Hierarchical Decision-Modeling Framework for Air Transportation System

Conference Paper · June 2016

DOI: 10.2514/6.2016-3154

Towards A Hierarchical Decision-Centric Modeling Framework for Air Transportation Systems

Kushal Moolchandani*, Zhenghui Sha†, Apoorv Maheshwari‡, Joseph Thekinen§, Navindran Davendralingam¶, Jitesh H. Panchal‖, Daniel A. DeLaurentis**
Purdue University, West Lafayette, IN 47907

This work is part of ongoing research towards the development of a multi-level framework that is envisioned to provide the ability to holistically study the future evolution of the ATS as a result of decisions made by relevant stakeholders. In prior work, the authors have developed a discrete choice model to mimic the airlines route-planning decisions. In this paper, we present a passengers choice model, which forms the post-airline decision phase in the resulting bi-level decision-making framework. The passengers in our model make decisions on the type of itineraries being offers by the airlines, including both the choice of number of layovers to make and where to make those layovers. Specifically, we show separate multiple linear regression models for passengers choices on three different types of itineraries non-stop, one-stop, and two-or-more-stops. We present results of our model using data from the Bureau of Transportation Statistics.

I. Introduction

THE air transportation system (ATS) is a complex, evolving system with many stakeholders, all of who act in accordance with their respective objectives, and thus make the analysis of such systems challenging. Sgouridis et al. [1] explain that the "global air transportation systems are governed by feedbacks and time dependencies, stakeholder interactions and decision processes, and non-linearity that make it hard for simple extrapolation models to capture and test future dynamics of the system." DeLaurentis [2] argues it is necessary to "understand air transportation as a system of systems." Similarly, Donohue [3] describes the ATS as a complex adaptive system which evolves due to a combination of technology, regulatory, and economic activities. Given this interdependent nature, it is useful for the stakeholders such as the Federal Aviation Administration (FAA), NASA, and many others, to predict the behaviors of other stakeholders in the system while making their own decisions.

With this motivation, we propose a multi-level multi-stakeholder decision framework, which is notionally shown in Fig. 1, including the three levels of passengers, airlines, and the FAA. As seen in this figure, passengers' decisions on their choice of itinerary will influence the demand distribution in the network, while the airlines' decisions on route selection would affect the available options to the passengers. Choices of both these entities would finally affect the network topology as shown by arrows pointing to the middle layer in the figure. As shown, we will make this framework modular, thus providing an ability to increase or decrease its fidelity depending on the requirements and available computational resources. With such a multi-level framework we can realistically assess how the entities at the higher levels in the hierarchy can account for the decisions of those at the lower levels in their own model, and how, in turn, their policies will affect those at the lower level.

We have previously developed a model that approximates airlines' decisions on route selection [4], i.e., the decisions on whether or not to add a route between a given city-pair, and whether or not to delete an existing route between a city-pair. This model uses publicly available data, from the Bureau of Transportation Statistics (BTS), on the airlines' past route selection decisions, market demand (i.e., the demand between origin-destination city pairs), direct operating cost (DOC), distance between airports, and the hub / non-hub nature of terminal airports as input, and employs a discrete choice model based approach to predict which routes would be operated on by the airlines. Note that, in this model, the airlines decide on whether or not to add a *direct* route between a given city-pair. The BTS T-100 data tables [5] are the source of input data.

*Graduate Research Assistant, School of Aeronautics and Astronautics, AIAA Member
†Post-doc Research Associate, School of Mechanical Engineering
‡Graduate Research Assistant, School of Aeronautics and Astronautics
§Graduate Research Assistant, School of Mechanical Engineering
¶Research Scientist, School of Aeronautics and Astronautics, AIAA Member
‖Associate Professor, School of Mechanical Engineering
**Professor, School of Aeronautics and Astronautics, AIAA Associate Fellow

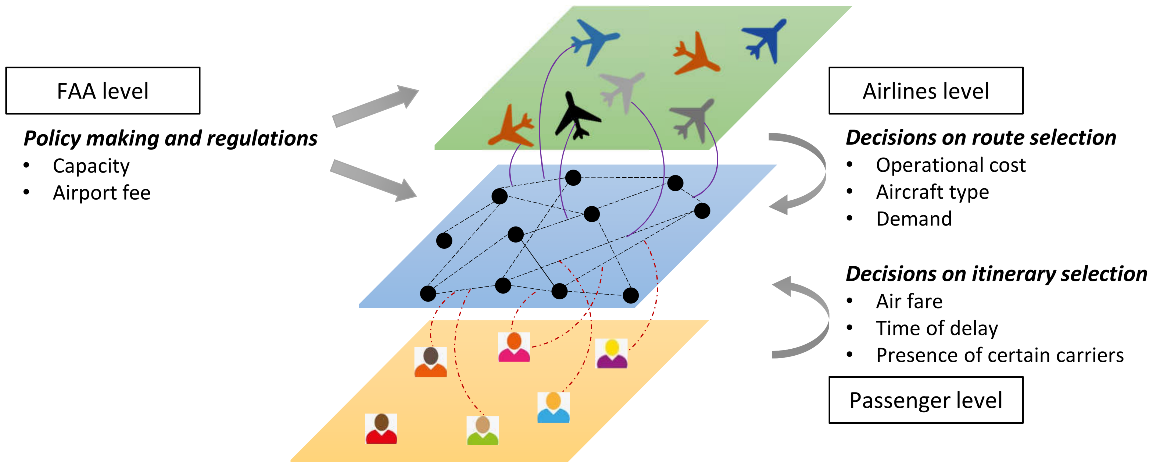


Figure 1. Multilevel decision-making in air transportation system

While useful, this kind of decision modeling does not take into account that the airlines may choose to route passengers through routes that already exist rather than add one on every high demand market. This is particularly true for network carriers that make use of the hub-and-spoke network structure in their operations. Work in this paper is based on the premise that an understanding of passenger behavior regarding itinerary selection given the alternatives would be useful for airlines to study the effects of their network structure on their profitability. By modeling passengers' preferences, the airlines can seek to offer those routes that are most desirable to the passengers while also economically favorable to the airlines themselves. Further, they can observe passengers' response to different values of the modeled explanatory variables such as fare.

The *research goal* is therefore to establish a multilevel decision-centric modeling framework for air transportation network. Specifically, in the current work, we describe the level of passengers' decision making within the framework; the 'FAA level', as well as integration of all these levels, would form part of our future efforts in the development of a multilevel decision-centric framework. The output from our model would be the distribution of passenger demand in the network based on passengers' preferences.

Once we have a multilevel framework for modeling decisions in the ATS, it can be used within a simulation environment for testing hypotheses and for forecasting network evolution in response to the stakeholders' decision-making. For example, we could test the hypothesis that if the passengers prefer flying direct routes to those with stops, then the airlines would be compelled to offer more non-stop flights which would lead to a higher density network. While we do not conduct any simulations, our models can be used within tools that do so. Furthermore, a multilevel framework facilitates the integration of theories from different domains for understanding the dynamics in complex systems, here the ATS, in a systems science perspective [6].

II. Technical Approach

A. Hierarchical Decision-Modeling

First, we describe the bi-level framework comprised of two models – the airlines' decision model on route selection and the demand forecasting model on itinerary selection, as shown in Fig. 2. The former model is the one we have developed previously [4], and the development of the latter is the contribution of this paper. The development of demand forecasting model is inspired by the work done by Bhadra and Hogan [7], wherein the authors used multinomial logit models to predict the probability of a specific itinerary type for a given route (e.g., non-stop or one-stop), thereby forecasting the distribution of itinerary types. Our process takes a similar approach, as shown in Fig. 2, by taking the data on market demand in our network from BTS DB1B database [8] and splitting it into three different itinerary types, viz. non-stop, one-stop, and two-or-more-stop. Note that we label all itineraries with two or more stops as two-stop. With the observed data, we build our models for demand forecasting which predict the number of passengers on each itinerary type. However, the major difference between our work and Bhadra and Hogans work is that our work focuses

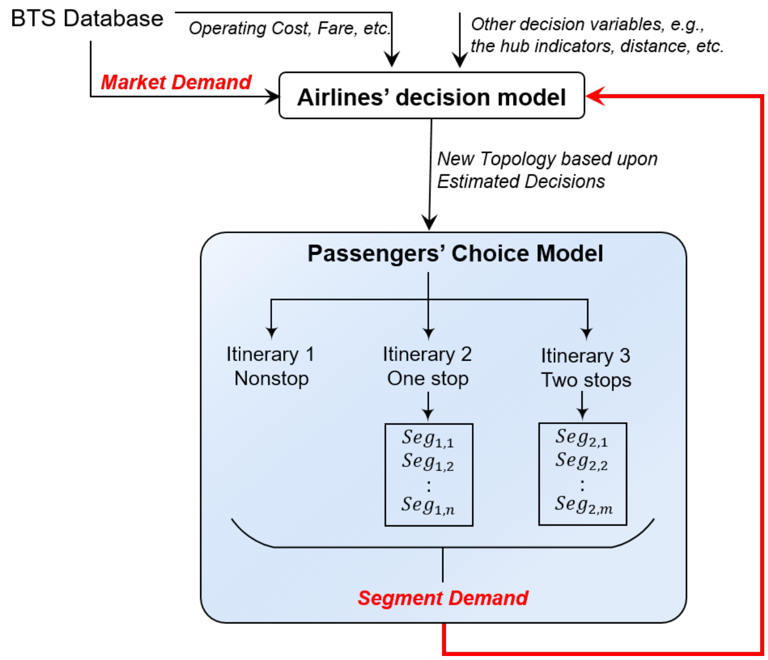


Figure 2. Flowchart of the hierarchical decision-modeling framework

on the forecast of demand, whereas their work provides us the insights of identifying appropriate explanatory variables for model development, as discussed in what follows.

The detailed procedure of such demand forecasting models is illustrated in Fig. 3. The objective is to build a regression model in which the dependent variables are the fraction of passengers on different type of itineraries, and the explanatory variables (independent variables) include market demand, geographic distance, flight fare, and carrier indicator which reflects whether a route is operated by a low-cost carries or a network carrier. The observation data of these explanatory variables can be obtained from the DB1B market and coupon datasets. To obtain the observation data of the dependent variables, we follow three steps: a) for each route in the BTS dataset, there is a unique route ID which can be used as an identifier to extract all the coupon information associated with that route; b) From the information on coupons, we can obtain the itinerary information as well as the number of coupons and number of passengers on each type of itinerary; c) finally, the fraction of passengers (demand ratio) is calculated for each of the three types of itinerary for a specific route. The demand forecasting model is then developed by regression between dependent variables and the explanatory variables; in this study, we use the multiple linear regression procedure.

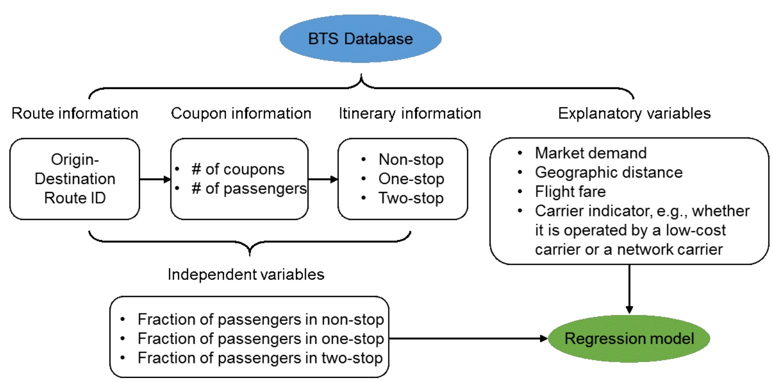


Figure 3. Detail of the demand forecasting framework

To get the distribution of demand in the network, we need information on the actual segments on which the passengers travel. Thus our next step is to model the passengers' decisions regarding the segments that would constitute their trip, as shown in Fig. 2. Therefore, the main objective of developing the demand models is not only for forecasting the demand on each itinerary, but also for estimating the demand at each segment. This is because the segment demand is critical to the following airlines route selection model. While doing the segment demand forecast, an implicit assumption in the above process of itinerary selection is that the passengers decide between the non-stop, one-stop, and two-stop options, without considering which segment is involved.

For any given city pair, there could be multiple ways in which a one-stop or a two-stop trip can be completed, which in Fig. 2, are shown as the division of one-stop and two-stop itineraries into multiple segments. We refer to a combination of segments required to complete a trip as a *ticket*, and preset the set of tickets for each city pair in our network. Also, unlike the passengers' choice on number of stops to make on a trip, we divide the passengers within one itinerary type into different tickets according to a probability distribution based on existing historical demand data. Finally, this output gives the segment demand on all the routes in the network, which is then fed to the airlines' decision model that we developed previously along with the cost data and other decision variables from the BTS Database. In this following subsection, we introduce the theoretical foundations for developing the demand forecasting models and the airlines route selection models.

B. Multiple linear regression

In reference [9], the authors have used logistic regression to model passengers' choices. In contrast, we adopt the multiple linear regression model, as shown in a standard formulation Eq. 1. We found that our linear regression model provided with better estimates of passengers' choices, and this led to opt for this model. A possible reason for the poor performance of our logistic regression model is that we do not have data on all alternatives available to the passengers. For example, a given passenger may have taken a one-stop itinerary between a particular origin-destination airports pair. In this case, we have data on the actual fare paid by the passenger for this one-stop itinerary, but have to estimate the values that would have been paid in cases of non-stop and two-stop itineraries. Further exploration of a logistic regression model is part of our planned future efforts.

$$Y_i = \beta_0 + \beta_1 X_{i,1} + ... + \beta_{p-1} X_{i,p-1} + \epsilon_i \tag{1}$$

In the above equation, X are the explanatory variables, also referred to as the predictor variables, and β are the coefficients corresponding to each one of those predictor variables. ϵ are the error terms, and these are assumed to be normally distributed. The predictor variables can be either continuous or discrete, and in the general case, can be functions of independent predictor variables.

A regression is classified linear when it is linear in its coefficients. The selection of the appropriate predictor variables is the task of model selection procedure. In our modeling, we used the same predictor variables as in [9] to include in our model. While other variables might be useful in modeling a linear function such as the one we desire, our current choice seemed to give satisfactory results and hence was adopted for use. As explained in a later section, part of future work would involve exploring other predictor variables as well as forms of linear regression models. Finally, for the purpose of estimation we made use of the 'lm' function in R. The outputs and the selected linear models are presented in Section IV.

C. The Passengers' Choice Model

Our multiple linear regression model of the passenger choice estimation made use of the following set of predictor variables in its final form:

- $average\_market\_fare$: This is the average fare paid by all passengers on a given route, regardless of class of travel.
- $average\_market\_miles$: This value equals the average number of miles flown by the passengers on a given route. In case of non-stop passengers, this value equals the track miles, while the passengers on one- and two-stop itineraries travel greater distances.
- leg_car: This is an indicator variable that identifies whether a network (legacy) carrier operates on the given route (value = 1) or not (value = 0).
- low_car: Similar to above, this indicator variable identifies the operation of low-cost carriers on the given route.
- $total\_market\_pax$: This variable reflects the demand on the given city-pair market. We expect that the markets with higher demand would correspond to those with higher number of flights, and possibly more itinerary options for passengers to choose from.

Using the 'lm' function in R, we obtained coefficients for the linear model, as shown in Table 1.

III. Data Collection and Analysis for Decision Variables

The first step in our process was that of getting data from all three tables in the "Airline Origin and Destination Survey (DB1B)" from BTS [8], including the coupon, market, and ticket tables. Together these three tables for the fourth quarter of 2014, the year of our analysis presented in this paper, amounted to 2.12 GB in size. Considering this size, we made use of SQL server to save and manage these tables, and interfaced it with R for the purpose of processing. Using the "RMySql" package we queried the raw tables to read in only a subset of columns that are relevant for our analysis as well as apply the following filters:

1. Read only those rows which have positive values of passengers

2. Remove rows which correspond to values of 'Bulk fare' of 1

The second filter of bulk fare is the one suggested by BTS since such fares do not reflect the true amount paid by passengers for their journey.

Finally, we considered only the markets between 132 major airports which constitute primary commercial operations in the US ATS [15]. These airports are regarded as either large, medium, or small hubs based on the number of enplanements; in 2013, for example, these airports handled over 96% of all enplanements in the US. With the 132 selected airports as nodes, an edge is added into the network if it passes four filters:

1) It has at least 8 scheduled departures over a period of any two consecutive months in a given year.

2) It has non-zero passenger demand.

3) It is classified as 'domestic' category in the BTS table.

4) It is not exclusively served by aircraft with freight configuration as reported in the BTS.

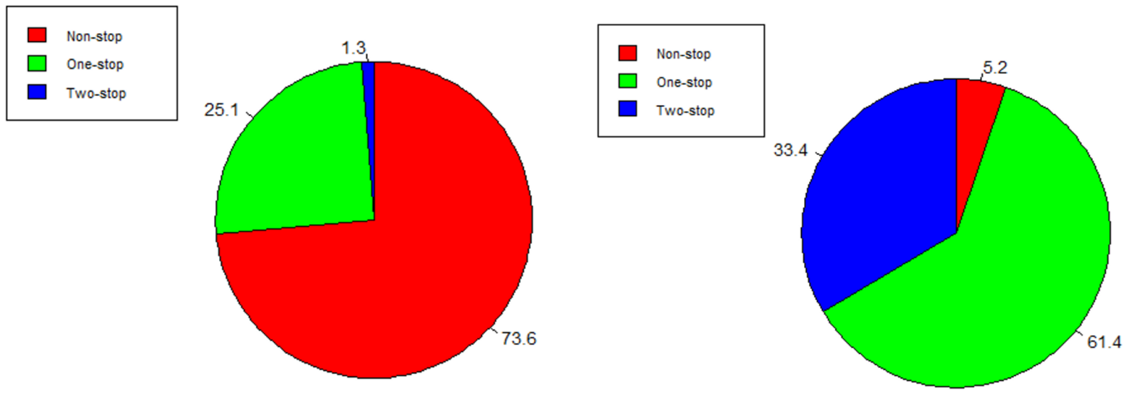
These filters are described in detail in Ref. [4]. Finally, we filtered these tables further to retain data on only those carriers and itineraries that were common among all three tables. On application of all of these filters, we were left with 764 MB of data in all three tables combined.

Following this, our next step was to aggregate data on every route in our network. Using the coupon table, we identified the segment demand on each route, while from the market table, we identified the market demand. Together, these two tables also provided us with the data on all our explanatory variables, viz. geographic distance, average fare, and the presence of network and low-cost carriers on every route. To limit the number of choices available to the passengers, we decided that itineraries could only be one of the following three choices – non-stop, one-stop, and two or more stops which we label as two-stop. The number of itineraries with greater than two stops forms a small fraction of the total, and hence this simplification does not lead to much loss of data. Figure 4 shows a snapshot of data used in multiple linear regression model for non-stop routes; similar data sets were used for regression on one- and two-stop routes.

ALL_POSSIBLE_ROUTES	TOT_SEG_PAX	TOT_MKT_PAX	AVG_MARKET_MILES_0	MKT_PAX_0	AVG_MKT_FARE_0	MKT_COUPONS_0	LEG_CAR	LOW_CAR	FRAC_COUPONS
1014010397	2194	736	1269	391	304.7625	240	1	1	0.433212996
1014010529	4	197	1886	4	167	4	1	1	0.021390374
1014010599	1	90	1138	1	350	1	1	1	0.012195122
1014010693	3	258	1123	2	305.5	2	1	1	0.008333333
1014010713	3	164	781	2	239.5	2	1	1	0.012578616
1014010721	62	642	1974	62	225.6666667	57	1	1	0.094214876

Figure 4. Example of data used for multiple linear regression on non-stop routes

Figure 5a shows a pie chart of distribution of demand by itinerary type. Clearly, the number of passengers who travel on non-stop itineraries are the most dominant. They are followed by those on the one-stop itinerary, and finally the two-or-more stop itineraries. Likewise, Fig. 5b show the number of unique itineraries of each type obtained from DB1B data for quarter 4 of 2014. Fig. 6 shows the histograms of market fares that we used in our modeling. All of these histograms follow normal distribution.



a) Distribution of passenger demand by itinerary types **b) Distribution of unique itineraries by itinerary types**

Figure 5. Distribution of filtered data for fourth quarter of 2014 where total number of passengers is 4,352,707 and total number of unique itineraries is 66,803. (Note: "Two-stop" indicates all itineraries that are two-or-more stops long.)

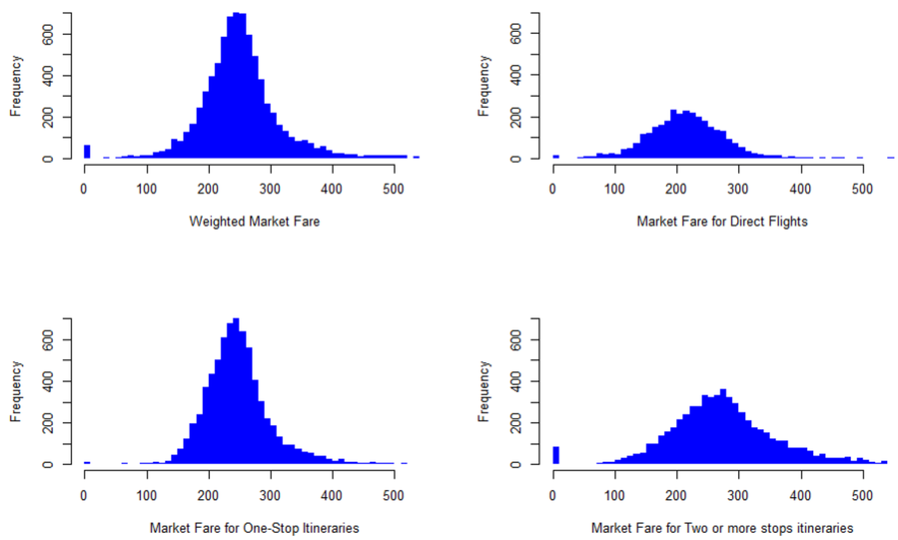


Figure 6. Distribution of market fares by itinerary types

IV. Modeling results

In this section, we first present our models of passenger itinerary selection. Following this, we use these models to predict the passenger demand over all of the segments in our data set. Finally, we compare our predicted segment demand with the actual observed demand, and this serves as a form of validation of our models.

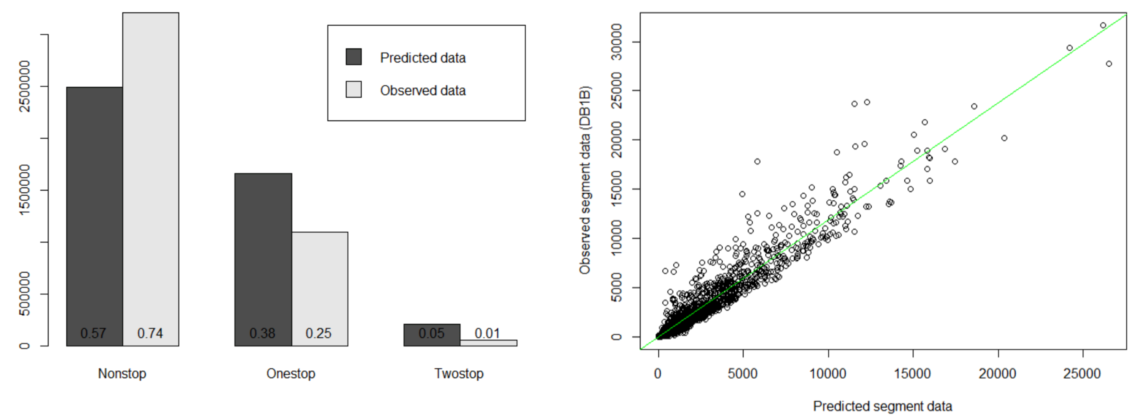
The multiple linear models for each of the three itinerary types is given in Table 1. All of the three models are statistically significant, though there are some differences with regards to the predictor variables. For example, we see that the model for non-stop itinerary choices has an R^2 of 0.38, and all predictor variables excluding average market fare are significant. For this model, the variable $average_{m}arket_{f}are$ has a p-value of 0.107. This could be because the passengers who choose to fly non-stop are less sensitive to ticket prices and likely consider other factors, such as time of travel, in their itinerary selection. Of the three models, the one-stop model has the least R^2 of 0.31. In this case, neither of the two indicator variables for network or low-cost carriers are significant. Like the passengers on non-stop itinerary, other factors influence passenger choices when choosing a one-stop itinerary, and exploration and inclusion of those factors in our model is part of our future work. Finally, the two-stop model has a R^2 value of 0.47,

Table 1. Coefficients of multiple linear regression for itinerary-type selection (data for fourth quarter of 2014)

Parameter	Non-stop model		One-stop model		Two-stop model	
	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value
Intercept	0.985	0	0.757	0	0.654	0
Average_market_fare	0.000155	0.1070	0.000626	0	0.000205	0
Average_market_miles	-0.000176	0	-0.000053	0	0.000108	0
Leg_car	-0.272	0	-0.00958	0.43	-0.352	0
Low_car	-0.244	0.0061	-0.0184	0.32	-0.431	0
Total_market_pax	0.0000743	0	-0.000099	0	-0.0000152	0
	$R^2 = 0.38$		$R^2 = 0.31$		$R^2 = 0.47$	

the highest of the three, and all predictor variables are significant in this case.

Having obtained this model, we compare its predicted values with the real segment data values for validation. The procedure to do so, in brief, is as follows. First, we read both the market demand and the segment demand data from the T-100 table for all origin-destination pairs in our network. The values of market demand data serve as inputs to the above regression models to estimate the number of passengers that choose each one of the three itinerary types. Using segment demand data, we identify our network topology, i.e., the actual routes in operation in the network, and distribute passengers over these routes based on their preferred itinerary-type choice. This gives us the predicted segment demand data which we then compare with actual segment demand data.



a) Predicted distribution of passengers on itinerary types **b) Scatter plot of actual vs. predicted segment demand values**

Figure 7. Comparison of predicted results with actual data from T-100 tables

Figure 7a shows the comparison of distribution of passengers on different itinerary choices in our network; the values within the bars in this figure are the fraction of total itineraries of that type. We notice that our models under-predict the number of passengers flying non-stop and over-predict those flying on one- and two-or-more-stop itineraries. While this figure shows the predicted demand distribution on itinerary types, our ultimate objective is to be able to forecast demand on every segment in the network. Figure 7b shows the scatter plot of actual segment demand values from DB1B market table versus those from our predictions. The correlation between observed and predicted data is 0.97. Thus, we conclude that the passengers' choice model estimates the distribution of market demand into different segments reasonably well. This can also be observed from the constant-variance scatter of data in this figure around the linear fit line as shown in the same figure. Figure 8 shows the density distribution of observed versus predicted segment demand; note that in this plot the horizontal axis is shown on a log-scale. This plot shows that our model tends to underestimate the number of segments with low demand values, but overestimates those that have higher demand values.

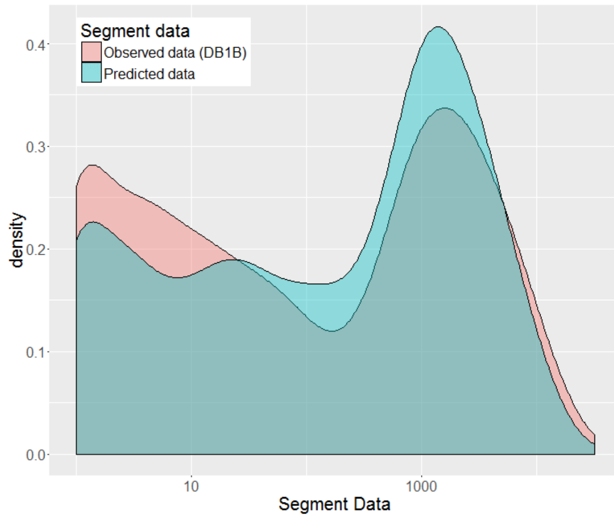


Figure 8. Density of segment data observed vs. predicted (Horizontal axis is log-scale)

V. Conclusions

In this paper we have demonstrated the development of a passengers' choice model in an air transportation systems context, which together with the airlines' route selection model [4] serve as two levels of stakeholders' hierarchical decision-making model, as shown in Fig. 1. Our approach has been to develop separate multiple linear regression models for each type of itinerary – non-stop, one-stop, and two-stop. In each one of these models, we used the average market fare, market miles, presence of legacy or low-cost carriers, and the total number of market passengers as the explanatory variables. In the results presented in this paper, we observed that our models tend to over-estimate the number of passengers on one- and two-or-more-stop itineraries while under-estimating those on non-stop itineraries. Further studies will help refine these models to include additional explanatory variables and thereby improve accuracy of our predictions.

While these models are useful, the airline passengers likely use many other factors in their decision-making. Hence, we will continue to explore additional factors that can included in the model to further refine its predictive capabilities. An important consideration is that making a choice between different itinerary types amounts to a discrete choice. For this purpose, application of logistic regression, instead of linear regression as done here, would be more appropriate. The process of developing an alternate model of passengers' choices based on logistic regression is part of our planned future efforts, and given appropriate results, we would replace our linear regression model with a logistic regression model in our hierarchical framework.

Our next steps in the development of the hierarchical model include development and integration of a model to forecast demand on the network. Currently we use only the available demand from BTS database, and the addition of a demand forecast module will provide the framework with predictive capability. A third level focused on policy-making would then be added. This is the level where entities such as the FAA would be modeled. Finally, and in keeping with the overall project objectives, we are also in the process of extending our work to account for the presence of competition in the airline industry which we will model using game theory. Such an hierarchical model of decision-making will provide the ability to assess effect of network-level policies on the constituent nodes, and the effect of decisions of nodes on network performance.

Acknowledgments

The authors gratefully acknowledge financial support from the National Science Foundation through NSF CMMI-SYS grant 1360361.

References

[1] Sgouridis, S., Bonnefoy, P. A., and Hansman, R. J., "Air transportation in a carbon constrained world: Long-term dynamics of policies and strategies for mitigating the carbon footprint of commercial aviation," *Transportation Research Part A: Policy*

and Practice, Vol. 45, No. 10, 2011, pp. 1077–1091.

[2] DeLaurentis, D., "Understanding transportation as a system-of-systems design problem," *43rd AIAA Aerospace Sciences Meeting and Exhibit*, Vol. 1, AIAA Reno, NV. New York, 2005.

[3] Donohue, G., "Air transportation is a complex adaptive system: Not an aircraft design," *AIAA/ICAS International Air and Space Symposium and Exposition*, 2003.

[4] Sha, Zhenghui, M. K. D. D. and Panchal, J., "Modeling Airlines Decisions on Route Selection Using Discrete Choice Models," *American Institute of Aeronautics and Astronautics (AIAA)-15th Aviation, Technology, Integration, and Operations Conference (ATIO)*, 2015.

[5] of Transportation Statistics, B., "Air Carriers : T-100 Domestic Market (U.S. Carriers)," online resource, Oct. 2014, http://www.transtats.bts.gov/Fields.asp?Table_ID=258.

[6] Davendralingam, N., Sha, Z., Moolchandani, K., Maheshwari, A., Panchal, J. H., and DeLaurentis, D., "Scientific Foundations for Systems Engineering – Challenges and Strategies," ASME, Boston, MA, 2015.

[7] Bhadra, D. and Hogan, B., "A Preliminary Analysis of the Evolution of US Air Transportation Network," *American Institute of Aeronautics and Astronautics (AIAA)-5th Aviation, Technology, Integration, and Operations Conference (ATIO)*, 2005.

[8] of Transportation Statistics, B., "Airline Origin and Destination Survey (DB1B)," online resource, Nov. 2015, http://transtats.bts.gov/Tables.asp?DB_ID=125&DB_Name=Airline Origin and Destination Survey (DB1B) &DB_Short_Name=Origin and Destination Survey.

[9] Bhadra, Dipasis, H. B., "A Preliminary Analysis of the Evolution of US Air Transportation Network," *AIAA ATIO 2005*, AIAA, Arlington, VA, 2005.

[10] Train, K., *Discrete Choice Methods with Simulation*, Cambridge University Press, 2009.

[11] Ben-Akiva, M. E. and Lerman, S. R., *Discrete choice analysis: theory and application to travel demand*, Vol. 9, MIT press, 1985.

[12] Sha, Z. and Panchal, J. H., "Estimating Local Decision-Making Behavior in Complex Evolutionary Systems," *Journal of Mechanical Design*, Vol. 136, No. 6, 2014, pp. 061003.

[13] Sha, Z. and Panchal, J., "Estimating linking preferences and behaviors of autonomous systems in the internet using discrete choice models," *2014 IEEE International Conference on Systems, Man, and Cybernetics Conference*, San Diego, CA., 2014.

[14] Chen, W., Hoyle, C., and Wassenaar, H. J., *Decision-based design: Integrating consumer preferences into engineering design*, Springer Science & Business Media, 2012.

[15] Administration, F. A., "Passenger Boarding (Enplanement) and All-Cargo Data for U.S. Airports," online resource, Oct. 2014.