will also be inefficient for future model updating in practice. To automatically select the most proper direct share time series forecasting model for each O&D pair, a general modeling framework is proposed in this research.

A. General modeling framework

The modeling framework is based on both classical time series models and the M_{hybrid} models proposed in this research. Illustrated in Fig.13 is the workflow of the modeling framework. The classical time series models and the M_{hybrid} models are developed individually for each O&D, then the best model is automatically selected based on the prediction performance on the validation set.

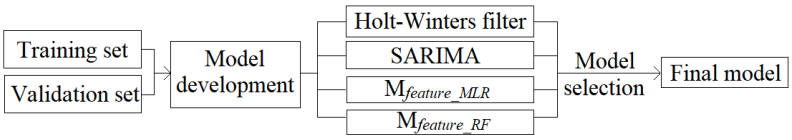

Fig. 13 General modeling framework for direct share time series modeling

The direct share time series forecasting models are developed for the 1295 O&D pairs based on the general modeling framework. Shown in Fig.14 are examples of models developed for different O&D pairs based on the general modeling framework. For direct share time series on different O&Ds, different models are selected. Shown in Table 7 is the model performance summary of the models developed based on the modeling framework. For 72% of the O&D pairs, the classical time series models are selected. The average RMSE of the twenty-step forward forecasting is 0.0898. In

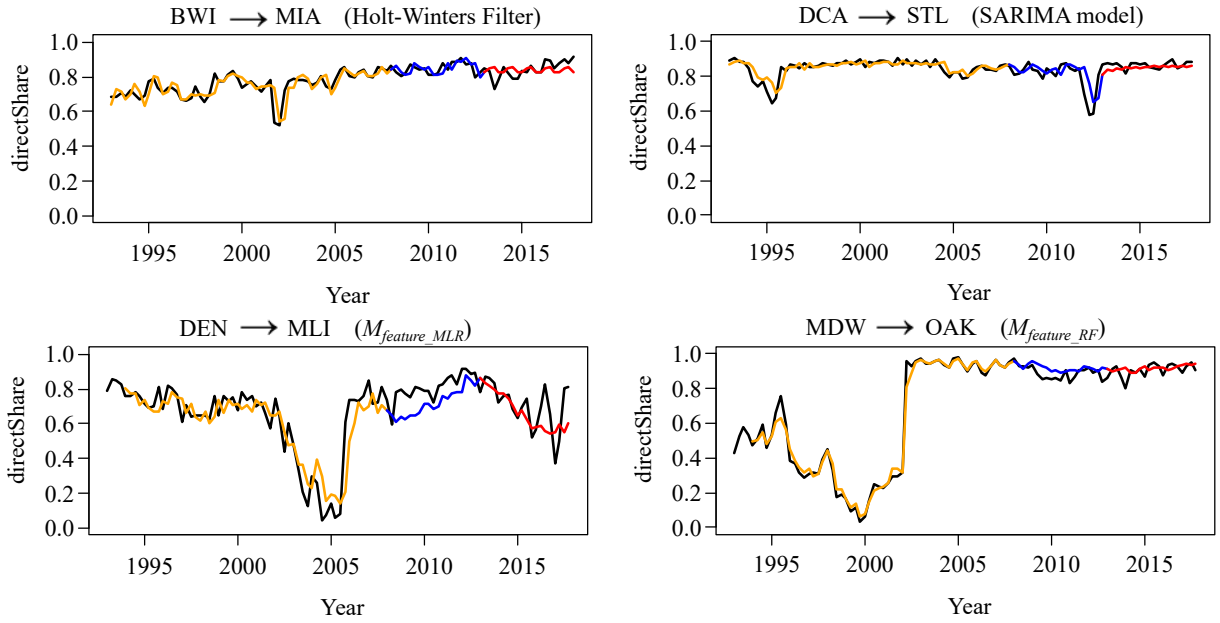

Fig. 14 Modeling performance of models developed based on the modeling framework

practical work, for the O&Ds using the hybrid models, we apply the forecasting of the involved variables from FAA TAF to create the long term forecasting of direct share.

Table 7 Direct share time series modeling based on the modeling framework

Model selected	Num of O&Ds	Average training RMSE	Average forecasting RMSE
Holt-Winters Filter	472 (36.44%)	0.0695	0.0986
SARIMA	461 (35.59%)	0.0750	0.0956
$M_{feature\_MLR}$	153 (11.81%)	0.0442	0.0721
$M_{feature\_RF}$	209 (16.13%)	0.0186	0.0703
Modeling framework	1295	0.0603	0.0898

B. The benchmark: *direct share* forecasting in the FAA TAF

One of the major objectives of this research is to propose a method for direct share time series forecasting which can be a reliable and promising replacement for the model used for direct share time series forecasting in the FAA TAF. The model in the FAA TAF can be denoted as M_{TAF}, in which the forecasting of the O&D direct share time series is a constant same as the latest observation in the historical direct share time series. Shown in Fig.15 are modeling performance of M_{TAF} for $directShare_{\text{CLT}\rightarrow\text{PHX}}$ and $directShare_{\text{ANC}\rightarrow\text{SFO}}$.

The forecasting of direct share time series based on M_{TAF} is constant. When the direct share time series fluctuates within a small range, such as $directShare_{\text{CLT}\rightarrow\text{PHX}}$ in Fig.15, the M_{TAF} can provide the estimation of direct share time series relatively close to the reality. However, when forecasting direct share time series with seasonality and randomness,

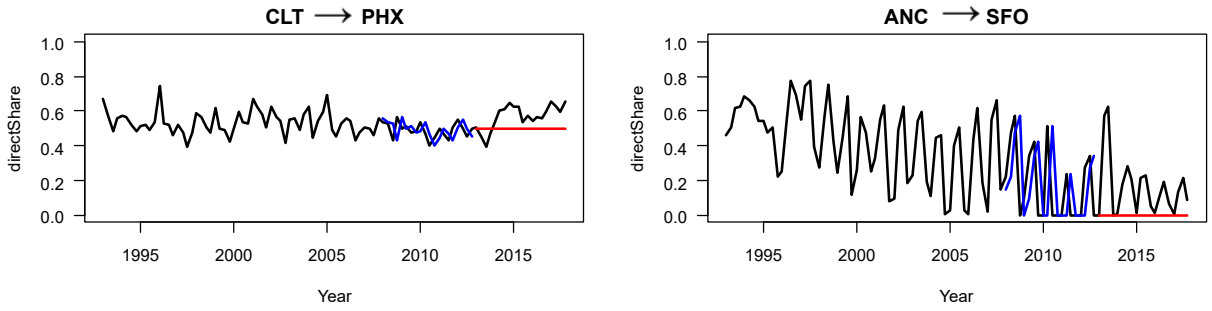

Fig. 15 Prediction and forecasting performance of M_{TAF}

such as $directShare_{\text{ANC}\rightarrow\text{SFO}}$ in Fig.15, the M_{TAF} model is not reliable anymore. The forecasting of direct share time series on 1295 O&D pairs is generated by M_{TAF} for comparison. Shown in Table 8 is the comparison between the M_{TAF} and the general model framework introduced previously. To measure the forecasting accuracy from multiple perspectives, the Mean Absolute Error (MAE) and Root Relative Squared Error (RRSE) are introduced. MAE is an un-scaled measurement of modeling accuracy, which is as Eq. (10). MAE is commonly used to measure the difference between the prediction and actual observations. A smaller MAE shows better modeling performance. RRSE normalizes the total square error of the forecasting by a certain model by dividing the total square error of a simple predictor, which is the average of the observations. A smaller RRSE shows better modeling capability of a developed model [58]. The RRSE equation used in this research is as Eq. (11).

$$MAE = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n} \tag{10}$$

$$RRSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{11}$$

Shown in Table 8 is the modeling performance comparison between the M_{TAF} and the proposed modeling framework for direct share time series forecasting on the 1295 O&D pairs. The average forecasting RMSE, average MAE, and average RRSE are compared, based on which we can safely conclude that the proposed modeling framework can provide better forecasting performance compared to the M_{TAF}.

Table 8 Modeling performance comparison between M_{TAF} and the modeling framework

	M_{TAF}	Modeling framework
Average forecasting RMSE	0.1058	0.0898
Average MAE	0.0897	0.0731
Average RRSE	1.5883	1.3466

Besides the better forecasting performance, the model developed based on the modeling framework can capture the characteristics of the direct share time series on different O&Ds and can reveal the major factors which have significant

impacts on the direct flight market as well. Additional to the historical direct share, the proposed hybrid model can use the additional information, such as the forecast of passengers at the origin and destination, the legacy share on the O&D market, and the price difference between direct and non-direct flight services to provide more accurate forecasting of direct share.

V. Conclusions

Air transportation direct share is an essential factor in air transportation planning, airlines' market strategy making, and airport operations scheduling. Direct share shows the passengers' distribution on direct and non-direct itineraries on an O&D. It indicates the air travelers' general preference for direct flight services on an O&D market under a certain supply. The O&D direct share time series reveals the evolution of a direct flight market. Based on the study of 1295 O&D pairs, we found direct share time series is O&D specific, which means for different O&Ds, the characteristics of the direct share time series may vary significantly from each other. An in-depth understanding of the characteristics of O&D direct share and accurate forecasting of direct share time series can benefit the air transportation planners, airlines, and airports in multiple ways. The object of this research is to develop accurate forecasting models of direct share time series on different O&Ds, which can be a promising replacement for the model used by FAA TAF.

To exploit the modeling capability of different models, both classical time series models and machine learning models are explored in this research. Based on the model comparison and analysis, we find the following characteristics of the models explored and proposed in this research.

- Holt-Winters Filter is a proper model for direct share time series with distinct seasonality and constant overall trend.
- SARIMA model is more capable of modeling seasonal and nonseasonal direct share time series, in which the underlying trend is more dynamic.
- The hybrid model proposed in this research, M_{hybrid}, is based on the combination of time series concept and machine learning techniques. The M_{hybrid} model can take advantage of additional information about the factors which have significant impacts on O&D direct share to provide more promising forecasting for certain O&D pairs.
- The developed M_{hybrid} model which applying Multiple Linear Regression (M_{hybrid_MLR}) and Random Forest model (M_{hybrid_RF}) shows different modeling capability of modeling seasonal and non-seasonal direct share time series. The M_{hybrid_MLR} is relatively more interpretable, which can provide insights into the direct flight market. While the M_{hybrid_RF} is more capable of handling the randomness in the direct share time series.

To fully exploit the modeling capability of different models and efficiently select the direct share forecasting model for different O&Ds, a modeling framework is proposed in this research. Based on the model performance comparison, the proposed modeling framework is a reliable replacement for the direct share forecasting model used by the FAA TAF.

VI. Future Work

In this research, we explored different models for time series modeling and forecasting. More models and varieties of models will be explored, such as the SARIMA model with covariates. In the future work, more parametric (e.g. ridge regression, Lasso regression, and beta regression, etc.) and non-parametric machine learning models (e.g. gradient boosting machine, support vector machines, and neural networks, etc.) can be explored to further improve the model performance. Meanwhile, we will try to include more features that may have impacts on the O&D direct share. We also plan to explore the fractional models in future work, if the hypothesis is validated.

Funding Sources

This research is partially funded by the Federal Aviation Administration project Passenger Route Share Forecast.

Acknowledgments

This research is under great support and help from the Office of Aviation Policy and Plans, Federal Aviation Administration. We thank all the colleagues who provided valuable insights and suggestions, especially David Chien, Roger Schaufele, and Chia-Mei Liu. We also want to send our thanks to Thea Graham, Robert Nazareth, and James Bouse for their great help on the DB1B data issues.

References

[1] Office of Aviation Policy and Plans, "Terminal Area Forecast Summary, Fiscal Year 2017-2045," *Federal Aviation Administration*, 2017.

[2] Lijesen, M. G., Nijkamp, P., and Rietveld, P., "Measuring Competition in Civil Aviation," *Journal of Air Transport Management*, Vol. 8, No. 3, 2002, pp. 189–197. (https://doi.org/10.1016/s0969-6997(01)00048-5).

[3] Pels, E., Nijkamp, P., and Rietveld, P., "Airport and Airline Choice in A Multiple Airport Region: An Empirical Analysis for the San Francisco Bay Area," *Regional Studies*, Vol. 35, No. 1, 2001, pp. 1–9. (https://doi.org/10.1080/00343400120025637).

[4] Bradley, M., "Behavioural Models of Airport Choice and Air Route Choice (Chapter 9 of Travel Behaviour Research: Updating the State of Play)," *Publication of: Elsevier Science, Limited*, 1998. (https://doi.org/10.1016/b978-008043360-8/50009-1).

[5] Tsui, W. H. K., Balli, H. O., Gilbey, A., and Gow, H., "Forecasting of Hong Kong Airport's Passenger Throughput," *Tourism Management*, Vol. 42, 2014, pp. 62–76. (https://doi.org/10.1016/j.tourman.2013.10.008).

[6] Terekhov, I., and Gollnick, V., "A Concept of Forecasting Origin-destination Air Passenger Demand between Global City Pairs Using Future Socio-economic Scenarios," *53rd AIAA Aerospace Sciences Meeting*, 2015, p. 1640. (https://doi.org/10.2514/6.2015-1640).

[7] Xie, G., Wang, S., and Lai, K. K., "Short-term Forecasting of Air Passenger by using Hybrid Seasonal Decomposition and Least Squares Support Vector Regression Approaches," *Journal of Air Transport Management*, Vol. 37, 2014, pp. 20–26. (https://doi.org/10.1016/j.jairtraman.2014.01.009).

[8] Dantas, T. M., Oliveira, F. L. C., and Repolho, H. M. V., "Air Transportation Demand Forecast through Bagging Holt Winters Methods," *Journal of Air Transport Management*, Vol. 59, 2017, pp. 116–123. (https://doi.org/10.1016/j.jairtraman.2016.12.006).

[9] Buaphiban, T., and Truong, D., "Evaluation of Passengers' Buying Behaviors toward Low Cost Carriers in Southeast Asia," *Journal of Air Transport Management*, Vol. 59, 2017, pp. 124–133. (https://doi.org/10.1016/j.jairtraman.2016.12.003).

[10] Srisaeng, P., Baxter, G. S., and Wild, G., "Forecasting Demand for Low Cost Carriers in Australia using an Artificial Neural Network Approach," *Aviation*, Vol. 19, No. 2, 2015, pp. 90–103. (https://doi.org/10.3846/16487788.2015.1054157).

[11] Marazzo, M., Scherre, R., and Fernandes, E., "Air Transport Demand and Economic Growth in Brazil: A Time Series Analysis," *Transportation Research Part E: Logistics and Transportation Review*, Vol. 46, No. 2, 2010, pp. 261–269. (https://doi.org/10.1016/j.tre.2009.08.008).

[12] Fildes, R., Wei, Y., and Ismail, S., "Evaluating the Forecasting Performance of Econometric Models of Air Passenger Traffic Flows using Multiple Error Measures," *International Journal of Forecasting*, Vol. 27, No. 3, 2011, pp. 902–922. (https://doi.org/10.1016/j.ijforecast.2009.06.002).

[13] Xiao, Y., Liu, J. J., Hu, Y., Wang, Y., Lai, K. K., and Wang, S., "A Neuro-fuzzy Combination Model based on Singular Spectrum Analysis for Air Transport Demand Forecasting," *Journal of Air Transport Management*, Vol. 39, 2014, pp. 1–11. (https://doi.org/10.1016/j.jairtraman.2014.03.004).

[14] Saâdaoui, F., Saadaoui, H., and Rabbouch, H., "Hybrid Feedforward ANN with NLS-based Regression Curve Fitting for US Air Traffic Forecasting," *Neural Computing and Applications*, 2019, pp. 1–13. (https://doi.org/10.1007/s00521-019-04539-5).

[15] Warburg, V., Bhat, C., and Adler, T., "Modeling Demographic and Unobserved Heterogeneity in Air Passengers' Sensitivity to Service Attributes in Itinerary Choice," *Transportation Research Record: Journal of the Transportation Research Board*, , No. 1951, 2006, pp. 7–16. (https://doi.org/10.1177/0361198106195100102).

[16] Garrow, L. A., *Discrete Choice Modelling and Air Travel Demand: Theory and Applications*, Routledge, 2010. (https://doi.org/10.4324/9781315577548).

[17] Freund-Feinstein, U., and Bekhor, S., "An Airline Itinerary Choice Model that Includes the Option to Delay the Decision," *Transportation Research Part A: Policy and Practice*, Vol. 96, 2017, pp. 64–78. (https://doi.org/10.1016/j.tra.2016.12.004).

[18] Coldren, G. M., Koppelman, F. S., Kasturirangan, K., and Mukherjee, A., "Modeling Aggregate Air-travel Itinerary Shares: Logit Model Development at a Major US Airline," *Journal of Air Transport Management*, Vol. 9, No. 6, 2003, pp. 361–369. (https://doi.org/10.1016/s0969-6997(03)00042-5).

[19] Montgomery, D. C., Jennings, C. L., and Kulahci, M., *Introduction to Time Series Analysis and Forecasting*, John Wiley & Sons, 2015.

[20] De Gooijer, J. G., and Hyndman, R. J., "25 Years of Time Series Forecasting," *International Journal of Forecasting*, Vol. 22, No. 3, 2006, pp. 443–473. (https://doi.org/10.1016/j.ijforecast.2006.01.001).

[21] De Faria, E., Albuquerque, M. P., Gonzalez, J., Cavalcante, J., and Albuquerque, M. P., "Predicting the Brazilian Stock Market Through Neural Networks and Adaptive Exponential Smoothing Methods," *Expert Systems with Applications*, Vol. 36, No. 10, 2009, pp. 12506–12509. (https://doi.org/10.1016/j.eswa.2009.04.032).

[22] Mahajan, S., Chen, L.-J., and Tsai, T.-C., "Short-Term PM2. 5 Forecasting Using Exponential Smoothing Method: A Comparative Analysis," *Sensors*, Vol. 18, No. 10, 2018, p. 3223. (https://doi.org/10.3390/s18103223).

[23] de Oliveira, E. M., and Oliveira, F. L. C., "Forecasting Mid-long Term Electric Energy Consumption through Bagging ARIMA and Exponential Smoothing Methods," *Energy*, Vol. 144, 2018, pp. 776–788. (https://doi.org/10.1016/j.energy.2017.12.049).

[24] Holt, C. C., "Forecasting Seasonals and Trends by Exponentially Weighted Moving Averages," *International Journal of Forecasting*, Vol. 20, No. 1, 2004, pp. 5–10. (https://doi.org/10.1016/j.ijforecast.2003.09.015).

[25] Gelper, S., Fried, R., and Croux, C., "Robust Forecasting with Exponential and Holt–Winters Smoothing," *Journal of Forecasting*, Vol. 29, No. 3, 2010, pp. 285–300. (https://doi.org/10.2139/ssrn.1089403).

[26] Wang, Y., Xu, C., Wang, Z., and Yuan, J., "Seasonality and Trend Prediction of Scarlet Fever Incidence in Mainland China from 2004 to 2018 using a Hybrid SARIMA-NARX Model," *PeerJ*, Vol. 7, 2019, p. e6165. (https://doi.org/10.7717/peerj.6165).

[27] Nacy, N. G., Badal, M. A., and Ibrahim, S. T., "Future Estimation for Electricity Interruption in Dohuk Governorate (Kurdistan-Iraq) using SARIMA Model," *Journal of Humanity Sciences*, Vol. 23, No. 1, 2019, pp. 201–215. (https://doi.org/10.21271/zjhs.23.1.14).

[28] Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M., *Time Series Analysis: Forecasting and Control*, John Wiley & Sons, 2015.

[29] Hansson, J., Jansson, P., and Löf, M., "Business Survey Data: Do They Help in Forecasting GDP Growth?" *International Journal of Forecasting*, Vol. 21, No. 2, 2005, pp. 377–389. (https://doi.org/10.1016/j.ijforecast.2004.11.003).

[30] Khandelwal, I., Adhikari, R., and Verma, G., "Time Series Forecasting using Hybrid ARIMA and ANN Models based on DWT Decomposition," *Procedia Computer Science*, Vol. 48, 2015, pp. 173–179. (https://doi.org/10.1016/j.procs.2015.04.167).

[31] Chen, C.-F., Chang, Y.-H., and Chang, Y.-W., "Seasonal ARIMA forecasting of inbound air travel arrivals to Taiwan," *Transportmetrica*, Vol. 5, No. 2, 2009, pp. 125–140. (https://doi.org/10.1080/18128600802591210).

[32] Milenković, M., Švadlenka, L., Melichar, V., Bojović, N., and Avramović, Z., "SARIMA modelling approach for railway passenger flow forecasting," *Transport*, Vol. 33, No. 5, 2018, pp. 1113–1120. (https://doi.org/10.3846/16484142.2016.1139623).

[33] Alpaydin, E., *Introduction to Machine Learning*, MIT press, 2009.

[34] James, G., Witten, D., Hastie, T., and Tibshirani, R., *An Introduction to Statistical Learning*, Springer, 2013, chapter and pages.

[35] Saâdaoui, F., "A Seasonal Feedforward Neural Network to Forecast Electricity Prices," *Neural Computing and Applications*, Vol. 28, No. 4, 2017, pp. 835–847. (https://doi.org/10.1007/s00521-016-2356-y).

[36] Hardle, W., Mammen, E., et al., "Comparing Nonparametric versus Parametric Regression Fits," *The Annals of Statistics*, Vol. 21, No. 4, 1993, pp. 1926–1947. (https://doi.org/10.1214/aos/1176349403).

[37] McCune, B., "Non-parametric Habitat Models with Automatic Interactions," *Journal of Vegetation Science*, Vol. 17, No. 6, 2006, pp. 819–830. (https://doi.org/10.1111/j.1654-1103.2006.tb02505.x).

[38] Seber, G. A., and Lee, A. J., *Linear Regression Analysis*, Vol. 329, John Wiley & Sons, 2012.

[39] Breiman, L., and Friedman, J. H., "Predicting Multivariate Responses in Multiple Linear Regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 59, No. 1, 1997, pp. 3–54. (https://doi.org/10.1111/1467-9868.00054).

[40] Faraway, J. J., *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*, Chapman and Hall/CRC, 2016. (https://doi.org/10.1201/b21296).

[41] Friedman, J., Hastie, T., and Tibshirani, R., *The Elements of Statistical Learning*, Springer Series in Statistics New York, 2001, chapter and pages.

[42] Svetnik, V., Liaw, A., Tong, C., and Wang, T., "Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules," *International Workshop on Multiple Classifier Systems*, Springer, 2004, pp. 334–343. (https://doi.org/10.1007/978-3-540-25966-4_33).

[43] Chang, L.-Y., and Chen, W.-C., "Data Mining of Tree-based Models to Analyze Freeway Accident Frequency," *Journal of Safety Research*, Vol. 36, No. 4, 2005, pp. 365–375. (https://doi.org/10.1016/j.jsr.2005.06.013).

[44] Quan, Z., Gan, G., and Valdez, E. A., "Tree-based Models for Variable Annuity Valuation: Parameter Tuning and Empirical Analysis," 2019. (https://doi.org/10.2139/ssrn.3247100).

[45] Khalilia, M., Chakraborty, S., and Popescu, M., "Predicting Disease Risks from Highly Imbalanced Data using Random Forest," *BMC Medical Informatics and Decision Making*, Vol. 11, No. 1, 2011, p. 51. (https://doi.org/10.1186/1472-6947-11-51).

[46] Bureau of Transportation Statistics, "Data Profile: Airline Origin and Destination Survey (DB1B)," , 2018. URL https://www.transtats.bts.gov/DatabaseInfo.aspDB_ID=125&DB_Name=Airline%20Origin%20and%20Destination%20Survey%20%28DB1B%29.

[47] Federal Aviation Administration, "Airport Categories Airports," , 2018. URL https://www.faa.gov/airports/planning_capacity/passenger_allcargo_stats/categories/.

[48] Lizotte, D. J., *Practical Bayesian Optimization*, University of Alberta, 2008.

[49] Snoek, J., Larochelle, H., and Adams, R. P., "Practical Bayesian Optimization of Machine Learning Algorithms," *Advances in Neural Information Processing Systems*, 2012, pp. 2951–2959.

[50] Winters, P. R., "Forecasting Sales by Exponentially Weighted Moving Averages," *Management science*, Vol. 6, No. 3, 1960, pp. 324–342. (https://doi.org/10.1287/mnsc.6.3.324).

[51] Hyndman, R., Koehler, A., Ord, J., and Snyder, R., *Forecasting with Exponential Smoothing: The State Space Approach*, Springer, Berlin, 2008.

[52] Zheng, X., Liu, C.-M., and Wei, P., "Air Transportation Direct Share Analysis and Forecast," *Journal of Advanced Transportation*, Vol. 2020, 2020. (https://doi.org/10.2514/6.2019-3187).

[53] Bureau of Transportation Statistics, "Database Name: Air Carrier Statistics (Form 41 Traffic)- U.S. Carriers," , 2018. URL https://www.transtats.bts.gov/Tables.aspDB_ID=110&DB_Name=Air%20Carrier%20Statistics%20%28Form%2041%20Traffic%29-%20%20U.S.%20Carriers&DB_Short_Name=Air%20Carriers.

[54] Hutengs, C., and Vohland, M., "Downscaling land surface temperatures at regional scales with random forest regression," *Remote Sensing of Environment*, Vol. 178, 2016, pp. 127–141. (https://doi.org/10.1016/j.rse.2016.03.006).

[55] Xia, X., Togneri, R., Sohel, F., and Huang, D., "Random forest classification based acoustic event detection utilizing contextual-information and bottleneck features," *Pattern Recognition*, Vol. 81, 2018, pp. 1–13. (https://doi.org/10.1016/j.patcog.2018.03.025).

[56] Díaz-Uriarte, R., and De Andres, S. A., "Gene Selection and Classification of Microarray Data using Random Forest," *BMC bioinformatics*, Vol. 7, No. 1, 2006, p. 3.

[57] Aiello, S., Eckstrand, E., Fu, A., Landry, M., and Aboyoun, P., "Machine Learning with R and H2O," *H2O booklet*, 2016.

[58] Hamner, B., Frasco, M., and LeDell, E., "Evaluation Metrics for Machine Learning," , 2018.