Application of Data Mining in Air Traffic Forecasting

Judit G. Busquets[*] and Dr. Eduardo Alonso[†]
City University London, London, EC1V 0HB, UK

Dr. Antony D. Evans[‡]
University of California, Santa Cruz, Moffett Field, CA 94035, USA

The main goal of the study centers on developing a model for the purpose of air traffic forecasting by using off-the-shelf data mining and machine learning techniques. Although data driven modeling has been extensively applied in the aviation sector, little research has been done in the area of air traffic forecasting. This study is inspired by previous research focused on improving the Federal Aviation Administration (FAA) Terminal Area Forecasting (TAF) methodology, which historically assumed that the US air transportation system (ATS) network structure was static. Recent developments use data mining algorithms to predict the likelihood of previously un-connected airport-pairs being connected in the future, and the likelihood of connected airport-pairs becoming un-connected. Despite the innovation of this research, it does not focus on improving the FAA's existing methodology for forecasting future air traffic levels on existing routes, which is based on relatively simple regression and growth models. We investigate different approaches for improving and developing new features within the existing data mining applications in air traffic forecasting. We focus particularly on predicting detailed traffic information for the US ATS. Initially, a 2-stage log-log model is applied to establish the significance of different inputs and to identify issues of endogeneity and multi-colinearity, while maintaining the simplicity of current models. Although the model shows high goodness of fit, it tested positive for both mentioned issues as well as presenting problems with causality. With the objective of solving these issues, a 3-stage model that is under development is introduced. This model employs logistic regression and discrete choice modelling. As part of future work, machine learning techniques such as clustering and neural networks will be applied to improve this model's performance.

I. Introduction

AVIATION is the fastest growing mode of passenger transportation globally[1]. Global air travel growth has averaged approximately 5% per year over the past 30 years, which is an annual growth rate twice that of global Gross Domestic Product (GDP)[2]. Forecasts for future growth are also high. Airbus[3] predicts that air traffic levels will double in the next 15 years, in line with historical trends. In the US, medium-term predictions forecast that the industry will grow from 731 million passengers in 2011 to 1.2 billion in 2032[4]. Economically, the aviation sector had a global impact of £2.4 trillion in 2012, equivalent to 3.4% of the global GDP[5].

While growth in air transportation has significant economic benefits, it also has negative consequences, including increasing flight delays and environmental impacts at both the local level (e.g., air quality and noise) and global scale (e.g., climate change), as reported by the International Panel on Climate Change[6]. These positive and negative consequences of aviation must be traded-off by policy makers in the development of policy. A topical example is the UK air transportation sector. It is estimated that London, one of the key gateway cities to Europe, is at risk of losing its economic and social competitiveness to other European hubs because of airport capacity constraints[7]. Failing to supply capacity to serve the long-term air travel demand in the South-east of England is estimated to have a negative economic impact of £18 to £20 billion on users and providers of airport infrastructure, and £30 to £45 billion on the wider economy[7]. There are also, however, concerns about the consequences of adding

[*] PhD Student, Air Transport Engineering, Student Member.

[†] Reader in Computing, School of Mathematics, Computer Science & Engineering, Non-member.

[‡] Associate Research Scientist, University Affiliated Research Center, M/S 210-6, Senior Member.

capacity, including increased noise and reduced air quality associated with increasing capacity at Heathrow Airport, as well as questions about whether the UK could still meet its emissions targets given the associated rise in air traffic.

To make policy trade-offs, it is critical that good forecasts of future demand for air traffic exist, as well as good forecasts of how airlines are likely to serve this demand. This is particularly important given the long timescales associated with airport capacity expansion, especially in many developed economies where there is significant resistance to airport development. Good forecasts of future demand are also critical for airlines and airport authorities, which must plan their operations accordingly, and often need to order equipment well before it is required. Good forecasting requires a solid understanding of the most important drivers of supply and demand. Consequently, not only do historical trends in air transportation need to be studied, but the intrinsic drivers underlying passenger and airline behavior must also be understood.

Aviation stakeholders tend to generate their own air travel forecasts and forecasting methodologies. While a diversity of methodologies exist, econometric, gravity and time-series models prevail. Most of these models are based on correlating aviation growth and socio-economic growth, (e.g. Ref. 8), and are characterized by their relative simplicity. For example, the FAA[9] applies a simple growth factor algorithm to allocate traffic across the US ATS. These approaches also often use similar explanatory variables, generally chosen given the judgment of domain experts. More complex approaches from the literature are often not used because of the associated drawbacks, such as computational intensity or relatively low accuracy.

In this paper we attempt to improve current forecasting methodologies by better understanding the patterns underlying the historical supply and demand for air travel, using the US domestic air transportation system as an example. In order to achieve this, three innovations are investigated: the use of several data mining techniques to develop a forecasting methodology; the use of a larger range of explanatory variables than is commonly considered; and explicitly modeling the distribution of city-pair passenger demand between itineraries.

Data mining provides us with a variety of computational methods to analyze relationships that exist within large datasets, identifying dominant drivers of important outcomes, predicting the probability of new outcomes, and determining anomalous behavior. The application of data mining techniques in air traffic forecasting constitutes a recent trend started in the last decade. Aviation is an interesting system in which data mining can be applied since it is a large and complex system that involves the generation of a large scale and unstructured mixture of data in various data formats. Data mining aims to transform these datasets into applied knowledge. From the application of data mining techniques in air transportation data, several benefits can be obtained. This includes improved revenue management[10], advances in safety within the air transportation system[11,12] as well as improving on existing air travel demand forecasts. Although the potential benefits in applying this type of approach to air traffic forecasting are widely acknowledged, little research has been done up until now.

The remainder of the paper is structured as follows: a review on existing forecasting methodologies in air transportation is outlined in Section II. This is followed by the paper's objectives in Section III. The modelling approach used in this study is detailed in Section IV, including an explanation of the parameters identified as the key drivers of air transportation supply and demand. Information regarding the data sources is outlined in Section V. The corresponding results are shown in Section VI, followed by a discussion on future work in Section VII.

II. Literature Review

How much, how quickly and where air transportation will grow is driven by a number of factors, some economic-related and others linked to demographics and socio-economic evolution. The primary objective of forecasters is to understand how different drivers will contribute to explaining the future of air transportation. Methodologies used to forecast future air traffic differ significantly depending on the forecast's purposes.

Swan[13] identifies three common methods of forecasting passenger demand for air travel: trends; gravity models; and stimulation models, which forecast the increase in traffic from estimated changes in fares and service levels. Historical trends are the most common forecasting technique used to predict air travel demand. This involves the use of econometric equations in which passenger and freight demand is regressed against economic activity over historical time periods. Factors that induce economic growth are also sometimes taken into account, e.g., demographic variables such as population. Boeing and the ICAO Asia/Pacific Area Traffic Forecasting Group[14] use this approach. In the case of Boeing, future air travel demand, measured in Revenue Passenger Kilometre (RPK), is estimated through an equation that compares air travel growth mainly against economic growth measured in GDP (Ref. 8). Some of the forecasting methodologies are combined with qualitative techniques such as surveys and questionnaires. These are based on intuition and subjective evaluation including expert opinion[15]. For example, UK aviation forecasts, produced by the Department for Transport (DfT), are mainly based on econometric models.

However, DfT also uses Civil Aviation Authority's (CAA) Passenger Surveys to calibrate the weight values that represent the strength of each factor driving passenger airport choice[16].

In contrast, Eurocontrol's medium-term forecasts are produced by combining regression techniques and time-series analysis[17]. This combination is most appropriate for producing short and medium term forecasts[18] and is mainly based on analyzing historical data and trends[19]. The FAA's Terminal Area Forecast (TAF) is based upon historical local and national measures that influence aviation activity as well as those drivers within the industry itself. In this manner, passenger demand at a particular airport is derived independently of the ability of that airport and its supporting air traffic control system to furnish the capacity required for meeting that specific demand.

Broadly, the FAA's air traffic forecasting process is split into two stages. The first stage consists in modelling the true-origin ultimate-destination (O-D) passenger demand flows using econometrics models. These are based on regression analysis using historical segment-pair air traffic data. Information such as airfares, demographic and income parameters are included within the set of independent variables. The second stage consists in combining the TAF results, which account for the estimates in traffic change at individual airports, with the most recent airline schedules obtained from T-100 segment data. The allocation process is performed by the application of the Fratar algorithm. This is a type of trip distribution algorithm based on a growth factor method, by which the connectivity between an airport or city pair is evaluated[20]. By using the Fratar algorithm, the O-D passenger demand forecasted during the first stage of the process is distributed across the possible routes of the network constructed from the airline schedule in the way that best satisfies the forecast airport level growth. As a result, the future airline flight schedule is generated. The above approach has a few drawbacks. Firstly, there is an inherent assumption that the future route network structure will remain the same as the current network structure[21]. Secondly, the Fratar algorithm has a number of limitations, including that it does not account for changes in transport costs and assumes that resistance to travel will remain the same. Finally, there is significant uncertainty associated with econometric models, mainly because the behavior of the transport system is correlated with relatively few socio-economic and historical features.

Much of the existing research is in line with the industry's use of econometric models, time series and gravity models as the dominant quantitative approaches to estimate future air traffic demand and supply. However, research has also focused on improving the existing techniques to forecast future air traffic by applying alternative modelling approaches, such as agent-based modelling and discrete choice modelling. For example, Ref. 22 combines a gravity model of passenger O-D demand with an agent-based model of airline decision-making, simulating airline frequency competition using a myopic best response game, in order to model the airlines operational responses to environmental constraints. The key difference between this work and the econometric, gravity and time series models described above is that it attempts to model the airline decision making process explicitly, instead of estimating model parameters based on historical data. Results obtained show that airline response to any type of capacity constraint and competition between airlines is important when trying to understand the underlying principles behind the evolution of the air transportation system. Similarly, Ref. 23 uses a two-stage Nash best-response game to evaluate the most appropriate hub-and-spoke network for an airline to develop in a competitive environment. Given three different settings, the study examines when equilibria in the air transportation industry would occur. Results obtained show that demand plays an important role in the solution outcome. Finally, focusing on modeling competition to determine air-travel itinerary shares, Ref. 24 presents a 3-level weighted nested logit model to predict airline ridership at the itinerary level and help carriers in medium and long term decision-making. This model is applied at an aggregate level and variables included are chosen to capture the inter-itinerary competition dynamic along three dimensions: time of the day, carrier and level of service. Results obtained suggest that itineraries sharing the same time show a moderate level of competition while those sharing time and carrier or level of service show a high level of competition. Although results from these researches are promising, they are computational intensive, limiting their application to relatively small network sets. The ability of some of the models to reproduce existing air traffic is also limited and further model refinement and verification is still required to better capture passenger choice effects. However, important insights can be gained into what some of the key drivers of airline decision-making are, and therefore what decision variables could be added to existing approaches.

Little existing research has been identified linking data mining with air traffic forecasting. Nevertheless, a few key studies can be highlighted. Ref. 25, one of the first studies done in this field, used neural networks to forecast international airline passenger traffic between the US and South Korea. Results showed that neural networks enhanced forecasting accuracy and went beyond the capabilities of the more conventional statistical analysis used at the time. The decision variables used in this study include eleven dummy variables defining the monthly seasonality and one time variable reflecting the trend effect. Similarly, Ref. 26 applied a hybrid model of neural network and statistical analysis in order to forecast air traffic flow at fixes on a 30-min aggregation level within China's air traffic network. For this study, the decision variables were a combination of information provided by radar data and historical airline flight schedule data.

Ref. 27 uses Support Vector Machine (SVM) techniques to develop a model that improves on the simple time-series approach to air traffic forecasting. The study highlights the advantage of this approach over traditional econometric models proving that potential benefits can be obtained when applying data mining techniques in air traffic forecasting. Finally, Ref. 28 developed a model using data mining techniques to capture mechanisms by which the US ATS network evolves. Ref. 28's approach uses complex network theory quantitative parameters as explanatory variables in the input dataset, and trains logistic regressions and neural networks to predict the likelihood of previously un-connected airport-pairs being connected in the future, and the likelihood of connected airport-pairs becoming un-connected. The main objective of this study was to improve on the FAA TAF assumption of a static routing network, which was done by adding an initial step that models US network evolution. Notwithstanding the innovativeness of Ref. 28's method, the accuracy of the results was between 20% and 40%, leaving room for improvement. In addition, the work done by Ref. 28 did not improve on the current FAA methodology for forecasting air traffic levels on existing routes. Hence, further enhancements in the approach are possible. Such improvements are the focus of this paper.

III. Objectives

The primary objective of this research is to develop a model for forecasting future air traffic levels. This is inspired by previous research[28] that focused on improving the FAA's forecasting methodology and for which further potential improvements have been identified. Consequently, in an attempt to improve on the current FAA's forecasting methodology, the model described in this paper is expected to:

- Highlight the most important factors underlying the growth of the US ATS. This will allow identification of the key drivers of evolution in the US ATS.
- Predict future air traffic growth, and hence, the evolution of the ATS system.

In order to achieve these objectives, the developed model includes three new elements beyond that of the existing research:

- The use of data mining techniques to model the US ATS evolution. By the use of these techniques, the resulting model will predict air traffic with improved accuracy and precision levels while maintaining the simplicity of existing econometric, gravity and time-series models.
- The consideration of a larger set of explanatory variables than is typically considered in existing air traffic forecasting approaches.
- Explicitly modeling the distribution of city-pair passenger demand between itineraries, so as to better predict airport-pair flows.

IV. Approach

A. Factors Influencing Future Air Traffic

One of the means by which improvements in forecasting air traffic are expected to be achieved is the inclusion of a larger range of explanatory variables than is typically considered in existing approaches, based particularly on findings from the agent-based modeling of airline decision making by Ref. 22 as well as the findings of other studies[23,28,29]. This extended set of explanatory variables is expected to better capture the underlying behavior that drives the aviation industry, including the underlying drivers of demand for air travel, and the underlying drivers of airline decision making to supply flights to serve this demand.

The explanatory variables used as input data for the air traffic forecasting model developed using data mining algorithms are classified into three groups:

I. Network theory quantitative measures. Their inclusion is based on a representation of the air transportation system using topology and mathematical graph theory, as by Ref. 28 and Ref. 29. Given this, the ATS is represented as a natural network that consists in well-defined nodes (airports) and links (flights that connect these nodes or airports);

II. Socio-economic variables;

III. Aviation-related variables. These represent the airline response to capacity constraints derived from insights gained through the work of Ref. 22 and Ref. 23.

Table 1 shows a brief explanation of all input data considered as explanatory variables.

Table 1. Key factors to influence the future air traffic demand, where *node* represents an airport within the US ATS; A_{ij} is a non-weighted adjacency matrix with all its entries being binary, indicating whether a link between two nodes –i.e. *i* and *j*– is present ("1") or not ("0"); and $\boldsymbol{A^w_{ij}}$ is a weighted adjacency matrix in which each entry (corresponding to a link between node *i* and node *j*) has a scalar weight that signifies some distinguishing trait characterizing that link, in this case flight frequency.

Measure	Symbol and Equation	General Explanation
Network Theory metrics		
Node degree	$NodeDeg_i = \sum_j A_{ij}$	In the transport network, node degree accounts for the total number of connections node *i* has with other nodes.
Node weight	$NodeWeight_i = \sum_j A^{w}_{ij}$	In the transport network, node weight refers to the total number of flights associated to node *i* .
Eigenvector centrality, which is a measure at the collective level that accounts for the influence that neighbouring nodes have to a given node.	$ECV_i = \lambda^{-1}\sum_j A^w_{ij} x_j$ where "x" is an eigenvector of the adjacency matrix	It assumes that the importance or popularity of an airport is proportional to the sum of centralities of the neighbouring airports (nodes) to which it is connected.
Clustering Coefficient	$CC_i = \frac{1}{k_i(k_i-1)} \sum_{j,k} A_{ij} A_{ik} A_{jk}$	It quantifies how many times node *i* forms triangular sub-graphs with their adjacent nodes.
socio-economic parameters		
Population	*Pop*	N° of inhabitants within the Metropolitan Statistical Areas (MSA) linked to node *i* .
Income	*Inc*	Mean household income per capita by the MSA associated with node *i* .
Aviation-related factors		
How special is a city, Boolean parameter	$Special\ 1_{ij}$= "**0**" if both cities are not special; "**1**" otherwise. $Special\ 2_{ij}$= "**0**" if both cities are special; "**1**" otherwise.	This variable represents whether the cities associated with a specific airport-pair have any attractiveness, business and leisure wise, that would promote air travel demand.
Connectivity by rail and/or road, Boolean parameter	$Roadrail_{ij}$ = "**0**" if accesibility by road or/and by rail exised for both cities & stage length ≤160 mile; "**1**" otherwise	This variable refers to the accessibility by road and rail transport modes between node *i* and node *j*.
Hub/no hub, Boolean parameter	$Hub1_{ij}$= "**0**" if both airports are a hub; "**1**" otherwise. $Hub\ 2_{ij}$= "**0**" if both airports are not hubs; "**1**" otherwise.	$Hub1_{ij}$ and $Hub2_{ij}$ refer to whether airports associated with an airport-pair are operated as hubs or not.
Number of airports by city	$Num\_airprt\_city_i$	Measure of number of airports located in a city. It accounts for airport competitiveness.
Stage length	$Stage\_length_{ij}$	Distance between airport-pair (from airport *i* to airport *j).*
Fuel price	$Fuel\_\ price$	Average annual fuel price.
Number of airlines by airport	$Num\_airlines\_airpt_i$	Measure of number of airlines operating at airport *i*.

Table 1. Key factors to influence the future air traffic demand (cont.).

Measure	Symbol and Equation	General Explanation
	Aviation-related factors (Cont.)	
Delay	$Delay_{ij} = D(GateDep)_{i} + D(TO)_{i} + +D(Airbone)_{ij} + D(TI)j$	*Delay* gives a sense of the capacity constraints of a given airport. It is captured through computing the difference between expected gate-to-gate time and the real time that a specific airport-pair route service has related.
Flight frequency from previous year, the auto-regressive term.	$Fltfreq\_previousyr_{ij}$	The use of this variable comes from the consideration that current observations of the dependent variable (at time *t*) are partly explained and generated by a weighted average of past that variable (at time *t-1*). This accounts for the fleet constraint, in that the current schedule is impacted by how many flights were operated in the previous year.
Number of passengers	$ODdemand_{ij}$	This represents the O-D demand between city i and city j in number of passengers.
Generalised Cost	$Generalised\_Cost_{ij} = AirFare_{ij} + \left(T_{Flight(ij)} * T_{Value} + T_{Delay\,(ij)} * Tdelay_{Value}\right)$	This variable refers to the generalised cost that a passenger has to incur to travel from city i to city j. It considers the airfare, the value of travelling time and the value of delayed time

B. Detailed Forecasting Methodology

Along with forecasting future air travel demand, effort is directed at identifying those factors that drive the evolution of the air transportation system. Therefore, this research attempts to develop an approach using data mining techniques that would predict air traffic growth with significant accuracy and precision levels while maintaining the simplicity levels of existing econometric, gravity and time-series models. An attempt will also be made to incorporate the methods developed by Ref. 28 in modeling the evolution of the US ATS network by allowing the possibility of new routes emerging and existing routes disappearing.

In order to develop a successful forecasting methodology, a variety of approaches are considered. The performance of each approach will be evaluated and assessed, and consequently, the model with the most potential will be identified. This validation process will compare the results of each approach to observed historic air traffic levels.

Two forecasting approaches are presented in this paper. The first approach uses linear regression with logarithmic transformation in both the dependent and independent variables – i.e. a log-log model. The implementation of this extended version of the more basic linear model is based on its ability to handle non-linear relationships between dependent and independent variables. This is done while maintaining the simplicity of the linear model[30]. By applying the natural logarithm, rate of change is accounted for rather than the absolute values of the variables, removing the growth over time in the variance of the data[19]. The second approach is an extended version of the previous one in which the log-log models are combined with other techniques, including a classification algorithm and a discrete choice model that captures passenger itinerary choice.

The two forecasting approaches differ in the number of explanatory variables and in the means by which the data is handled. Each approach is as follows:

- 2-stage log-log model: In a first stage, the air travel demand by city-pair (*ODdemand*) is estimated using *Pop, Inc, Special1, Special2, Roadrail* and *Generalised_cost* as input variables (See Table 1 for explanation of variables). This follows the approaches used to predict O-D passenger demand described by Ref. 31 and Ref. 22. In the second stage, the predicted *ODdemand* is then used as an input variable to predict the flight frequency by airport-pair, along with the network theory metrics, *Hub1, Hub2, Num_airpt_city, Fltfreq_previousyr* and *Fuel_price*. While this approach is not as simple as a 1-stage log-log model, it is expected to have improved performance.
- 3-stage model: This approach adds to the 2-stage log-log model an intermediate step with the objective

of transforming the O-D demand by city-pair extracted from the 1st stage into passenger demand by airport-pair, which serves as a stronger driver of airport-pair air traffic. This predicted passenger demand by airport-pair is then used as input variable to predict the flight frequency by airport-pair, along with the network theory metrics and aviation-related variables – as in the 2nd stage of the previous model.

Initially, simpler models were tested, such as several 1-stage log-log models. However, they were discarded due to poor performance.

In all the approaches considered, at least 20 flights per year must be operated between an airport-pair for it to be considered operational. Similarly, when estimating the O-D passenger demand by city-pair, in the 2-stage approach, at least 3,000 passengers per year must travel between the cities for demand to be considered.

The instances used to train the log-log models are generated by airport-pair or city-pair, which in some cases requires data transformation. For those parameters that take continuous values and are unique to individual airports or cities, the square root of the product of the two airports' or cities' characteristic is computed – e.g., given that city *i* has population Pop_i and city *j* has population Pop_j, the explanatory variable related to population for city pair *ij* is computed as shown in Eq. (1). In addition, all data is normalized, which helps to reduce skewness.

$$Pop_{ij} = \left(Pop_i * Pop_j\right)^{1/2} \tag{1}$$

For binary variables, which represent an attribute with two distinct categories, such as the variable *Special1*, dummy variables have been used, as described in Table 1.

Considering all these assumptions, the two equations used in the 2-stage log-log model are derived. The first one is presented in Eq. (2.a) and estimates the O-D passenger demand by city-pair in number of passengers. *Constant1*, ε, θ, μ, π, τ, κ are the coefficients to be estimated. The second equation predicts the flight segment frequency by airport-pair using the O-D passenger demand estimated in the first stage of the process as an explanatory variable. This is shown in Eq. (2.b) where *Constant2*, α, β, γ, δ, ρ, σ, ζ, φ, ι, υ are the coefficients to be estimated. Note that while Eq. (2.a) estimates city-pair demand, Eq. (2.b) estimates airports-pair flight frequency. The city-pair demand used in Eq. (2.b) is the city pair demand associated with the airports under question.

$$\begin{aligned} ODdemand_{ij} = {} & Constant1 + \left(Pop_{ij}\right)^{\varepsilon} + \left(Inc_{ij}\right)^{\theta} + \mu(Special1_{ij}) \\ & + \pi(Special2_{ij}) + \tau(RoadRail_{ij}) + \left(Generalised\_Cost_{ij}\right)^{\kappa} \end{aligned} \tag{2.a}$$

$$\begin{aligned} Fltfreq_{ij} = {} & Constant2 + \left(NodeDeg_{ij}\right)^{\alpha} + \left(NodeWeight_{ij}\right)^{\beta} + \left(CC_{ij}\right)^{\gamma} + \left(ECV_{ij}\right)^{\delta} + (ODdemand_{ij})^{\phi} \\ & + \rho(Hub1_{ij}) + \sigma(Hub2_{ij}) + \left(Num\_airprt\_city_{ij}\right)^{\varphi} + \left(Fltfreq\_previousyr_{ij}\right)^{\iota} \\ & + \left(Fuel\_price_{ij}\right)^{\upsilon} \end{aligned} \tag{2.b}$$

For the 3-stage model 4 equations are used. A flowchart showing the 3-stage model is presented in Fig. 1. In the 1st stage, Eq. (2.a) is applied to estimate air travel demand by city-pair. During the 2nd stage, in which estimated O-D demand by city-pair is transformed into passenger demand by airport-pair, a discrete choice model is applied to distribute the city-pair passenger demand predicted during the 1st stage across the available itineraries. This allows passenger demand by airport-pair to be calculated.

The availability of itineraries is identified using a classification algorithm similar to that used by Ref. 28 – i.e. logistic regression –, and will predict the likelihood of previously connected airport-pairs being disconnected in the future, as well as the likelihood of un-connected airport-pairs being connected in the future. For a non-stop itinerary, if a non-stop flight between the cities is predicted to be served by air traffic, then this itinerary is feasible. For a connecting itinerary, the itinerary will only be considered feasible if every flight leg in the itinerary (e.g., from the origin city to the hub, from the hub to the destination city, etc.) is predicted to be served by air traffic. The probability of a flight segment being connected is calculated applying Eq. (3) where Θ^T represents the set of parameters to be estimated.

$$h_{\theta}(x) = \frac{1}{1 + e^{-Z}} \tag{3}$$

Where

$$Z = (\theta^T x) = \theta_1 + \theta_2 (NodeDeg_{ij}) + \theta_3 (NodeWeight_{ij}) + \theta_4 (CC_{ij}) + \theta_5 (ECV_{ij}) + \theta_6 (OD_{(AP)} demand_{ij}) + \theta_7 (Pop_{ij}) + \theta_8 (Special1_{ij}) + \theta_9 (Special2_{ij}) + \theta_{10} (Hub1_{ij}) + \theta_{11} (Hub2_{ij}) + \theta_{12} (Fuel\_price_{ij})$$

After identifying the available itineraries, the O-D demand by city-pair obtained from the 1st stage is distributed across the available itineraries identified by the classification algorithm using a discrete choice model. This allows the flight segment passenger demand by airport-pair to be estimated, based on the passenger itinerary demand from all O-D city-pairs. In order to apply the discrete choice model, the US was divided into five regions, as done by Ref. 24: four Continental time zones (Central, East, Mountain and West) and a region for Alaska and Hawaii. The number and nature of these regional clusters will be modified using clustering techniques in future work. Given these regions, 18 entities have been defined: considering all 16 possible combinations of the Continental time zones – e.g., Central-Central (C-C), Central-East (C-E), Central-Mountain (C-M), Central-West (C-W), […], West-Mountain (W-M), West-West (W-W) –; as well as an entity for Alaska and Hawaii to Continental US and an entity for the Continental US to Alaska and Hawaii.

At this early stage of the work presented in this paper, a total of 26 itinerary options have been considered as the universal choice set which considers a non-stop service and 25 one-stop services in one of the 25 US hub airports considered in this study. For each entity, the set of alternatives is a subset of the universal choice set. The passenger choice of itinerary is modeled based on travel time (*TT*) and travel cost (*TC*).

The annual share of passenger demand assigned to each itinerary between a given city-pair is modeled as an aggregate multinomial logit (MNL) function and is given by Eq. (4) where S_i is the passenger share assigned to itinerary *i*, V_i is the utility function or value of itinerary *i* and the summation is over all itineraries for a given airport-pair. The utility function (V_i) is a linear function of the explanatory variables – i.e. *TT* and *TC* –. Equation (5) shows the V_i used in this study where β_{1i} is the intercept corresponding to itinerary *i*, and β_2 and β_3 are the coefficients of TT and TC respectively. β_{1i}, β_2 and β_3 are the coefficients to be estimated.

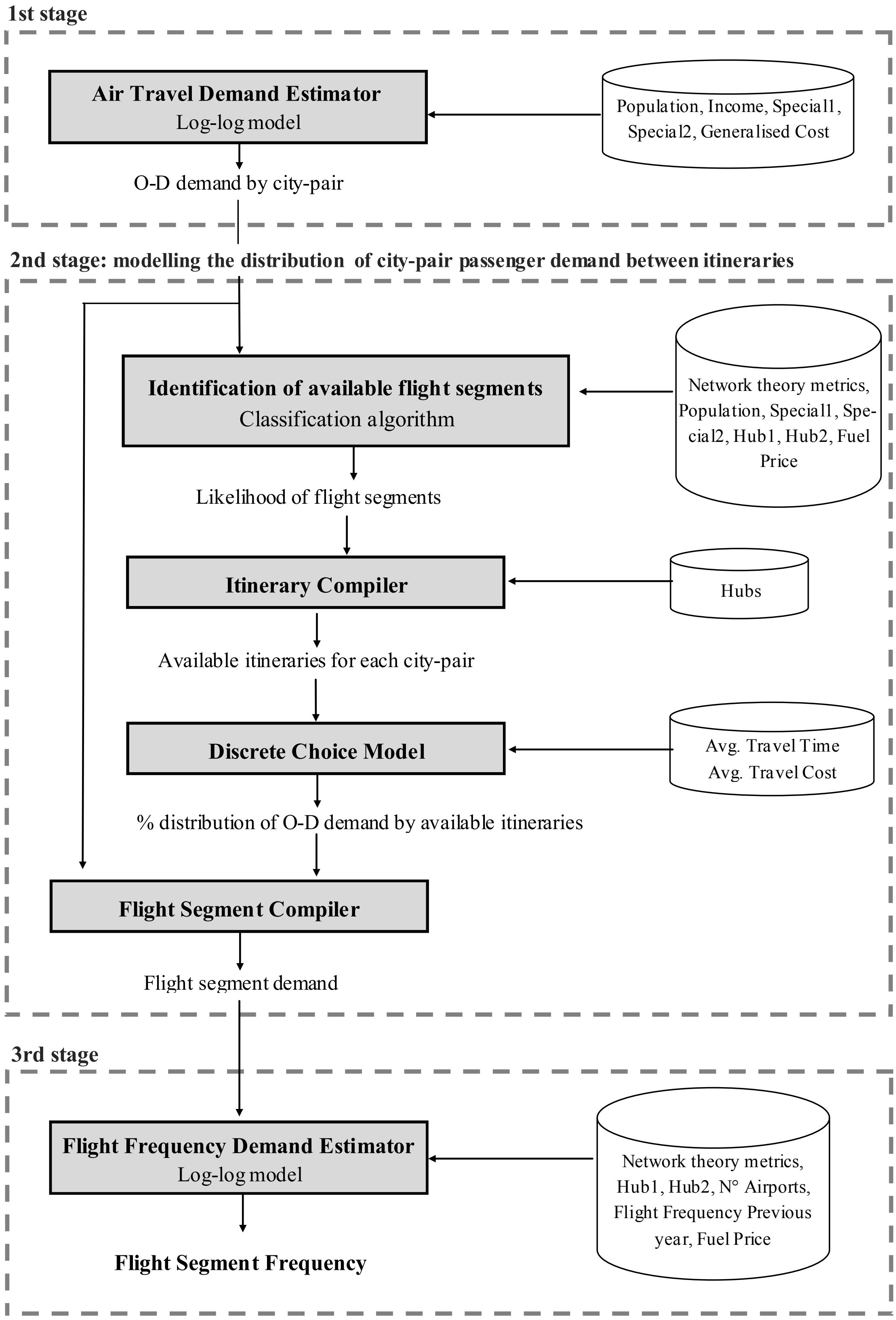
$$S_i = exp(V_i) / \sum_j exp(V_j) \tag{4}$$

$$V_i = \beta_{1i} + \beta_2 TT + \beta_3 TC \tag{5}$$

The estimated flight segment passenger demand by airport-pair will then be used as one of the input variables during the 3rd stage of the 3-stage model to estimate the flight segment frequencies by applying Eq. (6). Note that the only difference between this 3rd stage and the 2nd stage of the 2-stage log-log model (Eq. 2b) is that flight segment demand (*APdemand$_{ij}$*) is used as one of the explanatory variables instead of O-D demand by city-pair. The rest of the input variables are identical.

$$Fltfreq_{ij} = Constant\,3 + (NodeDeg_{ij})^{\alpha} + (NodeWeight_{ij})^{\beta} + (CC_{ij})^{\gamma} + (ECV_{ij})^{\delta} + (APdemand_{ij})^{\omega} + \rho(Hub1_{ij}) + \sigma(Hub2_{ij}) + (Num\_airprt\_city_{ij})^{\varphi} + (Fltfreq\_previousyr_{ij})^{\iota} + (Fuel\_price_{ij})^{\upsilon} \tag{6}$$

Figure 1. Flowchart of the 3-stage model.


1st stage
Air Travel Demand Estimator
Log-log model
Population, Income, Special1, Special2, Generalised Cost
O-D demand by city-pair
2nd stage: modelling the distribution of city-pair passenger demand between itineraries
Identification of available flight segments
Classification algorithm
Network theory metrics, Population, Special1, Special2, Hub1, Hub2, Fuel Price
Likelihood of flight segments
Itinerary Compiler
Hubs
Available itineraries for each city-pair
Discrete Choice Model
Avg. Travel Time
Avg. Travel Cost
% distribution of O-D demand by available itineraries
Flight Segment Compiler
Flight segment demand
3rd stage
Flight Frequency Demand Estimator
Log-log model
Network theory metrics, Hub1, Hub2, N° Airports, Flight Frequency Previous year, Fuel Price
Flight Segment Frequency

As a final note, two issues that can affect the validity of the approaches followed in this study have been studied: endogeneity[§] and multi-collinearity[**]. In order to test for endogeneity the *Hausman specification* test is performed[19]. Endogeneity tests are carried out for two explanatory variables, *num_airlines_airpt* and *delay*.

The second potential issue, multi-collinearity or ill-conditioned data, is tested by computing a condition number, the Variance Inflation Factor (VIF). If the VIF is above a specific threshold, it suggests the presence of ill-conditioned data[32]. A more in-depth technique proposed and explained in Ref. 32 is also applied. Multi-collinearity tests are carried out on all explanatory variables.

V. Application

The models described above are applied to a network of 337 airports within the US ATS. This represents a significantly larger dataset than that modeled by Ref. 22, and is comparable to that modeled by Ref. 28. The exact choice of airports is made for compatibility with the AIM Project[33], allowing for integration into the AIM modeling framework in the future. These 337 US airports represent the US airport set included in a global set of 1,277 airports, serving 95% of global RPK. Along with the airport set mentioned, the compilation of the corresponding US cities, special city variables, and road-rail variables are identical to those in Ref. 33.

For the purposes of estimating model parameters, historic socio-economic data, covering population and mean household income per capita, is extracted from the US census Bureau[34] and Bureau of Economic Analysis[35], respectively. Historical flight frequency data is extracted from US Department of Transport T-100 data[36], while historical information on passenger demand data and airfares is extracted from the Airline Origin and Destination Survey (DB1B), which contains a 10% sample of airline tickets from reporting carriers[37]. Travel times used in the discrete choice model using *Biogeme*[38], are also extracted from Ref. 37. Flight delay information is obtained from the FAA Aviation System Performance Metrics (ASPM) database[39].

The time-step considered in this study is annual. The period taken into account for estimating the model corresponds to the year range from 2003 to 2007. This period was considered because it represents a reasonably stable period not affected by any major external factor such as a terrorist attack or an economic crisis. Because more than one year is considered, it is necessary to convert monetary data to the same year, which is done based on the Consumer Price Index (CPI)[40] to 2007 US dollar values.

Once the models are estimated, they will be applied to forecast traffic growth in the same network of 337 airports into the future, using population, income and oil price growth forecasts from the US Department of Transport[41]. These results will then be compared to those of the TAF in future work.

VI. Model Estimation Results

A. First approach: 2-stage log-log model

As described above, the first stage of the 2-stage log-log model consists of estimating the O-D passenger demand between city-pairs. The model estimation results for this stage are presented in Table 2. The second stage consists of taking the predicted air travel demand by city-pair along with other parameters to form the input matrix for predicting flight frequency by airport-pair. Results for this second stage are presented in Table 3.

The estimated model predicting O-D passenger demand by city-pair, presented in Table 2, has an adjusted R^2 of 0.61. Given that there is no auto-regressive term, this is considered acceptable. Estimated coefficients are consistent with expectations and there is no sign of multi-collinearity or endogeneity. In addition, all coefficients are statistically significant at the 95% level of confidence, and are similar in value to those estimated by Ref. 22**Error! Bookmark not defined.** and Ref. 31.

For the second stage, the estimated model is run for two cases, with the estimated coefficients and the corresponding t-statistics, presented in Table 3, shown in each case. In the first case, the input dataset as described in §IV.B is included. In the second, Node degree (*NodeDeg*) and Clustering Coefficient (*CC*) are dropped, because they are not found to be statistically significant at 95% confidence level in the first case. The variable *num_airlines* is dropped in both cases due to issues with multi-collinearity. The estimated coefficients do not vary drastically between the two cases. However, in the second case, variable *Hub1* becomes only statistically significant at the 86th

[§] Endogeneity refers to the situation when one or more explanations variables are correlated to the dependent variable and consequently, to the disturbance term.

[**] Multi-collinearity refers to the situation when two or more explanatory variables are highly, but not perfectly, correlated with each other. Therefore, any change on one of the collinear variable will cause a change on the rest of collinear variables.

percentile confidence level. In both cases, the adjusted R^2 is 0.84, which is considered reasonable given the inclusion of the auto-regressive term within the explanatory variables.

The estimated coefficients are all of the expected sign, suggesting no causation problems. With the removal of *num_airlines* no endogeneity or multi-collinearity issues were detected. However, four variables dominate, with significantly larger coefficients compared to the rest of the variables. These are the auto-regressive term (*Fltfreq_previousyr*) (0.83), *Fuel_price* (-0.16), *NodeDeg* (0.16), and *ECV* (0.16). This is a cause for concern given that parameters such *ODdemand*, which would be expected to have a large influence on predicting future flight frequency, have limited influence.

These model estimation results therefore represent an improvement with respect to the discarded 1-stage log-log models which presented endogeneity, collinearity and causation issues, but further improvements are still possible.

Table 2. 2-stage log-log model. 1st stage model estimation results: estimated coefficients and corresponding t-statistics for the air travel demand forecasting.

	Coefficients	*t Stat*
Constant	10.5443	*503.6*
Pop (ε)	1.0145	*119.1*
Inc (θ)	1.3181	*30.5*
Special1 (μ)	0.5848	*42.7*
Special2 (π)	-1.3398	*-75.2*
RoadRail (τ)	-1.5550	*-28.5*
Generalised_Cost (κ)	-0.8735	*-54.5*
Adjusted R^2	0.6134	
Collinearity	No	
Num. Observations	22225	

Table 3. 2-stage log-log model. 2nd stage model estimation results: estimated coefficients and corresponding t-statistics for flight frequency forecasting.

	Coefficients	*t Stat*	*Coefficients*	*t Stat*
Constant	7.4913	*1152*	7.4920	*1158*
NodeDeg (α)	-0.1635	*-3.21*	-0.1368	*-5.13*
NodeWeight (β)	-0.0385	*-1.29**	- - -	- - -
CC (γ)	-0.0129	*-0.65***	- - -	- - -
ECV (δ)	0.1598	*4.38*	0.1044	*5.96*
ODdemand (Φ)	0.0501	*9.04*	0.0481	*8.85*
Hub1 (ρ)	0.0330	*1.98*	0.0233	*1.5****
Hub2 (σ)	-0.0431	*-3.40*	-0.0413	*-3.46*
Num_airprt_city (ϕ)	-0.0478	*-4.18*	-0.0444	*-3.95*
Fuel_price (υ)	-0.1571	*-9.17*	-0.1681	*-10.64*
Fltfreq_previousyr (ι)	0.8289	*168.10*	0.8271	*171.24*
Adjusted R^2	0.8384		0.8383	
Endogeneity	No		No	
Collinearity	No		No	
Num. Observations	9187		9187	

* significant at the 80.2% confidence levels.
**significant at the 48.1% confidence levels.
*** significant at the 85.9% confidence level.

B. Estimation of flight frequency as a function of observed airport-pair demand

A further analysis was also carried out to identify the value of modeling passenger routing. As described above, the first stage of the 2-stage log-log model estimates O-D passenger demand by city-pair. The second-stage, however, estimates airport-pair flight frequency, which is likely to be a stronger function of flight segment passenger flows by airport-pair than O-D city-pair demand. It is therefore important to identify the value in estimating airport-pair flight frequency as a function of airport-pair passenger flows. Therefore, observed flight segment passenger flows by airport-pair are extracted from Ref. 36, and are used as an explanatory variable in the second stage of the 2-stage log-log model (as in Eq. 6) instead of the estimated O-D demand predicted during the 1st stage of the 2-stage log-log model (Eq. 2.a). Note that these segment passenger flows include passengers travelling on only that flight segment as well as connecting passengers for which that segment constitutes one of the legs of their multi-stop trip.

Through the above analysis it is possible to evaluate whether or not a further step should be included in the flight frequency forecasting methodology. The objective of this intermediate step would be to transform the estimated O-D passenger demand by city-pair, extracted from the 1st stage of the process, into passenger segment demand, by airport pair. This could be done by allocating O-D passenger demand by city-pair to passenger itineraries, describing passenger routing, as described for the 3-stage model in §IV.B. This would then allow for the calculation of passenger segment demand by airport-pair. The resulting estimated airport-pair demand would then be included within the set of explanatory variables of the second stage of the 2-stage log-log model, as described above.

Table 4 shows the model estimation results for the modified flight frequency estimation using observed flight segment passenger flows. These results show a significant improvement in goodness of fit relative to the previous model, with the adjusted R^2 increasing from 0.84 to 0.89 – an increase of 5.3%. The values of the estimated coefficients are also more balanced, with less dominance from the auto-regressive term and network parameters, and increased impact from passenger demand (the *segment demand data* has an estimated coefficient of 0.35). Furthermore, all estimated coefficients are significant at the 95th percentile confidence level, show the expected sign, and there are no signs of endogeneity or multi-collinearity.

It is therefore clear that the use of flight segment demand by airport-pair in the estimation of flight frequency is superior to using O-D passenger demand by city-pair. Consequently, it appears that there would be significant benefit in adding a further step in the forecasting process that transforms the O-D passenger demand estimated in the first stage of the process into segment passenger demand by airport-pair.

Table 4. Model estimation results when considering T-100 segment demand data instead of the estimated *ODdemand.*

	Coefficients	*t-Stat*
Constant	7.5702	1667
NodeDeg (α)	-0.2373	-6.45
NodeWeight (β)	-0.4688	-21.23
CC (γ)	-0.0408	-2.88
ECV (δ)	0.4787	18.88
T-100 real segment demand data (Φ)	0.3504	62.16
Hub1 (ρ)	0.0436	3.75
Hub2 (σ)	-0.1431	-15.45
Num_airprt_city (ϕ)	-0.0629	-9.37
Fuel_price (υ)	-0.0504	-4.02
Fltfreq_previousyr (ι)	0.5841	110.63
Adjusted R^2	0.8857	
Endogeneity	No	
Collinearity	No	
Num. Observations	11295	

C. Second approach: 3-stage model

From the comparison of the results obtained from the 2-stage log-log model and the case when observed flight segment demand by airport-pair is used, it is clear that further improvement could be achieved by introducing a further step to the forecasting process, resulting in a 3-stage model. The aim of this intermediate step is to transform O-D passenger demand estimated in the first stage of the process into segment passenger demand by airport-pair.

At the 1st stage of this 3-stage model Eq. (2.a) is used, and therefore results from Table 2 apply. Similarly, for the 3rd stage, in which flight segment frequency is predicted using segment passenger demand, Eq. (6) applied, which is a modified Eq. (2.b) – i.e. using segment passenger demand instead of O-D demand –. The intermediate step, as described above, consists of a classification algorithm applied to identify the available itineraries and a discrete choice model to assign the passenger share across these identified available itineraries. Table 5 presents the model estimation results for the logistic regression obtained by applying Eq. (3). These results show all estimated coefficients significant at the 95th percentile confidence level, and no collinearity is observed. However, two causation problems can be identified. Firstly, the estimated coefficient of Population *(Pop)* is negative. One would expect the population coefficient to be positive since higher numbers of inhabitants associated with a city-pair would be expected to correlate to flight segments that exist. A similar case is observed with the *Hub2* variable. *Hub2* is 0 if both airports are not a hub and 1 otherwise, so a positive correlation is expected.

To evaluate the performance of the logistic regression model, the confusion matrix is computed along with sensitivity and specificity tests. Sensitivity is also known as True Positive Rate (TPR) while specificity is also known as True Negative Rate (TNR). Table 6 shows the results of these tests. Results obtained for the model's TPR and TNR have high values of 89.24% and 89.29% respectively, which is good.

Table 5. 3-stage model. 2nd stage model estimation results: estimated coefficients and corresponding t-statistics for logistic regression model.

	Coefficients	t-stats
Constant	-1.8085	-13.93
NodeDeg (α)	2.0754	15.83
NodeWeight (β)	1.5081	7.89
CC (γ)	0.2195	5.04
EVC (δ)	0.5364	2.99
APdemand	0.3176	13.81
Pop (ε)	-0.6640	-15.34
Special1 (μ)	0.3758	7.16
Special2 (π)	-0.4314	-5.49
Hub1 (ρ)	-1.1467	-2.20
Hub2 (σ)	-1.3016	-26.61
Fuel_price (ν)	-0.6895	-7.84
Collinearity	No	
N of observations	24692	

Table 6. 2nd stage model estimation results: Confusion matrix, True Positive Rate and True Negative Rate for logistic regression model.

		Predicted		**Percentatge correct**	
		Connected	Unconnected		
Observed	Connected	11025	1321	**TPR**	89.24%
	Unconnected	1329	11017	**TNR**	89.29%

The second part of the intermediate stage – i.e. the discrete choice model – is under development and the model estimation results have not been completed at the time of writing. However, some descriptive parameters for the 18 entities have been computed. This data summary is presented in Table 7. Note that 'Available Itineraries' refers to the number of itineraries available to any traveller between the regions described by the entity. These itineraries are typically a non-stop flight from the traveller's origin to destination, plus a number of connecting itineraries through a set of hub airports. The set of hub airports varies between entities, as some hubs do not make sense for geographical reasons. As mentioned in §IV.B the number of alternative itineraries for each entity is a subset of the universal choice set and varies across the entities. 'Number of Observations' shows the busiest flows in the US ATS network, i.e., the East Coast corridor (East-East entity), the West Coast corridor (West-West entity), between the Mid-West and Each Coast (Central-East and East-Central entities), and between the East and West Coasts (East-West and West-East entities).

Table 7. Data summary for the 18 entities used in the discrete choice model.

Entity	Available Itineraries	Num. Observations
Alaska&Hawaii - Continental US	19	23,901,990
Continental US-Alaska&Hawaii	25	23,809,500
Central-Central	23	118,972,940
Central-East	24	179,885,310
Central-Mountain	26	24,509,400
Central-West	26	88,268,740
East-Central	26	179,337,760
East-East	20	491,361,300
East-Mountain	26	29,771,100
East-West	26	128,209,700
Mountain-Central	26	24,650,610
Mountain-East	26	29,930,000
Mountain-Mountain	12	5,846,370
Mountain-West	13	36,508,810
West-Central	26	88,802,810
West-East	26	129,306,700
West-Mountain	20	36,415,180
West-West	10	201,024,320

VII. Conclusions and Future Work

Research described in this paper provides an initial effort to improve on existing air traffic forecasting methodologies through a better understanding of the factors driving demand, supply and network dynamics. In order to achieve this, three enhancements are being pursued: the use of data mining techniques in air traffic forecasting; and the use of a larger range of explanatory variables not considered in existing approaches, and explicitly modeling the distribution of city-pair passenger demand between itineraries. These enhancements are applied to identify the different influences underlying the US ATS evolution.

Initially, results obtained from the 2-stage log-log model show high goodness of fit. No endogeneity and multi-collinearity issues were detected and the estimated coefficients are all of the expected sign, suggesting no causation problems. However, four variables – the auto-regressive term and network parameters – dominate, with significantly larger coefficients compared to the rest of variables. This is a cause of concern given the limited influence of other parameters, such as passenger demand, which is expected to have a large influence on predicting future flight frequency. From the comparison of the results obtained from the 2-stage log-log model and the case when observed flight segment demand by airport pair is used instead of predicted city pair demand, it is clear that further improvement could be achieved by introducing a further step to the forecasting process. The aim of this step would be to transform O-D passenger demand estimated in the first stage of the process into segment passenger demand by airport-pair. Considering this intermediate step, this paper introduces a 3-stage model.

Currently, this 2nd stage of the 3-stage model is under development, involving the distribution of city-pair passenger demand between itineraries and the evolution in the US ATS network. This is achieved in two steps. First, the existence of potential itineraries for any given O-D city-pair is estimated by using a classification algorithm. Then, a discrete choice model is applied to estimate the passenger itinerary share across the previously identified available itineraries. Results for the first step of the process, which estimates a logistic regression model as classification algorithm, look promising, with an overall percentage of correct classifications of 89.25%. Initial development of the second step of the process, the discrete choice model, is still in progress.

Model estimation results obtained to date look promising. However, there is room for improvement and further work is planned. A line of work from which improvements could be achieved is the application of a feature extraction process prior to the training of the algorithm. Specifically, clustering US airports with similar properties is being considered to improve the model performance.

Other machine learning techniques than regression will also be considered for each of the models described. Those under consideration include neural networks using various learning algorithms such as backpropagation and backpropagation through time (BPTT). In these approaches a feature extraction technique could also be applied as a previous step, to reduce computational complexity. Also, other model structures for the discrete choice model, such as nested-logit models, are considered.

The best performing model will be used to predict air traffic in the US ATS into the future, so that the results can be compared directly to the TAF. In this way, the benefit of improved air traffic forecasting will be identified.

Acknowledgments

The authors would like to gratefully acknowledge Dr. Lynnette Dray and Dr. Maria Kamargianni from University College London for their advice on data sources and approach.

References

[1] Schäfer, A., "Long-term trends in global passenger mobility," *The Bridge, Linking Engineering and Society,* The National Academies Press, Vol. 36, No.4, Washington, D.C., 2006, pp. 24-32.

[2] Henckels, E., 2011. “Airline Industry Overview,” *Columbia Graduate Consulting Club* [online database], URL: http://www.columbia.edu/cu/consultingclub/resources.html [cited 21 October, 2014].

[3] Airbus, “Future Journeys. Global Market Forecast 2013-2032,” URL: http://www.airbus.com/company/market/forecast/ [cited 21 October 2014].

[4] FAA, “FAA Aerospace Forecast: Fiscal Years 2012-2032,” March 2012, URL: https://www.faa.gov/about/office_org/headquarters_offices/apl/aviation_forecasts/aerospace_forecasts/2012-2032/media/2012%20FAA%20Aerospace%20Forecast.pdf [cited 21 October 2014].

[5] ATAG, “Aviation Benefits Beyond Borders,” April 2014, URL: http://aviationbenefits.org/media/26786/ATAG__AviationBenefits2014_FULL_LowRes.pdf [cited 21 October 2014].

[6] Intergovernmental Panel on Climate Change, “IPCC Special Report – Aviation and the Global Atmosphere, Working Group I and III,” IPCC, Geneva, Switzerland, 1999.

[7] Airports Commission, “Airports Commission: Interim Report,” December 2013, URL: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/271231/airports-commission-interim-report.pdf. [cited 21 October 2014].

[8] Boeing, “Current Market Outlook 2013-2032,” 2013, URL: http://search-www.boeing.com/search?q=Current+Market+Outlook+2013-2032%E2%80%9D.+&site=www_boeing&client=www_boeing&proxystylesheet=www_boeing&output=xml_no_dtd&btnG.x=0&btnG.y=0 [cited 21 October 2014].

[9] FAA, “Forecast Process for 2013 TAF”, URL: https://aspm.faa.gov/main/taf.asp [cited 21 October 2014].

[10] Cumming, S., “Data Mining at British Airways,” Royal Statistical Society, Reading Local Group, February 2005, [PowerPoint slides] URL: http://www.reading.ac.uk/RSSlocal/Synopses/2005february.html [cited 7 November 2014].

[11] Smalley, E., “NASA Applies Text Analytics to Airline Safety,” Data Informed, July 30, 2012, URL: http://data-informed.com/nasa-applies-text-analytics-to-airline-safety/ [cited 23 October 2014].

[12] Srivastava, A., “NASA Chat: Data Mining Digs Up Clues to Safer Flights,” NASA Ames Research Center, 2011, URL: http://www.nasa.gov/connect/chat/data_mining_chat.html#.U-owc_ldWSo [cited 23 October 2014].

[13] Swan, W. M., “Forecasting Air Travel with Open Skies,” Chief Economist Airline Planning Group, 2008.

[14] ICAO, “Report of the Asia/Pacific area traffic forecasting Group (APA TFG),” 16th meeting, Montreal, 19-21 September 2012.

[15] Teyssier, N., “How the consumer confidence index could increase air travel demand forecast accuracy?,” PhD dissertation, Air Transport Group, Cranfield University, UK, September 2012.

[16] Department for Transport (DfT), “UK Aviation Forecasts,” January 2013, URL: https://www.gov.uk/government/publications/uk-aviation-forecasts-2013 [cited 21 October 2014].

[17] Eurocontrol, “Challenges of Growth 2013: Task 4: European Air Traffic in 2035,” 2013, URL: https://www.eurocontrol.int/articles/challenges-growth [cited 21 October 2014].

[18] Garvett, D. S., and Taneja, N. K., “New directions for forecasting air travel passenger demand,” Department of aeronautics and astronautics, Flight transportation laboratory, Cambridge, Massachussets 02139, 1974.

[19] Pindyck, R. S., and Rubinfeld, D. L., “Econometric Models and Economic Forecasts,” 4th Edition, Irwin, McGraw-Hill, Unites States of America, 1998.

[20] Viken, J., Dollyhigh, S., Smith, J., Trani, A., Baik, H., Hinze, N., and Ashiabor, S., "Utilizing Traveler Demand Modeling to Predict Future Commercial Flight Schedules in the NAS," *11th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*, Virginia, 2006.

[21] FAA, “Terminal Area Forecast Summary: Fiscal Years 2013-2040”, 2014, URL: https://www.faa.gov/about/office_org/headquarters_offices/apl/aviation_forecasts/taf_reports/ [cited 21 October 2014].

[22] Evans, A. D., and Schäfer, A. W., “Simulating airline operational responses to airport capacity constraints,” *Transport Policy*, vol. 34, 2014, pp. 5-13.

[23] Adler, N., “Hub-Spoke Network Choice Under Competition with an Application to Western Europe,” *Transportation Science*, V. 39, 2005, pp58-72.

[24] Coldren, G. M., and Koppelman, F. S., "Modeling the competition among air-travel itinerary shares: GEV model development," *Transportation Research Part A: Policy and Practice* 39.4 (2005): 345-365.

[25] Nam, K., and Schaefer, T., "Forecasting international airline passenger traffic using neural networks," *Logistics and Transportation Review 31.3*, 1995.

[26] Cheng, T., Cui, D., and Cheng, P., “Data Mining for Air Traffic Flow Forecasting: A Hybrid Model of Neural Network and Statisticl Analysis,” *Intelligent Transportation Systems, Proceedings*, Vol. 1, IEEE 2003, pp. 211-245.

[27] Bao, Y., Xiong, T., and Zhongyi, H., “Forecasting Air Passenger Traffic by Support Vector Machines with Ensemble Empirical Mode Decomposition and Slope-Based Method,” *Discrete Dynamics in Nature and Society*, Vol. 2012, 2012, Article ID 431512.

[28] Kotegawa T., “Analyzing the evolutionary mechanism of the air transportation system-of-system using network theory and machine learning algorithm.” PhD dissertation, Faculty of Purdue University, West Lafayette, Indiana, 2012.

[29] Guimera, R., Mossa S., Turtschi, A., and Amaral, L. A. N., “The worldwide air transportation network: Anomalous centrality, community structure, and cities’ global roles,” *Proceedings of the National Academy of Sciences (PNAS)*, Vol. 102, May 31, 2005, no.22.

[30] Benoit, K., “Linear Regression Models with Logarithmic Transformations,” Methodology Institute London School of Economics, March 17, 2011, URL: http://www.kenbenoit.net/courses/ME104/logmodels2.pdf [cited 24 October 2014].

[31] Dray, L. M., Evans, A. D., Reynolds, T., Rogers, H., Schäfer, A., and Vera-Morales, M., “Air Transport Within An Emissions Trading Regime: A Network-Based Analysis of the United States and India,” *TRB 88th Annual Meeting*, Washington DC, 11-15 January 2009.

[32] Belsley, D., Kuh, E., and Welseh, R.E “Regression Diagnostics: Identifying Influential Data and Sources of Collinearity,” Chapter 3, John Wiley & Sons, Inc, New York, 1980.

[33] Aviation Integrated Modelling (AIM) Project, URL: http://www.aimproject.aero/ [cited 27 October 2014].

[34] US Census Bureau, 2014. “Annual Estimates of the Population of Metropolitan and Micropolitan Statistical Areas: April 1 2000 to July 1 2009,” Vintage 2009: Metropolitan and Micropolitan Statistical Areas Tables (CBSA-EST2009-01) [online database], URL: http://www.census.gov/popest/data/historical/2000s/vintage_2009/metro.html [cited 17 September 2014].

[35] Bureau of Economic Analysis, US Department of Commerce, “CA1-3 Personal income summary: 2003 to 2007,” [online database] URL: http://bea.gov/iTable/iTable.cfm?reqid=70&step=1&isuri=1&acrdn=2#reqid=70&step=25&isuri=1&7022=20&7023=7&7024=non-industry&7001=720&7029=20&7090=70 [cited 17 September 2014].

[36] Bureau of Transportation Statistics - Research and Innovative Technology Administration (BTW-RITA), “T-100 Domestic Segment (U.S. Carriers): 2003 to 2007,” [online database] URL: http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=259&DB_Short_Name=Air%20Carriers [cited 13 August 2014].

[37] Bureau of Transportation Statistics - Research and Innovative Technology Administration (BTS-RITA), “Origin and Destination Survey: DB1BMarket for 2003 to 2007,” [online database] URL: http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=247&DB_Short_Name=Origin%20and%20Destination%20Survey [cited 17 September 2014].

[38] Bierlaire, M., “Biogeme: A free package for the estimation of discrete choice models,” *Proceedings of the 3rd Swiss Transportation Research Conference*, Ascona, Switzerland (2003).

[39] FAA, “Aviation System Performance Metrics (ASPM) Manuals”. URL: http://aspmhelp.faa.gov/index.php/ASPM_Manuals#User_Manuals [cited 20 September 2014].

[40] Bureau of Labour Statistics, US Department of Labor, “Annual Average Indexes 2007 (Tables 1A-23A),” URL: http://www.bls.gov/cpi/cpi_dr.htm#2007 [cited 20 September 2014].

[41] Bureau of Transportation Statistics - Research and Innovative Technology Administration (BTS-RITA), "Airline Fuel Cost and Consumption (US Carriers – Scheduled Service)," [online database] URL: http://www.transtats.bts.gov/fuel.asp?pn=1 [cited 20 September 2014].