Application of Data Mining to forecast Air Traffic: A 3-Stage Model using Discrete Choice Modeling

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 this study centers on developing an aggregate air itinerary share model estimated at the city-pair level within the US air transportation system. This route demand assignment model is part of a new modeling approach that has as its ultimate output the prediction of detailed traffic information for the US air transportation system. In this approach, city-pair demand generation, route demand assignment and air traffic levels estimations are completed in 3 different stages within a single framework. Aiming to fully develop the overall model, in this paper we focus on estimating the 2nd stage, the air itinerary choice model. In order to achieve this, the first approach taken applies a multinomial logit model and uses a combination of stated preferences (SP) and revealed preferences (RP) data to estimate the model. By using a mixed dataset, we attempt to improve the RP model results, which often perform poorly due to high demand inelasticity. Preliminary results show the potential of this approach, although further analysis is required to understand the results obtained. For the final paper, different approaches and further interactions among the model attributes will be applied to improve the model's performance.

I. Introduction

AIR TRAFFIC FORECASTS are crucial for planning in the aviation industry, allowing trade-offs between benefits – e.g., economic growth – and negative consequences – e.g., the associated environmental impact of aviation–. To enable supply to adapt to growth in demand, good forecasts of future demand for air traffic as well as good forecasts of how airlines are likely to serve this demand are essential. The latter are 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.1), and are characterized by their relative simplicity. For example, the FAA[2] applies a simple growth factor algorithm to allocate traffic across the US ATS. These approaches also often use similar explanatory variables, generally chosen through the judgment of domain experts. More complex approaches from the literature are often not used because of drawbacks such as computational intensity or relatively low accuracy.

This paper is an extension of previous work[3], which introduced a model 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 that work, three innovations were proposed: 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

[*] 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.

between itineraries. As a result, a 3-stage model, still in development at the time, was introduced, as follows: In the first stage, travel demand by city-pair is estimated; in the 2nd stage, the predicted travel demand by city-pair is transformed into travel demand by airport-pair; and finally, traffic levels are estimated in the third stage. In this paper, efforts center on more fully developing the 3-stage model, with a particular focus on stage 2. While a modeling approach for this second step was identified in the previous work, no model was developed. In this paper, we develop and apply a discrete choice model to distribute the city-pair passenger demand across the available itineraries, identified using the classification algorithm described in the previous work.

The approach described in Ref. 3 for distributing city-pair passenger demand across the available itineraries aimed to estimate itinerary shares, defining the available itineraries based on the level of service and the 25 hubs considered in the study. These assumptions limited the number of itinerary options to 26: a non-stop service plus 25 one-stop services to each of the 25 US hub airports. Although the model presented a valid solution for modeling aggregate air-travel itinerary shares, it fails to directly model the distribution of market shares across the available itineraries serving a given OD city-pair. Consequently, in this paper, we attempt to solve the issue by developing an aggregate air passenger itinerary choice model estimated at the city-pair level within the domestic US air transportation network.

The remainder of the paper is structured as follows: a brief review of existing discrete choice modeling applied in air transportation is outlined in Section II. This is followed by the paper's objectives in Section III. The modeling approach is detailed in Section IV. Information regarding the data sources is outlined in Section V. Modeling results are presented in Section VI, followed by a discussion on future work in Section VII.

II. Literature Review

While the majority of existing research focuses on improving air travel demand models, a growing interest in understanding the behavior of air travelers has become evident in aviation. Consequently, in recent years some researchers focus on modeling competition and customer behavior to determine air-travel itinerary shares using discrete choice methodologies – also known as demand assignment models –. In the air transportation context, understanding passengers' choice behavior can become crucial to support airlines in their network planning and scheduling. While most of the discrete choice models applied in urban transportation are built using disaggregate data and include information about the individual making the decision – i.e. the passenger –; in air transportation, data disaggregation as well as data accessibility are limiting factors. Moreover, most of the early studies on demand assignment for air travel focus on studying the distribution of demand across one single dimension. These early models were mostly applied to analyze air travelers' choice within multi-airport cities or regions – i.e. airport choice models[4,5] – or across airlines – airline choice models[6] –. Although the former is the most widely studied topic in discrete choice model within air transportation, and has given a deeper understanding of the relationship between airport's attributes and airport's market share, a more disaggregated assignment of the air travel volumes is needed.

Consequently, air itinerary choice models have become a growing trend in recent years. Several approaches can be identified in the existing itinerary demand allocation literature. Early models used a multinomial logit (MNL) approach[7,8], which are characterized by their simplicity. However, MNL models are based on the assumption that alternatives –i.e. itinerary options – compete equally with each other. Since itineraries sharing similar attributes are expected to experience more competition among themselves than with other itineraries not sharing these attributes' similarities, this consideration does not hold. In order to solve this uniformity assumption, nested logit (NL) models, mixed multinomial logit (MMNL) models and other alternatives approaches have been used. Using NL models, Ref.9, as described in Ref. 3, presents a 3-level weighted nested logit model to predict airline ridership at the itinerary level. This model is applied at an aggregate level and variables included are chosen to capture the inter-itinerary competition dynamic along three dimensions. Similarly, Ref. 10 presents a route demand assignment model, which is included in an air passenger model that also deals with city-pair demand generation within the same framework. For the route demand assignment model, the study introduces different approaches, with the 3-level nested logit model being the one with most promising results. The three dimensions considered within the nesting structure proposed are transport mode, airport choice and route choice. Using a MMNL model, Ref. 11 presents an approach to analyze the itinerary choice behavior of business travelers. The advantage of using this modeling approach is the fact that it allows random taste variations across the individuals making the decision. Finally, more recent studies include alternatives approaches. For example, Ref. 12 applies a multinomial probit (MNP) model to analyze the demand distribution across itineraries to support airlines' management process, while Ref. 13 introduces a latent class model for airline itinerary and fare product choice.

Alternatively, other studies look at different types of data when developing the modeling approaches. Inspiring the approach presented in this paper, Ref. 14 focuses its work on studying the use of a mixed dataset – containing

revealed preferences (RP) and stated preferences (SP) data – in order to develop an itinerary choice model. Ref. 14's approach aims to exploit the variability of SP data for the estimation of the RP model parameters, which often perform poorly due to high demand inelasticity. Even though further analysis on the results obtained need to be generated, results show the potential advantage of using the SP data to model route demand assignment.

Although results from these studies are promising, the approaches used 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. Taking as a reference the work done by Ref. 14, further enhancements on existing route demand assignment models are still possible. Such improvements are the focus of this paper. Hence, efforts center on developing an aggregate air itinerary share model estimated at the city-pair level within the US ATS, to be eventually included within the single modeling framework to produce future air traffic levels forecasts presented in Ref. 3.

III. 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. Eventually, this model will be combined along with models for forecasting air travel demand and air traffic levels, all within the same framework (described in Ref. 3). The ultimate air traffic forecasting model is inspired by previous research[15] that focused on improving the FAA's forecasting methodology and for which further potential improvements have been identified. Consequently, in an attempt to fully develop the 3-stage model presented in Ref. 3, and therefore, centering the efforts in the model's stage 2, the approach described in this paper is expected to:

- Highlight the most important factors underlying the air traveler's choice behavior within the domestic 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 elements beyond that of the existing research:

- Explicitly modeling the distribution of city-pair passenger demand between itineraries within the US ATS, so that when integrating this approach with the air traffic demand model presented in Ref. 3, better predictions of airport-pair flows will be generated.
- The consideration of a combined RP and SP dataset rather than just RP data as typically considered.
- Ultimately, presenting a 3-stage model that deals with city-pair demand generation, route demand assignment and air traffic levels estimations within a single framework. The use of data mining and machine learning techniques to develop this model will allow modeling the US ATS evolution, as described in Ref. 3; as well as producing prediction of air traffic with improved accuracy and precision levels while maintaining the simplicity of existing econometric, gravity and time-series models.

IV. Approach

A. Detailed Forecasting Methodology

Following the work presented in Ref. 3, which introduced a 3-stage model for the purpose of forecasting future air traffic levels, this paper focuses on fully developing its stage 2. The objective of this phase is to transform Origin-Destination (O-D) demand by city-pair into passenger demand by airport-pair by using an air itinerary choice model. The overall approach of the 3-stage model is shown in Figure 1 in Appendix A. Each stage is as follows:

- 1st stage: Air travel demand by city-pair is estimated using population, income, dummy variables indicating attractiveness of the city, availability of other transport modes and generalized cost as input variables, as described in Ref. 3. This follows the approaches used to predict O-D passenger demand described by Ref. 16 and Ref. 17.
- 2nd stage: O-D demand by city-pair extracted from the 1st stage is transformed into passenger demand by airport-pair, which serves as a stronger driver of airport-pair air traffic. This is the main focus of this paper, estimating an itinerary choice model that directly models the distribution of market shares across the available itineraries serving a given OD city-pair.
- 3rd stage: The 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 - i.e. flight frequency from previous year, number of airports serving a city, fuel price and dummy variable indicating whether an airport serving a city is a hub or not –, as described in Ref. 3.

Different approaches are used in each one of the phases of the 3-stage model. Linear regression with logarithmic transformation in both the dependent and independent variables – i.e. log-log model – is used in stage 1 and 3, as described in Ref. 3. Stage 2 consists in 2 steps: identification of available itineraries estimated using logistic regression (described in detail in Ref. 3); followed by the distribution of the O-D demand by city-pair obtained from the 1st stage across the available itineraries using a discrete choice model. The latter phase of this stage 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 Ref. 9: 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 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), etc., 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.

An aggregate itinerary choice model is developed, as done by Ref. 9 and Ref. 14. In contrast to previous work[3], this approach aims to directly model the distribution of passenger demand amongst all available routes serving a given city-pair, instead of only focusing on the level of service. 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 characteristics of the alternatives. Although the aggregate approach presents some disadvantages compared to the disaggregate approach, the detailed data for each individual required for the disaggregate approach is not available. In this extended abstract of the study, a simplified approach using a mixed dataset formed of RP and SP data to model air itinerary choice is presented. The RP data is booking data from airlines operating within the US domestic market and the SP data is based on an internet survey in US. City-pairs, *M*, considered are those within the domestic US ATS and are defined by origin and a destination. The universal choice set, *C*, is form for all possible itineraries within the entire ATS. The choice problem is defined for each city-pair, $m \in M$, with choice set of all the possible itineraries in that given city-pair represented by I_m. Each itinerary *i* is characterized by a set of attributes such as level of service, price and time. As a simplification, only two possible level of service are considered, non-stop and one-stop flights. For the one-stop flights, the connections available are through one of the 25 US hubs considered in this study[§].

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. (1) 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-in-parameters 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 individual will choose the alternative with the highest value, maximizing their utility. Equation (2) 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*[14].

$$S_i = \frac{Exp(V_i)}{\sum_i exp(V_i)} \tag{1}$$

$$V_i = \beta' \cdot X_i = \beta_1 \cdot X_{i1} + \beta_2 \cdot X_{i2} + \ldots + \beta_k \cdot X_{ik} \tag{2}$$

Attributes included in vector X_i are as follows:

- Travel time, TT_i, is the travel time of itinerary *i* in minutes.
- Travel cost, TC_i, is the travel time of itinerary *i* in $, which is normalized by 100 for scaling purposes.
- *non-stop*$_i$, is a dummy variable which is 1 if itinerary *i* is composed of 1 leg, 0 otherwise.
- *one-stop*$_i$, is a dummy variable which is 1 if itinerary *i* is composed of 2 leg, 0 otherwise.

[§] IATA codes for the 25 hubs considered in this study are: ORD, ATL, DFW, LAX, IAH, DEN, DTW, PHL, CVG, MSP, PHX, EWR, CLT IAD, JFK, LAS, MIA, SFO, SLC, SEA, BWI, STL, CLE, MEM, PIT.

Interactions among the attributes are accounted for by the model. Table 1 shows the specifications of the utilities. It can be seen how *TT* and *TC* interact with dummy variables *non-stop* and *one-stop* – i.e. level of service – representing the strong correlation between the number of stops and the travel cost of the itinerary. These interactions are an attempt to capture the correlation between the specific hub and both variables, price and time. Moreover, a logarithmic transformation has been considered for the variable price, assuming that the effect of increasing the price is not linear across the several price levels. For further improvements, other interactions are to be considered for the final paper. For example, the importance of an itinerary origin and/or destination airport can be correlated with time and price for non-stop itineraries. Similarly, for one-stop itineraries, the importance of hub location on travel time and cost can also be accounted.

To better understand the results obtained from the air itinerary choice model, indicators such as willingness to pay can be computed. Value of time (VOT) is the willingness of passengers to pay for one hour of travel. VOT is given by Eq. (3), which is computed for each given alternative *i*. Note that due to including the travel price in logarithmic transformation when computing the utility, travel cost is also included in the formulation of VOT.

Table 1. Specifications of the utilities for the itinerary choice model.

Constant	ASC_i	$1 \text{ x alternative}_i$
Travel Cost	β_{TC}^{NS}	$ln(TC_i/100) \text{ x non-stop}_i$
	β_{TC}^{OS}	$ln(TC_i/100) \text{ x one-stop}_i$
Travel Time	β_{TT}^{NS}	$TT_i \text{ x non-stop}_i$
	β_{TT}^{OS}	$TT_i \text{ x one-stop}_i$

$$VOT_i = \frac{\partial V_i/\partial time_i}{\partial V_i/\partial price_i} = \frac{\beta_{TT} \cdot TC}{\beta_{TC}} \tag{3}$$

During the estimation of the model, for each city-pair considered, the utility and likelihood function are computed, with the latter being combined in order to give the estimated log likelihood. Once the itinerary choice model is estimated using the MNL algorithm, Eq. (1) is applied to compute the market share of passengers for each itinerary. The estimated passenger demand per itinerary is then used to compute segment demand, i.e. passenger demand per airport-pair, which will eventually be used as input for the 3rd stage within the 3-stage model as described in detail in Ref. 3.

V. Application

The models described above are applied to a network of 337 airports within the US ATS, as in Ref.18. 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. 18.

Historical flight frequency data and airlines schedule are extracted from US Department of Transport T-100 data[19], 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[20]. Travel times and costs are also extracted from Ref. 20 while the SP data is obtained from an Internet survey conducted by the Boeing Company in the fall of 2004[21]. The air itinerary choice model is estimated using *Biogeme*[22]. Flight delay information is obtained from the FAA Aviation System Performance Metrics (ASPM) database[23].

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 in Ref. 3.

Once the model is estimated, it will be applied 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 TAF in future work.

VI. Model Estimation Results

From the 3-stage model, results obtained for stage 1 and 3, as well as results for the identification of available itineraries between city-pairs, are presented and described in Ref. 3.

The air itinerary choice model introduced in this paper is under development and the model estimation results will be included in the final paper. However, results have been generated on a reduced dataset for the entity Continental US to Alaska and Hawaii. Two different datasets have been used to generate these results, differing by the number of city-pairs included from RP data, with one considering 7 OD pairs and the other one 15 OD pairs. Both datasets are then combined with SP data. There are a maximum of 41 alternatives available for each OD pair. The estimation results model when considering both datasets are presented in Table 2. For the first case, when only 7 OD pairs from RP data are included, all estimated coefficients are statistically significant with a 95% confidence level. All travel time and cost coefficients have negative signs, as one would expect since the increase on time and/or on price of a given itinerary will decrease its attractiveness, and therefore, its utility. Finally, the magnitude of the travel cost parameter for the non-stop itineraries is larger than the cost parameter for one-stop itineraries. However, for the travel time parameter the opposite effect is true, although the difference in this case is quite small. This indicates that passengers on connecting itineraries are less affected by an increase in price than on non-stop itineraries. Contrariwise, passengers on non-stop itineraries are less affected by an increase on travel time compared to one-stop itineraries. Both these results are unexpected, as non-stop passengers would typically be expected to be less price sensitive and more travel time sensitive. Regarding the price, the fact that in some observations within the RP data one-stop itineraries are more expensive might be the reason why the results suggest that passengers are less affected by an increase in price on one-stop itineraries compared to non-stop itineraries. In the case of travel time, the limited number of available one-stop itineraries, resulting from the consideration of only 25 connecting hubs when creating the dataset, can be the cause of lack on variability for the travel time variable, leading to the results described. More analysis is needed to better understand these results.

For the case when 15 OD pairs are considered, estimated coefficients for the travel cost have negative signs as expected. However, the estimated travel time coefficients have positive signs, which is opposite to what one would expect since an increase in travel time is expected to decrease the attractiveness of the itinerary. When considering the magnitude of the estimated parameters, price and time for non-stop itineraries are larger compared to those for one-stop itineraries. Similarly to the earlier case, most of the one-stop itineraries in the RP data have higher prices and as expected travel times are longer. All estimated coefficients are statistically significant with a 95% confidence level.

Table 2. Estimated coefficients and corresponding t-statistics for the air itinerary choice model corresponding to entity Continental US to Alaska and Hawaii.

		7 ODs		*15 ODs*	
	Parameters	*Coefficient*	*t-test*	*Coefficient*	*t-test*
Travel Cost	β_{TC}^{NS}	-9.84	*-20.36*	-7.79	*-35.92*
	β_{TC}^{OS}	-3.22	*-11.54*	-3.2	*-22.13*
Travel Time	β_{TT}^{NS}	-2.13	*-17.52*	1.07	*37.66*
	β_{TT}^{OS}	-2.81	*-20.56*	0.174	*6.34*
Likelihood ratio test		118966		214255	
ρ^2		0.257		0.304	

Comparing results from the two cases, when 7 and 15 OD pairs are considered respectively, it is important to note the type of data used. The lack of variability in the RP data may be the cause for poorer results, such as the unexpected positive sign for the travel time in the 2nd case. Despite this, the latter model shows a slight improvement on the goodness of fit with a ρ^2 of 0.304, which compared to the 1st case (0.257) corresponds to an increase of 19%. 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. (3) - for two itineraries differing by the number of stops, but with

the same price. For example, taking a non-stop and a one-stop itinerary both costing $400, VOT values computed for both datasets are shown in Table 3. It can be seen how in the first case, VOT values obtained are opposite to what one would expect since it shows that passengers are willing to pay a much higher price for one hour reduction on their one-stop itineraries than on their non-stop itineraries. Looking at the second case, when 15 OD pairs from RP data are included, VOT values in absolute terms make more sense since a higher price is expected to be paid for an hour reduction of travel time of non-stop itineraries compared to one-stop itineraries. However, in this case, the signs obtained are opposite to what one would expect.

Table 3. Value of Time for the case when itineraries with different levels of service cost the same price of $400.

	7 ODs	*15 ODs*
VOT_{NS}	87	-55
VOT_{OS}	349	-22

VII. Conclusions and Future Work

Research described in this paper provides an 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, an aggregate air itinerary choice model, which is part of a 3-stage model for air traffic forecasting, is presented. The model introduced aims to directly estimate the distribution of the O-D demand across the available itineraries serving a given city-pair using the domestic US ATS as example application. The modeling approach explores the use of a mixed SP and RP dataset, which is expected to improve results compared to existing researches only using RP data.

Initial development of the route demand assignment model is still in progress. However, some preliminary results have been produced for one of the entities – Continental US to Alaska and Hawaii –. From these results some aspects can be highlighted. When using a reduced dataset with a limited RP data, results are quite promising, showing estimated coefficients for travel cost and time with negative signs and all being statistically significant. When increasing the number of OD pairs included from the RP data, results change slightly. While some aspects worsen compared to previous estimated results - e.g. sign for travel time estimated coefficients become positive -; others improve, such as the goodness of fit for the model that goes from 0.257 to 0.304. These discrepancies are believed to be caused by lack of variability within RP data as well as the nature of the data itself, having in some cases higher prices for one-stop itineraries than for non-stop itineraries.

Model estimation results obtained to date look promising. However, there is room for improvement and further work is planned to be included in the final paper. New attributes and new correlations will be considered, as well as better understanding the data used. Alternative modeling approaches, such as NL or MMNL models, will also be investigated.

Ultimately, the estimated air itinerary choice model will be included within the 3-stage model, which aims to model air travel demand, route demand assignment and air traffic demand within a single framework. The proposed modeling framework provides with an effort to improve on existing air traffic forecasting methodologies by using an innovative approach. After the full model is developed, it 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.

Acknowledgments

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

[1]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].

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

[3]Busquets, J. G., Alonso, E., and Evans, A. D, "Application of Data Mining in Air Traffic Forecasting," *AIAA Aviation Technology, Integration and Operations Conference*, Dallas 2015.

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

[5]Windle, R., Dreesner, M., "Airport choice in multiple-airport regions," *ASCE Journal of Transportation Engineering,* 121 (4), 1995, pp. 332-337.

[6]Proussaloglou, K, Koppelman, F., "Air carrier demand: an analysis of market share determinants," *Transportation,* vol. 22. 4, 1995, pp. 371-388.

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

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

[9]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.

[10]Hsiao, C., Hanse, M., "A passenger demand model for air transportation in a hub-and-spoke network," *Transportation Research Part E: Logistics and Transportation Review* 47.6, 2011, pp. 1112-1125.

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

[12]Gramming, J., Hujer, R., Scheidler, M., "Discrete choice modelling in airline network management." *Journal of Applied Economics* 20, 2005, pp. 467-486.

[13]Carrier, E., "Modeling the Choice of an Airline Itinerary and Fare Product using Booking and Seat Availability Data." PhD dissertation, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, 2008.

[14]Atasoy, B., Bierlaire, M., "An itinerary choice model based on a mixed RP/SP dataset" *Technical report TRANSP-OR 120426.* Transport and Mobility Laboratory, ENAC, EPFL.

[15]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.

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

[17]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.

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

[19]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].

[20]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].

[21]Boeing, "Mathematical Modeling of Behaviour course", *Transport and Mobility Laboratory,* Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland (2015).

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

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

Appendix A

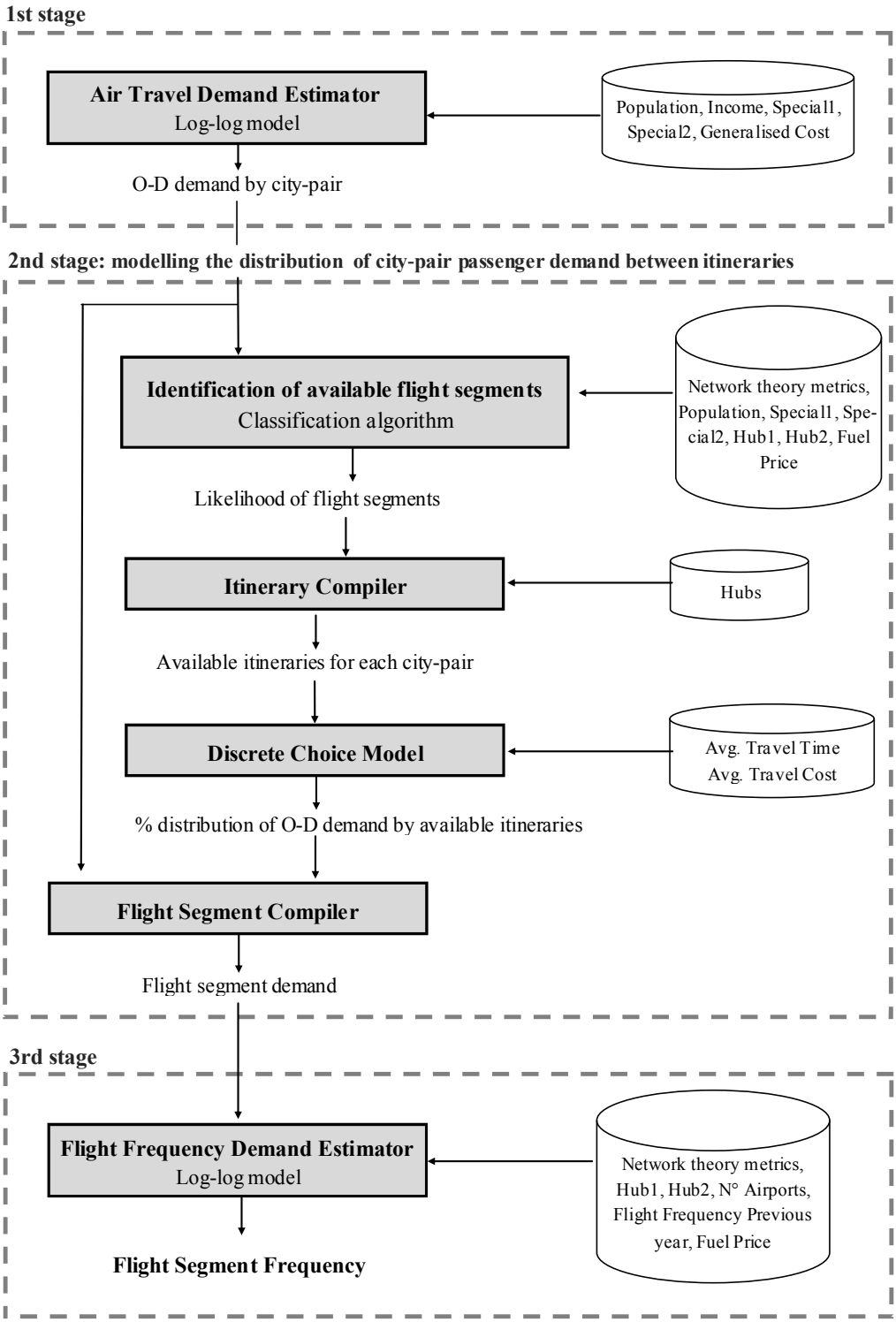

Figure 1. Flowchart of the 3-stage model.