City Research Online

City, University of London Institutional Repository

Citation: Guimera Busquets, J., Alonso, E. and Evans, A. (2018). Air itinerary shares estimation using multinomial logit models. Transportation Planning and Technology, 41(1), pp. 3-16. doi: 10.1080/03081060.2018.1402742

Air Itinerary Shares estimation using Multinomial Logit Models

Ms. Judit Guimera Busquets[1]

Department of Mechanical Engineering & Aeronautics, City, University of London, EC1V 0HB London, United Kingdom

Dr. Eduardo Alonso

Department of Computer Science, City, University of London, EC1V 0HB London, United Kingdom

Dr. Antony D. Evans

Crown Consulting, Inc., Moffett field, CA 94035, USA

[1] Corresponding author: Judit Guimera Busquets

Email: Judit.Guimera-Busquets.2@city.ac.uk

Air Itinerary Shares estimation using Multinomial Logit Models

The main goal of this study is the development of an aggregate air itinerary market share model. In order to achieve this, multinomial logit models are applied to distribute the city-pair passenger demand across the available itineraries. The models are developed at an aggregate level using open-source booking data for a large group of city-pairs within the US Air Transport System. Although there is a growing trend in the use of discrete choice models in the aviation industry, existing air-itinerary share models are mostly focused on supporting carrier decision-making. Consequently, those studies define itineraries at a more disaggregate level, using variables describing airlines and time preferences. In this study, we define itineraries at a more aggregate level, i.e., as a combination of flight segments between an origin and destination, without further insight into service preferences. Although results show some potential for this approach, there are challenges associated with prediction performance and computational intensity.

Keywords: word; air itinerary shares; discrete choice models; multinomial logit; aggregation level;

1. Introduction

Good forecasts of future demand for air traffic as well as good forecasts of how airlines are likely to serve this demand are essential to enable supply to adapt to growth in demand. While the majority of existing research focuses on improving air travel demand models, there is a growing interest in developing better itinerary share models than those that already exist. Itinerary share models can be crucial to support airline network planning and scheduling since important decisions on resources allocation and pricing are made based on itinerary demand. These decisions are essential as airlines plan their operations, purchase equipment and make strategic decisions. Airport authorities also benefit from good forecasts, given the long timescales associated with airport development and capacity expansion. Improving the accuracy of itinerary share models is therefore a powerful tool for airline and airport authority planning and decision making, translating into more efficient operations, improved revenue management and increase profitability. Consequently, for the past 15 years, efforts have focused on developing this type of model, shifting away from the Quality of Service indices (QSI) used during the period when the industry was regulated, and/or more simplistic approaches – such as time-series and simplistic probability models based on historical trends – (Garrow, 2010). In contrast, discrete choice models model demand by capturing *how* individuals make decisions and trade-offs among airports, airlines, price, level of service and other factors that define the air passenger journey.

Most of the current research centres on developing innovative approaches using such discrete choice modelling. These approaches, which aim to model competition and customer behaviour to determine air-travel itinerary shares (also known as demand assignment models), are expected to more accurately predict air travel demand. While most of the discrete choice models applied in urban transport are built using disaggregate data and include information about the individual making the decision – i.e. the passenger –; in air transport, data disaggregation as well as data accessibility are limiting factors. The need to quickly adapt to changes in demand makes flexibility crucial for carriers and other stakeholders in the industry. For this reason, most of the models built to support decision-making rely on booking data, which is generally proprietary. Furthermore, airlines do not typically record much of the passenger data that is relevant to passenger decision making, such as age, gender and income. This data is not typically available, except for a small subset of passengers through surveys, which are time consuming and costly to complete.

Most of the early studies on demand assignment for air travel focus on studying the distribution of demand across one single dimension, i.e. only focusing on modelling passenger choice in terms of one single criteria, such as airport-choice or airline choice. These early models were mostly applied to analyse air travellers' choice within multi-airport cities or regions – i.e., airport choice models (Hansen, 1995; Windle & Dreesner, 1995) – or across airlines – airline choice models (Proussaloglou & Koppelman, 1995) –. Although the former is the most widely studied topic in discrete choice modelling within air transport, and has given a deeper understanding to the relationship between airport attributes and airport market share, a more aggregated assignment of air travel volume is also needed. Only a few studies present approaches for itinerary market share estimation across multiple dimensions (i.e., modelling a passenger's simultaneous choice in terms of multiple criteria, e.g., airline, flight time, fare-class etc.) using discrete choice modelling. Of those, early models used a multinomial logit (MNL) approach (Adler, 2001; Coldren *et al.*, 2003; Grosche and Rothlauf, 2007; Atasoy and Bierlaire, 2012), while more recent models also apply nested logit (NL) models (Coldren and Koppelman, 2005; Hsiao and Hansen, 2011), mixed multinomial logit (MMNL) models (Warburg *et al.*, 2006) and other alternatives approaches (Gramming *et al.* 2005; Carrier, 2008). The mentioned aggregate passenger-allocation studies can be classified according to the type of data they are based on: revealed preference data (RP) or booking data; stated preferences (SP) data or survey data; or a combination of both. Studies using RP data do not usually provide full insight into passenger choice behaviour since models are estimated based on real booking data, and no information regarding other alternatives at the moment of booking is generally available. This limitation often leads to RP models performing poorly due to the high demand inelasticity of the booking data used to estimate the model (Garrow, 2010). In contrast, SP data collected from surveys allows for modelling of new non-existing alternatives, as well as more accurate estimation of the sensitivity of travellers to characteristics of their trips. However, studies using SP data may be subject to bias due to the nature of the experiment in which the individuals are asked to make hypothetical choices by making trade-offs among the attributes of the choice set (e.g., level of service, air fare etc.). Although such surveys provide a customer response to a wider range of choices, providing a better estimate of how individuals make tradeoffs, they are tailored to the needs of the survey writer, which limits the natural range of choice sets to only those that the survery writer is aware of (Garrow , 2010; Louviere *et al.*, 1999). Studies based on SP data are also often limited to a small range of markets, limiting their application to a small network set.

Although the models applied in the studies described above are generally effective for the purposes to which they are applied, they do not allow for an estimation of how passenger market demand is distributed across the available itineraries at the most aggregate level, only considering average market air fare and travel time, level of service and basic airport attributes as inputs.

This paper presents the full air itinerary share model introduced by Busquets *et al.* (2016), refined to better capture passenger choice effects, model validation, and estimated at the most aggregate level possible, linking annual city-pair demand to the different itineraries available within the entire US Air Transport System (ATS).

The remainder of the paper is structured as follows: The paper’s objectives are presented in Section 2. The modelling approach is detailed in Section 3, with information regarding the input variables used to estimate the model. The model is estimated on one dataset, and validated on another. Section 4 provides information about these two datasets. Modelling results are presented in Section 5, followed by the model validation results in Section 6 and a discussion on future work in Section 7.

2. Objectives

The primary objective of this research is to develop an air itinerary choice model to directly estimate the distribution of passenger demand across available routes for a given O-D pair, using only aggregate data describing average air fare and travel time, level of service and basic airport characteristics. Ultimately, this model will be combined with models for forecasting air travel demand and air traffic, all within the same 3-stage framework (described in Busquets *et al.*, (2015)). This framework consists of the following stages:

(1) Forecast city-pair passenger demand;
(2) Distribute this demand across available itineraries; and
(3) Forecast air traffic as a function of route demand.

This modelling approach is inspired by previous research that focused on improving the Federal Aviation Administration's (FAA) forecasting methodology and for which further potential improvements have been identified. The 3-stage framework is expected to allow for identification of the key drivers of evolution in the US ATS as well as to predict future air traffic growth within the US ATS. In order to achieve these objectives, the approach includes three elements beyond that of the existing research:

- The use of data mining techniques to model the US ATS evolution in order to predict air traffic with improved accuracy and precision levels while maintaining the simplicity if existing econometrics, 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 modelling the distribution of city-pair passenger demand between itineraries.

This paper addresses the last of these three elements, which develops the framework's stage 2 – to distribute passenger demand across available itineraries. The approach described in this paper is therefore expected to:

- Highlight the most important factors underlying the air traveller's choice behaviour within the domestic US ATS;
- Perform air itinerary share model refinement and verification for the entire US ATS following previously work (Busquets *et al.*, 2016); and
- Explicitly model the distribution of city-pair passenger demand between itineraries within the US ATS.

The model presented in this paper is expected to generate better predictions of airport-pair air traffic flows once integrated with the air traffic demand model presented by Busquets *et al.*, (2015).

3. Approach

Data

Based on the literature review, there are a large number of factors that describe an itinerary. An itinerary, as defined in this paper, is a flight segment or combination of flight segments connecting a given city-pair. In this study, itineraries are either non-stop, or one-stop (i.e., a combination of two flight segments involving an aircraft change during the connection). Considering constraints in data availability and the different attributes that are considered to contain the most relevant information for an itinerary, the input variables for the itinerary market share model are chosen as described in Table 1.

[Table 1]

The output variable for the model developed in this paper is the market share (S_i) of a given itinerary i. This is defined as the ratio of the demand of the itinerary i (d_i), to the total demand for the market served by itinerary i (D_m), as shown in Eq. (1). The total demand for market m is given by the sum of passengers travelling on all itineraries that serve that market.

$$S_i = \frac{d_i}{D_m} \tag{1}$$

Detailed Forecasting Methodology

Following the work presented by Busquets *et al.* (2015), which introduced the 3-stage model described in §2 to forecasting future air traffic levels, this paper focuses on fully developing its stage 2 – to distribute passenger demand across available itineraries. The objective of this phase is therefore to transform Origin-Destination (O-D) demand by city-pair into passenger demand by airport-pair using an air itinerary choice model.

Stage 2 of the 3-stage model described by Busquets *et al.* (2015) consists of 2 steps: identification of available itineraries estimated using logistic regression (described in detail in Busquets *et al.* (2015)), followed by the distribution of the O-D demand by city-pair obtained from the O-D demand model (stage 1 in the 3-stage model described by Busquets *et al.* (2015)) across the available itineraries using a discrete choice model. The first step is motivated by the scope of this research to improve the current FAA's forecasting methodology while maintaining the simplicity of current models and is inspired by a previous research (Kotegawa, 2012). The second step is the focus of this paper. This air itinerary model allows the flight segment passenger demand by airport-pair to be estimated, based on the passenger itinerary demand from all O-D city-pairs. It is not feasible to develop a model for each possible O-D market, so in order to apply the discrete choice model, the US is divided into five regions, as done by

Coldren, et al. (2003): four Continental time zones (Central, East, Mountain and West) and a region for Alaska and Hawaii. This specific O-D market grouping is an attempt to capture similarities among all city-pairs. The number and nature of these regional clusters will be modified using clustering techniques in future work. Given these regions, 18 region-pairs 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), etc., West-Mountain (W-M), West-West (W-W) –; as well as a region-pair for Alaska and Hawaii to the Continental US and an region-pair for the Continental US to Alaska and Hawaii. For each region-pair, henceforth referred to as an 'entity', an air itinerary share model is developed.

This attempts to model the aggregate share of all or groups of decision makers - i.e., air travellers - choosing each alternative as a function of the trip characteristics. In constrast to existing research, the itinerary share estimation is done at the most aggregate level, without considering variables specific to the traveller, such as passenger preferences and perceptions, or variables specific to the service provider, such as airline operating the given route, departure time or aircraft type, among others. Instead, only attributes related to average air fare and travel time, level of service and basic airport characteristics are considered. The focus of the model is to estimate the distribution of annual passenger market demand among itineraries, which will be used as one of the input variables in the third stage of the air traffic estimation model described in §2, per annum.

In order to develop the air itinerary share model, RP data is used, avoiding the risk of response bias and allowing for the consideration of a much larger network of city-pairs within the US ATS. The RP data used is 10% ticket survey of booking data from airlines operating within the US domestic market (BTS-RITA, 2003-2010). The city-pairs considered, M, are all within the domestic US ATS and are defined by origin and destination. The universal choice set, C, is formed for all possible itineraries within the entire ATS connecting these city pairs. The choice problem is defined for each city-pair, $m \in M$, with the choice set being all the possible itineraries connecting that given city-pair, represented by I_m. Each itinerary i is characterised by a set of attributes such as level of service, price, time and basic airport characteristics. As a simplification, only two possible levels of service are considered, non-stop and one-stop flights. For the one-stop flights, the connections available are through one of a set of 24 US hub airports defined for this study.

The annual share of passenger demand assigned to each itinerary between a given city-pair is modelled as an aggregate multinomial logit (MNL) function and is given by Eq. (2) 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 city-pair. The utility function (V_i) is a linear function of the explanatory variables and assumes that each vector of attributes characterizing an alternative can be reduced to a scalar value, which expresses the attractiveness of each alternative. Consequently, it is expected that the individual or group of individuals will choose the alternative with the highest value, maximizing their utility. Equation (3) shows the general expression for V_i, where X_i is the vector of attributes defining alternative i; and β' represents the coefficients to be estimated capturing the influence of the corresponding attribute on the alternative i (Atasoy & Bierlaire, 2012).

$$S_i = \frac{Exp(V_i)}{\sum_j exp(V_j)} \quad (2)$$

$$V_i = \beta' \cdot X_i = \beta_1 \cdot X_{i1} + \beta_2 \cdot X_{i2} + \ldots + \beta_k \cdot X_{ik} \quad (3)$$

Attributes included in the X_i vector are described in Table 1 (§3). Some interactions between the attributes are accounted for by the model. After evaluating several model specifications, the interactions that define the utilities considered in this paper were identified as follows:

- Accessibility: The interaction between airport accessibility information and multi-airport city information is accounted for (i.e., the *masORIG* and *masDEST* variables). Four possible interactions are possible, two regarding the origin airport and two regarding the destination airport. However, because coefficients need to be normalised, the coefficients regarding accessibility for origin and destination airports within cities that are not multi-airport systems are set to 0.
- From/to hub variables: The interaction between the hub variables (i.e., whether the itinerary is from and to a hub, only the origin or destination airport is a hub, or none of the itinerary airports are hubs) and markets that contain at least one non-stop itinerary is considered. From/to hub variables are normalised by setting the variable from and to a hub (i.e., the *hub2hub* variable) to 0.

During the estimation of the model, for each city-pair considered, the utility and likelihood function are computed, with the latter being used to calculate the final estimated log likelihood.

Although all 18 air-itinerary share models have been developed, in this paper estimated results are only presented for six entities (the entities C-M, M-C, C-W, W-C, M-W and W-M). Due to issues with computational intensity during the estimation process for some entities, reduced estimation datasets were generated by sampling a subset of the total number of city-pairs within the given entity. The size of the reduced estimation datasets was chosen after evaluating preliminary model estimation results obtained when considering different estimation dataset sizes. Due to the aggregate nature of the data used in this study and the fact that this data represents only a 10% sample of real booking data, limiting assumptions are implicitly included when estimating the model. For example, some itineraries have a very small probability of occurring, heavily influencing the results obtained for the model estimated as well as its performance. Moreover, due to the large number of city-pairs considered in the estimation data and the large number of coefficients to be estimated, the model estimation becomes computationally too intensive. For these reasons, the data is reduced to 10 datasets containing information on 100 randomly chosen city-pairs, which are then each used to estimate the model, reducing the complexity of the problem. The final estimated model coefficients are computed as the average of the 10 different models. The performance of each of the entities' air itinerary share model is validated with data not used for the model estimation. Table 2 reports summary statistics for all the entities. The set of hub airports varies between entities, as some hubs do not make sense for some entities for geographical reasons. Table 2 shows the busiest flows in the US ATS network, i.e., the East Coast corridor (East - East entity), the Central corridor (Central – Central entity) and between the Central region and East Coast (Central-East and East-Central entities). A total of 17,200 city-pairs and 104,806 itineraries within the US ATS network are accounted for in the development of the air itinerary share models.

To better understand the results obtained from the air itinerary share model, indicators such as passenegers' 'willingness to pay' can be computed. Value of time (VOT) is the willingness of passengers to pay for one hour of travel and is defined by Eq. (4), which is computed for each given itinerary *i*. Note that because *Travel Fare Ratio* is a function of the average air fare in the market and *Travel Time Ratio* is a function of the minimum flight time possible in the market, when computing the utility

V_i, average air fare ($\overline{TF}$) and minimum flight time (TT_{sh}) are also included in the formulation of VOT.

$$VOT_i = \frac{\partial V_i/\partial time_i}{\partial V_i/\partial price_i} = \frac{\beta_{FlightTimeRatio}}{\beta_{AirFareRatio}} \cdot \frac{\overline{TF}}{TT_{sh}} \qquad (4)$$

[Table 2]

Once the itinerary choice model is estimated using the MNL function, Eq. (1) is applied to compute the market share of passengers on each itinerary. The estimated passenger demand per itinerary is then used to compute segment demand – i.e., passenger demand per airport-pair – which will ultimately be used as an input for stage 3 of the 3-stage model described in §2, as described in detail by Busquets *et al.* (2015).

4. Application

The models described above are applied to a network of 337 airports within the US ATS, as used in the Aviation Integrated Modelling (AIM) Project (2006). The choice of the US air transport network is motivated by improving the current FAA's forecasting methodology, and by the availability of data. The availability of data for the analysis of air transport systems can be challenging, with the US being one of the few countries to provide open source data.

The RP data used for this study includes passenger demand data and airfares extracted from the Airline Origin and Destination Survey (DB1B) (BTS-RITA, 2003-2010), which contains a 10% sample of airline tickets from reporting carriers. Travel times and costs are also extracted from BTS-RITA (2003-2010). The air itinerary choice model is estimated using Biogeme (Bierlaire, 2003). Flight delay information is obtained from the FAA Aviation System Performance Metrics (ASPM) database (FAA, 2007-2010).

The RP data considered for estimating the model is from 2007, to be in line with the period considered when estimating the ultimate 3-stage model described by Busquets *et al.* (2015). The data used to validate the model is from 2010.

Once the model is estimated, it will be applied in future work to estimate the itinerary shares in the same network of 337 airports into the future. These results will then be compared to those of the Terminal Area Forecasts (TAF) produced by the FAA.

5. Model Estimation Results

Parameter estimates for the six air itinerary share models mentioned above are reported in Table 3 below. From the entities shown, parameters for the C-W and W-C entities are estimated using 10 different folds of 100 randomly selected city-pairs. The estimated coefficients are averaged to define the final model coefficients. For the C-M, M-C, M-W and W-M entities, the entire estimation dataset is used to estimate the air itinerary share model. As Table 2 shows, the C-W and W-C entities have 724 city-pairs and just over 5,200 itineraries, while the rest of the entities' datasets reported in this paper contain a much lower number of city-pairs and itineraries, making the estimation process less computationally intensive.

Model performance is described using the likelihood ratio test and rho-squared parameter (ρ^2). The likelihood ratio test provides an evaluation of the entire estimated model by evaluating whether it is possible to reject the null hypothesis that a more restricted model (i.e., a model with zero coefficients) is equal to the estimated one. The ρ^2 metric is an indicator of overall goodness of fit.

All estimated coefficients are statistically significant at the 95th percentile confidence level.

The *Travel Fare Ratio* and *Travel Time Ratio* coefficients are both of the expected sign, negative, indicating that fares and travel time are a resistance to travel. In contrast, some of the coefficients associated with delay at the origin and destination airports are positive, suggesting a correlation between delay and itinerary attractiveness, which is unexpected. For entities C-M, M-C, M-W and W-M, the sign of the coefficients alternates between positive and negative, indicating a positive correlation between delay and itinerary attractiveness associated with Mountain (M) airports. For the C-W and W-C entities both delay parameters are positive. These results may be an indication of airport importance since larger and/or hub airports are expected to have more passengers and flights, and therefore higher delay. This suggests that passengers are more inclined to travel to and from large airports, which is likely because of the increased number of routing alternatives available at these airports.

The coefficients associated with airport accessibility are also positive, with the exception of the *AccessDEmas* coeffient for the C-M entity and the *AccessORmas* coeffient for the W-C entity. This is opposite to what one would expect since an increased travel time to/from an airport is a resistance to air travel, and given the influence on door-to-door travel time, a negative sign would be expected. However, the coefficients associated with all airport accessibility time variables are small, - with the exception of the *AccessORmas* coefficients for the M-C and M-W entities -, indicating low influence of passenger preferences on itinerary choice.

[Table 3]

The estimated *Airline Ratio* coefficients tend to be in the order of 10e-2 and positive - with the exception of the coefficient associated with the C-W entity -, indicating low influence of passenger preference on itinerary choice. Coefficients associated with level of service are represented by dummy variables in the models and are characteristics of every entity. These variables show the passengers preference in terms of level of service and connecting hub choice. Due to the fact that each entity has a specific set of hubs and different assumptions have been made in building the connection alternatives, a comparison of the estimated coefficients across entities is not possible.

For the variables associated with origin and destination hub information (*1hub* and *no_hub*), both coefficients are generally negative, except for the C-W and W-C entities. One would expect a negative correlation between itinerary attractiveness and traveling from or to a hub airport (i.e., *1hub*=1), and also between itinerary attractiveness and travelling from and to a non-hub airport (i.e., *no_hub*=1). In both cases fewer alternatives would exist than for an itinerary between two hubs. The positive correlation for entities C-W and W-C may be because these sets of variables interact only with itineraries belonging to markets in which non-stop options exist, and itineraries from or/and to a non-hub airport may be associated with lower delay as well as lower travel fare ratio than itineraries from and to a hub.

Regarding the model performance, both the likelihood ratio test and rho-squared parameters for the six entities show reasonable goodness of fit. Although all the models show a likelihood ratio test large enough to reject the null hypothesis that all coefficients are equal to zero; rho-squared values tend to be largest for those models for which the entire dataset has been used during estimation. While the C-M, M-C, C-W and W-C entities have a rho-squared value of about 0.7; the rho-squared values for the

C-W and W-C entities are lower than 0.6. The same trend is found for the other air itinerary models estimated.

To further analyze the results and understand the effect that the level of service has on the willingness to pay, VOT is computed – using Eq. (4) – for an example case. Table 4 shows the VOT values for the six air itinerary share models presented in this paper. For each of the entities an example case has been chosen and the corresponding VOT value has been computed. Considering that VOT values in the literature are typically under $100/hour (Hsiao & Hansen, 2011; Atasoy & Bierlaire, 2012) several observations can be highlighted from the results presented in Table 4. While the estimated VOT for the specified city-pair belonging to the W-M entity is high compared to the literature (i.e., $144.42/hr), the estimated values for the case examples from the other entities are well below $100/hr, and therefore comparable to those found in the literature. This may be because of a lack of differentiation between fare classes, the level of aggregation of the data used or the differences between the entities' estimation datasets.

[Table 4]

6. Model Results Validation

The estimated air itinerary share models are validated using data associated with city-pairs existing in the corresponding entity for the first quarter of 2010. To evaluate the performance of the model, the market share by itinerary predicted by the model is compared to the observed market share obtained directly from the DB1B dataset (BTS-RITA, 2003-2010). Absolute errors are averaged across itineraries, shown in Table 5. Validation results obtained show an average mean absolute error, expressed in terms of percentage deviation, of 14.2%, ranging from 7.5% for the W-E entity to 27.2% for the

M-M entity. Most of the percentage errors in itinerary share are lower than those in the literature (e.g., the model developed by Coldren et al. (2003) for 2010 passenger itinerary shares has a mean absolute error of 16.6%). Only the percentage errors accoriated with the M-M entity, the Hawaii & Alaska-US Continental entity and US Continental-Hawaii & Alaska entity are larger. The model specifications and data aggregation, however, differ markedly, so such a direct comparison of model performance is difficult.

It is believed that the primary differences lie in the fact of the estimation dataset used to estimate the M-M air itinerary share model has the smallest number of observations compared to the other entities, as shown in Table 2. The high mean absolute error values obtained for the Hawaii & Alaska-US Continental entity and the US Continental-Hawaii & Alaska entity, may be due to the different assumptions implicit in the datasets. While the rest of the entities contain city-pairs with the same time-zone difference, these two entities contain a variety of time zones, which may affect the estimation results.

[Table 5.]

7. Conclusion and Future Work

In this paper a step is made 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, an aggregate air itinerary share model is presented that only uses aggregate data, without further insight into service preferences, in contrast to other models in the literature. Given this aggragate input data, the developed model attempts to model demand effects and passenger travel decision more accurately than is possible using other methods. Ultimately, when integrated into a 3-stage model for air traffic forecasting, better predictions of airport-pair traffic flows are expected.

An aggregate multinomial logit model is estimated to predict how market demand is distributed across available itineraries. In an attempt to capture similarities between city-pairs, eighteen models are developed, each modelling traffic-flow between two major regions of the US ATS. In this paper, results for six entities are presented (C-M, M-C, C-W, W-C, M-W and W-M entities). Due to computational limitations some of the models are estimated using a reduced dataset containing information about 100 city-pairs in each of 10 runs. Results obtained from the estimated model show high goodness of fit. All estimated coefficients are significant at the 95th percentile confidence level and are generally of the expected sign.

The estimated models are validated by computing the mean absolute error between the predicted market share and the observed market share. Data for city-pairs from the 1st quarter of 2010 is used for validation. Validation results show an average mean absolute error of 14.2%, ranging from 7.5% for the W-E entity to 27.2% for the M-M entity. In general, the validation results obtained are slightly better than comparable numbers in the literature (Coldren et al., 2003). However, because of differences in model specifications and data aggregation, a direct comparison is difficult. Model evaluation parameters including likelihood ratio test and Rho-squared show reasonable values, with the likelihood ratio test values large enough to reject the null hypothesis and the Rho-squared values showing a reasonable goodness of fit. Estimated VOTs are found to be in line with those in the literature for all the entities, -i.e. under $100/hr -, with the exception of VOT for the W-M entity. This may be because of a lack of differentiation between fare classes, the level of aggregation of the data used or the differences between the entities' estimation datasets.

Model estimation results obtained to date look promising, showing that the application of multinomial logit modelling for air itinerary share estimation at the aggregate level is possible. However, computational intensity is a significant problem, requiring the approach to be adjusted to estimate the model with reduced datasets of 100 city-pairs in each of 10 runs. This leads to some issues with the estimated coefficients, and may reduce model performance. Hence, further work will focus on improving model estimation results through the use of alternative techniques. Those under consideration include neural networks using various learning algorithms such as backpropagation and Levenberg-Marquardt.

In future work the best performing model will be used to estimate the air itinerary shares between city-pairs, so that passenger demand by airport-pair can be predicted and ultimately used as one of the input variables for the final stage of the 3-stage model. Additionally, by providing more accurate itinerary shares, this model could be used to aid the decision making process across multiple stakeholders (e.g. airlines, airport providers, government' agencies, etc.). Route network expansion, equipment purchase or airport expansion are some examples in which its application could be beneficial. Moreover, subject to adequate model refinement, there is the potential of a broader model application to include other transport modes as one of the choice criteria. This would allow for the analysis of, e.g., competition between air and ground transport over short distances.

Acknowledgements

The authors would like to gratefully acknowledge Dr. Lynnette M. Dray from University College London, Dr. Bilge Atasoy from Massachusetts Institute of Technology and Dr. Gregory Coldren from Coldren Choice Consulting Ltd. for their advice on data sources and approach.

Tables

Table 1. Input variables considered to influence air itinerary market share.

Variable	Name	Description
Level of service	LoS	Dummy variable indicating the level of service of the itinerary i (non-stop or one-stop) with respect the best level of service within its market (either non-stop or one-stop with the best connection).
Travel Time Ratio	TT_i^{Ratio}	Ratio between travel time of itinerary *i* and travel time of shortest itinerary in the market *sh.*
Travel Fare Ratio	TF_i^{Ratio}	Average fare paid on a specific itinerary i divided by the market average fare.
Multi-airport system (MAS) Origin	$masORIG_i$	Dummy variable indicating whether the Origin airport is within a multi-airport system or not.
Multi-airport system (MAS) Destination	$masDEST_i$	Dummy variable indicating whether the Destination airport is within a multi-airport system or not.
Origin airport average delay	$\overline{\mathrm{Dly}}_{\mathrm{ORIG}}$	Average departure delay of origin airport for the previous year.
Destination airport average delay	$\overline{\mathrm{Dly}}_{\mathrm{DEST}}$	Average arrival delay of destination airport for the previous year.
Origin airport	$Access_{ORIG}$	Average distance between city center and origin

Accessibility		airport.
Destination airport Accessibility	$Access_{DEST}$	Average distance between city center and destination airport.
Origin and destination airports are hubs	$hub2hub_i$	Dummy variable indicating whether itinerary i is between two hub airports.
Either the origin or destination airport is a hub	$1hub_i$	Dummy variable indicating whether itinerary i is from or to a hub airport.
Neither origin nor destination airports are a hub	no_hub_i	Dummy variable indicating whether itinerary i is not from nor to a hub airport.
Airlines Ratio	$AirlinesRatio_i$	Ratio between the number of airlines serving itinerary i and the number of airlines serving the shortest itinerary sh.

Table 2. Summary statistics for all entities.

Origin Region	**Destination Region**	**City-pairs**	**Itineraries**	**N° itineraries per city-pair**	**N° Hubs**
Hawaii & Alaska	US Continental	438	2,063	19	15
US Continental	Hawaii & Alaska	437	2,052	19	15
Central	Central	1,547	6,335	16	11
Central	East	2,562	14,415	27	19

Central	Mountain	462	1,867	17	17
Central	West	724	5,216	38	19
East	Central	2,552	15,150	38	18
East	East	3,520	21,157	27	17
East	Mountain	508	2,895	18	20
East	West	867	9,268	87	24
Mountain	Central	463	1,899	15	18
Mountain	East	527	3,150	24	18
Mountain	Mountain	134	359	5	6
Mountain	West	252	1,230	27	11
West	Central	724	5,222	38	19
West	East	862	9,274	90	24
West	Mountain	265	1,313	29	11
West	West	356	1,941	31	9
Total		17,200	104,806		

Table 3. Estimated coefficients for the air itinerary choice model corresponding to entities C-M, M-C, C-W, W-C, M-W and W-M.

Variable Name	**C – M**	**M – C**	**C – W**	**W – C**	**M – W**	**W - M**
Level of Service (relevant to every entity)	--	--	--	--	--	--
Markets Containing Non-Stop itineraries:						

hub2hub	0.000	0.000	0.000	0.000	0.000	0.000
1hub	-1.590	-2.090	0.013	0.626	-1.090	-0.846
no_hub	-1.410	-2.650	0.095	0.928	-1.970	-1.380
Airlines Ratio	0.012	0.017	-0.550	0.017	0.010	0.023
Travel Fare Ratio (TF^{Ratio})	-3.840	-4.080	-0.789	-1.321	-1.970	-0.754
Travel Time Ratio (TT^{Ratio})	-1.030	-1.020	-0.170	-0.329	-0.844	-1.530
$\overline{Dly}_{ORIG}$	-0.174	3.950	0.086	0.627	2.010	-0.026
$\overline{Dly}_{DEST}$	2.930	-0.092	0.542	0.227	-0.067	1.110
AccessDEmas	-0.919	0.020	0.044	0.015	0.002	0.331
AccessORmas	0.098	0.864	0.098	-0.001	0.749	0.005
LogLikelihood Ratio Test	523,121	435,323	1,030,223	191,252	908,296	906,390
Rho-squared (ρ^2)	0.724	0.691	0.587	0.559	0.715	0.714

**Note: All variables are statistically significant at the 95% confidence level.*

Table 4. Comparison between Value of Time for the C-M, M-C, C-W, W-C, M-W and W-M entities.

Entity	**Origin City**	**Destination City**	$\overline{TF}$ (\$)	TT_{sh} (hr)	VOT $(\$/hr)$
C – M	Chicago	Denver	137.1	2.51	14.66
M – C	Denver	Chicago	136.6	2.24	15.26
C – W	Chicago	Reno	183.5	4.04	9.79
W – C	Reno	Chicago	184.3	3.59	12.78
M – W	Denver	Los Angeles	150.5	2.17	29.76
W – M	Los Angeles	Denver	151.0	2.12	144.42

Table 5. Mean absolute error in itinerary share computed in terms of percentage deviation.

Origin Region	**Destination Region**	**Number of City-pairs**	**Number of Itineraries**	**Mean absolute Error in Itinerary Share (%)**
Hawaii & Alaska	US Continental	422	1,889	22.60
US Continental	Hawaii & Alaska	435	1,963	24.17
Central	Central	1,490	6,088	13.46
Central	East	2,460	13,457	11.35
Central	Mountain	463	1,931	17.94
Central	West	679	4,814	9.03
East	Central	2,461	13,748	11.14
East	East	3,503	19,487	11.07
East	Mountain	523	3,066	14.06
East	West	785	7,622	8.63
Mountain	Central	464	1,895	16.69
Mountain	East	517	3,049	14.53
Mountain	Mountain	121	309	27.21
Mountain	West	250	1,130	13.22
West	Central	683	4,868	9.40
West	East	786	7,577	7.49
West	Mountain	262	1,243	11.42
West	West	343	1,653	11.97

References

Adler, N. (2001). Competition in a deregulated air transportation market. *European Journal of Operational Research, 129*(2), 337-345.

AIM. (2006). *Aviation Integrated Modelling Project*. Retrieved August 15, 2016, from Aviation Integrated Modelling Project: http://www.aimproject.aero/.

Atasoy, B., & Bierlaire, M. (2012). *An itinerary choice model based on a mixed RP/SP dataset.* ENAC, EPFL. Lausanne: Transport and Mobility Laboratory, ENAC.

Bierlaire, M. (2003). Biogeme: A free package for the estimation of discrete choice models. *3rd Swiss Transportation research conference.* Ascona, Switzerland. Retrieved from Biogeme: A free package for the estimation of discrete choice models.

BTS-RITA. (2003-2010). *Bureau of Transportation Statistics - Research and Innovative Technology Administration.* Retrieved September 14, 2014, from Origin and Destination Survey: DB1B Market for 1003 to 2007: http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=247&DB_Short_Name=Origin%20and%20Destination%20Survey

Busquets, J. G., Alonso, E., & Evans, A. D. (2015). Application of Data Mining in Air Traffic Forecasting. *AIAA Aviation Technology, integration and Operations Conference.* Dallas: AIAA.

Busquets, J. G., Alonso, E., & Evans, A. D. (2016). Predicting Aggregate Air Itinerary Shares Using Discrete Choice Modeling. *AIAA, Aviation Technology, Integration and Operations Conference.* Washington D.C.

Carrier, E. (2008). *Modeling the Choice of an Airline Itinerary and Fare Product using booking and Seat Avaiability Data.* Cambridge, Massachusetts: Massachusetts Institute of Technology.

Coldren, G. M., & Koppelman, F. S. (2005). Modeling the competition among air-travel itinerary shares. GEV model development. *Transportation Research Part A: Policy and Practice, 39*(4), 345-365.

Coldren, G. M., Koppelman, F. S., Kasturirangan, K., & Mukherjee, A. (2003). Modeling aggregate air travel itinerary shares: logit model development at a major US airline. *Journal of Air Transport Management, 9*(6), 361-369.

FAA. (2007-2010). *Federal Aviation Administration.* Retrieved September 20, 2015, from FAA Operations & Performance Data: https://aspm.faa.gov

FAA. (2014). *Federal Aviation Administration.* Retrieved October 10, 2015, from Aviation System Performance Metrics (ASPM) Manuals: https://aspm.faa.gov/

Garrow, L. A. (2010). *Discrete Choice Modelling And Air Travel Demand: theory.* Aldershot, United Kingdom: Ahsgate.

Gramming, J., Hujer, R., & Scheidler, M. (2005). Discrete choice modelling in airline network management. *Journal of Applied Economics, 20*, 467-486.

Grosche, T., & Rothlauf, F. (2007). *Air Travel Itinerary Market Share Estimation.* Manheim: University of Manheim.

Hansen, M. (1995). Positive feedback model of multiple-airport system. *ASCE Journal of Transportation Engineering, 121 (6)*, 453-460.

Hsiao, C., & Hansen, M. (2011). A passanger demand model for air transportation in a hub-and-spoke network. *Transportation Research Part E: Logistics and Transportation Review*(47), 1112-1125.

Kotegawa, T. (2012). *Analyzing the evolutionary mechanism of the air transportation system-of-system using network theory and machine learning algorithm.* West Lafayette, Indiana: Faculty of Purdue University.

Louviere, J. J., Meyer, R. J., Bunch, D. S., Carson, R., & Dellaert, B. (1999). Combinig sources of preference data for modeling complex decision provesses. *Marketting Letters 10*, 205-2017.

Proussaloglou, K., & Koppelman, F. S. (1995). Air carrier demand: an analysis of maket share determinants. *Transportation, 22*(4), 371-388.

Warburg, V., Bhat, C., & Adler, T. (2006). Modeling demographic and unobserved heterogeneity in air passengers' sensitivity to service attributes in itinerary choice. *Transportation Research Record: Journal of the Transportation Research Board, 1951*, 7-16.

Windle, R., & Dreesner, M. (1995). Airport Choice in multiple-aiport regions. *ASCE Journal of Transportation Engineering, 121*(4), 332-337.

548