Modeling Airline Decisions on Route Planning Using Discrete Choice Models

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

We propose a model for the airlines' decisions on route planning, i.e., the decision on selecting which route to add and delete, using discrete choice random-utility theory. The central hypothesis is that a discrete choice model can effectively model the airlines' decisions on route selection , and thereby help model the evolution of the air transportation network. We first model the airlines' utility function as a linear function of decision variables with interaction effects. The decision of route selection is then modeled using a binary choice model derived from the utility function. The preferences for each variable in the utility function are estimated using historical datasets. Advantages of this approach include the ability to use statistical techniques to quantitatively construct decision models as well as to account for the uncertainty in unobserved attributes of the decision model. The proposed model helps predict the airlines' decisions on routes addition and deletion which affect the network topology of air transportation and its future evolution. This capability can be beneficial to other stakeholders, such as Federal Aviation Administration, who may need to make their decisions in response to those made by the airlines, but do not have access to the airlines' true decision models.

I. Introduction

AIRLINES' decisions on route selection are, along with fleet planning and schedule development, the most important decisions they make [1]. Route selection decisions are concerned with where to offer service subject to fleet availability constraints. Such decisions are not only vital to airline profitability, they also have other network-wide effects such as propagation of delays, robustness of network to service disruptions, and the network's traffic flow capacity. This is why stakeholders like the Federal Aviation Administration (FAA) may want to access airline decisions to model their responses. For example, the FAA may want to understand how its investments at individual airports will translate to network restructuring, leading to reduction of network-wide delays. However, airline decision-making strategies are not publicly known and likely depend on a number of different criteria. Consequently, many researchers have tried to replicate these strategies using many different approaches.

The most common approach for airline route selection decisions is to model them as integer programming problems with an objective function of maximizing profit or traffic flow. Work done by Lederer and Nambimadom [2] is an example of this type of approach since they study choices of different network designs and schedules using profit maximization as the objective function. Jaillet, et al. [3] present three integer linear programming problems for airline network design, and propose heuristics for designing capacitated networks and routing problems. Magnanti and Wong [4] review some of the integer programming based approaches to network design and also describe both discrete and continuous choice models and algorithms.

Using a network analysis approach to study network evolution is a common alternative to the optimization approaches mentioned above. Song et al. [5] present a model that splits the air transportation network into two tiers and use it to estimate network evolution. Their model uses available demand data for its analysis with the result that it has to rely on future demand estimation for prediction of network evolution. Kotegawa [6] uses machine learning algorithms to study network evolution using patterns derived from historical data. He compares algorithms based on logistic regression, random forests, and support vector machines and shows high route addition and removal forecast accuracies.

The airlines' decision-making models are not only likely very elaborate with many criteria but proprietary as well, meaning that stakeholders other than the airline are left guessing and do not have access to the airline's true decision-

[*] Graduate Research Assistant, School of Mechanical Engineering
[†] Graduate Research Assistant, School of Aeronautics and Astronautics, AIAA Member
[‡] Graduate Research Assistant, School of Aeronautics and Astronautics
[§] Graduate Research Assistant, School of Mechanical Engineering
[¶] Assistant Professor, School of Mechanical Engineering
[||] Associate Professor, School of Aeronautics and Astronautics, AIAA Member

making models. To avoid the elaborate criteria, decision-making models in the literature make some simplifying assumptions. Under the rationale that economic considerations, especially profit-maximizing, are the primary drivers of decision-making for most airlines, the above mentioned studies offer ways of replicating airline decisions. There are two types of models: Models based on network optimization commonly use a single objective function such as cost minimization, while models based on network analysis usually rely on network metrics for their analysis, ignoring many airline-specific criteria in the process.

In this paper, we model the route selection problem as discrete choice decision problem using random-utility theory based approach. The objective is not to offer an alternative model to the airlines' true decision-making model, rather propose one that approximates the airlines' decisions based on known data. The two schematics shown in Fig. 1 help clarify our objective further. For an airline *i*, the schematic on the left of Fig. 1 shows its true decision-making model, $D_i(t)$, as a function of time *t*. This model takes into account the airline's decision criteria, *x*, as well as external factors that the airline does not directly control, and gives its decisions as the output, y_i. The schematic on the right of Fig. 1 shows our approximation of the airline model as $D_i'(t)$. We assume that the market demand, operating costs, and the hub or non-hub nature of airports in the network are the input criteria ($\mathbf{x}'$) to our model along with the known airline decision obtained from available data, y_i. This model gives the airline's predicted decisions as y_i'.

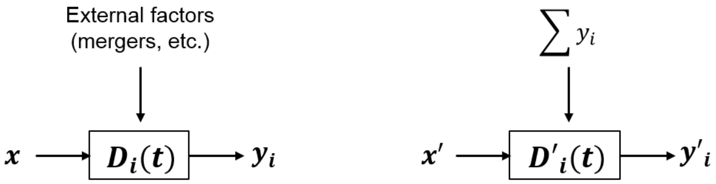


Figure 1. Schematic of actual airline decision model (left) vs. predicted decision model (right)

Given the proposed utility function, we quantitatively evaluate the impact of each variable on airlines' route planning decisions based on the historical dataset. The resulting decision-models can be adopted to implement the local mechanisms (e.g., route addition/deletion) in an air transportation network topology generator. This work improves on prior work that used machine learning on network topology time series to estimate network evolution [6]. That approach however neglected the true mechanism behind network restructuring: decision-making of airlines. Thus, this is a problem of decision-makers (agents) who act to determine the network topology. We are modeling the decisions of agents who are not nodes in the network but influence the network topology. The **research question** addressed in this paper is: How can airlines' preference structure for service network evolution be derived from available air transportation network data? The **central hypothesis** is that a discrete choice model can be used to effectively model the airlines' new route selection decisions.

The following section presents our technical approach to modeling the airline decision-making strategies, including a brief description of the discrete choice random-utility theory. Following this are studies on assessment of effects of airline decisions on network performance.

II. Technical Approach

A. Overview of discrete choice models

The core of the proposed approach is the discrete choice analysis based on the assumption of random-utility maximization. In discrete choice analysis, it is assumed that the decision-maker has complete knowledge about his/her own utility U_i when choosing the alternative *i*. This utility consists of two parts, the observed utility V_i and unobserved utility ϵ_i. This formulation is from the researcher's point of view because the researcher does not know the decision maker's utility completely, but is only able to observe the choices made by the decision maker. The observed utility is usually modeled as parameterized function of a set of explanatory variables that would affect the decision-maker's decisions. These variables are deterministic in nature from the researcher's point of view, and can be identified through different ways, such as survey and existing literature. The unobserved utility captures the randomness due to unobserved attributes, unobserved variations among decision makers, measurement errors, functional misspecification and bounded rationality of decision makers [7]. A common formulation for observed utility V_i is in linear form:

$$\mathbf{V}_i = \mathbf{x}'\beta_{\mathrm{i}} = \beta_{i1}x_{i1} + \beta_{i2}x_{i2} + ... + \beta_{ik}x_{ik} \qquad (1)$$

where $x = (x_i, x_2, ..., x_n)'$ is a vector that contains n explanatory variables for the utility, and $\beta_i = (\beta_{i1}, \beta_{i2}, ..., \beta_{ki})'$ is the set of weights of each variable that represents decision-maker's preference. Thereby, V_i quantitatively captures the decision-maker's preference structure. Note that the explanatory variables can be either alternative-specific or decision-maker-specific. We assume that each agent has complete knowledge about his/her own utility U, but since we can only observe the observed utility V which could be different from the total utility U, we must separate U into the observed utility V and the unobserved utility. The total utility is then the sum of the observed utility and the unobserved utility:

$$U_i = V_i + \epsilon_i \tag{2}$$

Therefore, the total utility is random and the decision-making process is modeled as a nondeterministic process. With random-utility maximization, the decision-maker chooses alternative *i* rather than *j* if and only if $U_i \geq U_j, \forall j \neq i$. Thus, the probability of decision-maker choosing alternative *i* is:

$$P_i = P(U_i \geq U_j) = P(V_i - V_j \geq \epsilon_j - \epsilon_i) \quad \forall j \neq i \tag{3}$$

This probability, P_i, is the cumulative distribution of $\epsilon_j - \epsilon_i$, thus can be determined once the density function of ϵ is specified. In this paper, a logit model with the following form is adopted:

$$P_i = \frac{e^{x'\beta_i}}{\sum\limits_{j=1}^{J} e^{x'\beta_j}} \tag{4}$$

In airline route planning context, whether to establish a route between a pair of cities can be regarded as a binary choice for the airlines (1=Yes, 0=No). With the above logit choice model in Eq. 4, the probability that an airline chooses to create the route is:

$$P_1 = \frac{e^{x'\beta_1}}{e^{x'\beta_1} + e^{x'\beta_0}} = \frac{e^{x'(\beta_1 - \beta_0)}}{1 + e^{x'(\beta_1 - \beta_0)}} = \frac{e^{x'\beta}}{1 + e^{x'\beta}} \tag{5}$$

where $\beta = \beta_1 - \beta_0$ captures the difference in preference of choosing yes and choosing no. The parameters β can be readily estimated from the choice data through statistical estimation techniques, such as maximum likelihood estimation and hierarchical Bayesian estimation. The key is to construct the preference structure in terms of the utility function V through the identification of the explanatory variables (i.e., x') which could have impact on the airlines' decision-making. In the following sections, after establishing the dataset, we discuss what explanatory variables are chosen in this paper to construct the airline's preference structure and the approaches for obtaining the data for these variables.

B. Proposed approach

The rest of the paper is organized in the order of execution of our process of utility modeling and subsequent analysis. The first step in our process was that of exploring the decision variables to be included in our predicted airline decision-making model. Based on our literature survey, we decided on using the demand, operating cost, and the hub or non-hub nature of the terminal airports as our decision variables. Following this selection, was the process of gathering data described in section III. This was followed by the application of discrete choice analysis, described in sections IV and V, and finally, model validation. Figure 2 gives the overview of these steps.

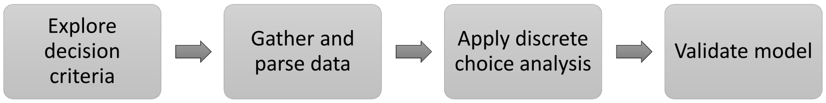


Figure 2. Schematic of our proposed approach

III. Data Collection and Preparation

To set up our models, we made use of the data on airline operations and economics in the US market available from the Bureau of Transportation Statistics. Then, we applied regression on the data on travel demand and operating costs of the airlines to estimate the airlines' utility function for the purpose of route selection. The details of regression are given in section V; this section describes the process of data collection for the purpose of analysis.

A. Establishing set of airports and initial network

All of our data on airline operations and economics is available from the Bureau of Transportation statistics website. Specifically, forms T-100 and T-2 provide data on airline operations, while Schedule P-5.2 provides data on their economics [8, 9, 10].

In addition to the BTS website, we make use of the list of airports provided by the Federal Aviation Administration (FAA) to select the set of airports in our network [11]. This dataset categorizes airports into those with primary and non-primary commercial operations. The airports in the former category are further subdivided as large, medium and small hubs, and non-hubs based on the number of enplanements at the airport. We start by selecting the set of 134 largest airports in the FAA dataset, which are listed as either large, medium or small hubs. After analyzing the market data from the BTS, we removed Phoenix-Mesa Gateway Airport (IWA), and Saipan International Airport (GSN) from consideration as there was no segment data (see [8] for details) available for these airports. Together, we find that the remaining 132 airports constitute over 96% of all enplanements in the US during calendar year 2013 [8].

With the 132 major airports, we extract the network topology from 2003 to 2009 based on the BTS T-100 segment dataset [8]. An edge in the network exists if there is a route between two airports in a calendar year. In this paper, we filter out routes which had either zero demand or were served by aircraft with the freight configuration as reported in the BTS database. Further, we include only those routes which are classified as those in the 'domestic' category in the BTS table. To account for those routes which may not be commercially operational, we count only those which had at least one flight in any consequent eight week period in a given year. As an example, these filters, when applied to the 2009 network data, result in the final set of 2015 routes (edges) out of the total 8646 routes required for the network to be fully connected.

From an airline perspective, hubs are the transfer points to regulate the traffic between the airports not connected directly in their network. We compare the list of 30 large hubs in the FAA dataset with the hubs listed by the three largest full-service airlines in US, viz. United Airlines (UA), American Airlines (AA) and Delta (DL) [12, 13, 14]. Based on this comparison, we finally select a set of 21 airports that are fixed as hubs in this work. Additionally, we add Chicago Midway to this list because of its large volume of operations, resulting in a final list of 22 hub airports.

Once the network topology is obtained, the next step is to extract the data related to the explanatory variables identified from the preference structure formulation. In the following subsection, we discuss how the data regarding the demand, cost and airport hub information is obtained from various data sources.

B. Demand and cost data

The raw market demand file from the BTS has more than 0.2 million entries for any given year. While we use segment data from the T-100 table to identify the network, the market data provides us with the ability to identify the need for new routes to be added. In other words, a high market demand between two city pairs would be a reason for the airline to add a direct flight between the two cities. This data table contains market demand estimates for all possible routes (even those outside the 132 airports taken in the analysis), month, and aircraft type, for a directed network. Since we are interested only in the market demand of our undirected reduced network, we filter out the demand for only the routes that we are interested in. For a particular route, say between airport A and B, we sum the market demand from A to B and B to A for all months and all airlines to get the actual market demand between the origin-destination (O-D) airport pair. This is based on the rationale that over long intervals (like one year considered in the paper), the directed demand would average out to be the same in either direction.

The BTS website reports zero demand between several airport pairs. This may either be due to reporting errors or due to no potential demand between the O-D markets. To account for this 'NULL value' in the reported data we directly used segment demand of the same year for the same route. This is justified in the sense that the market and segment demand is reasonably close on many routes.

The direct operating cost (DOC) of an airline is the second variable in our decision model. For every route in the current year, we calculate the DOC for all airlines that operated on a given route, weighted by their number of operations. We then use this calculated value as our estimate of operating cost on that route for that year. For the purpose of our analysis, we need an estimate of cost on all routes in the selected choice set, described below, including for the years in which a route may not be present. For such cases where the values are missing, we use an 'airline cost index' based on data from the Statistical Abstract of United States to estimate them [15]. Note that at the time of our analysis, we could only obtain cost indices until 2009, which was the reason we restricted our analysis years until 2009.

C. Hub and non-hub

We divide the routes into categories based on hub characteristics of the O-D airport pair. These categories are a) both the airports as hubs, identified by the label ‘2’, b) at least one airport as hub, labeled ‘1’ and c) both airports as non-hubs, labeled ‘0’. We assigned a value of 1 or 0 to each city based on whether the city qualifies as a hub or not. For example, for the network formed by 22 hubs and 112 non-hubs, there are 2062 routes in the year of 2005. In such network, there are 644 routes at hub level 0, 1198 routes at hub level 1, 220 routes at hub level 2.

D. Determine the choice set and observations of choices

We define the choice set for addition or deletion as the set of routes that are available for the airline in our model to add or delete, respectively. If a route was in operation for any number of years in the range of years under consideration, then it forms part of the choice set for addition, because, we reason that if any of the airlines found a route feasible for operation in the past, then any airline is likely to consider operating on it in the future provided sufficient demand. The choice set for route deletion is the set of routes that exists in the current year. In doing so, we found that some values of segment demand were 0, and some DOC values were ‘NA’, and hence, we filtered out these routes from our choice set. Note that our modeled airline makes use of the utility function approach presented herein for its decision on route addition or deletion. No factor outside of the model is used by the airline for its choice.

In order to analyze the airline’s decision-making preferences on route selection, we need observations of which routes are added or deleted yearly, i.e., the choices made by airlines. To obtain the observations, we perform the edge dynamic analysis on network data from two consecutive years. For instance from 2005 to 2006, with the networks after filtering, the choice set for adding has 634 routes in which 170 routes are selected. The choice set for deleting has 2062 routes (i.e., the number of routes in 2005), there are 138 routes are selected to delete. Table 1 lists the size of choice set and number of observations from 2002 to 2009.

Table 1. The choice set and number of observations of routes addition and deletion from 2002 to 2009

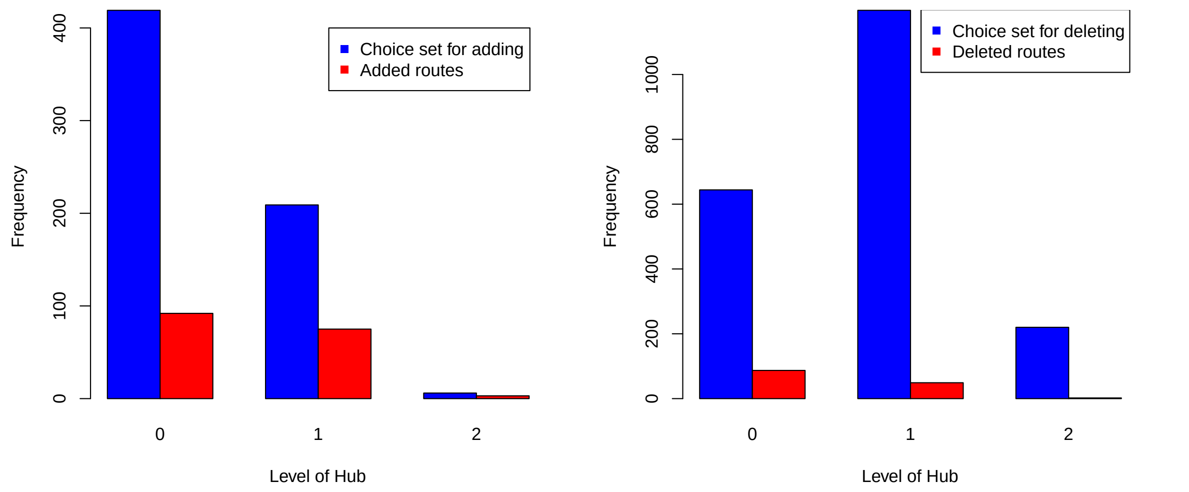
Year	02 - 03	03 - 04	04 - 05	05 - 06	06 - 07	07 - 08	08 - 09
Choice set for addition	905	868	719	634	603	486	478
Number of routes added	115	240	181	170	233	146	96
Choice set for deletion	1789	1831	1977	2062	2094	2210	2218
Number of routes deleted	75	94	96	138	116	138	299

IV. Data Analysis for Decision Variables in the Choice Set

In this section, we present the data analysis on the explanatory variables of the routes that are chosen for adding and deleting. The aim of this analysis is to have a preliminary understanding on the characteristics of routes being selected. The insights obtained from the analysis will guide the formulation of the discrete choice models. The analysis is performed based on the data of air transportation network’s evolution from two consecutive years. In this section, the results for the evolution of 2005-2006 are presented as an illustrative example. The analysis on other evolution instances follows the same approach.

With the network of our selected 132 airports, there are 2062 routes in 2005 and 2094 routes in 2006. From 2005 to 2006, 170 new routes were added and 138 routes were deleted across the US air transportation network. Using the method proposed in Section III.D, we identified that the size of choice set for adding routes is 634, and the size of choice set for deleting routes equals to the number of routes in 2005, i.e., 2062. Figures 3a and 3b show the distribution of hub levels of the routes being added and deleted among the correspond choice set, respectively. Figure 3a shows that most of the routes in the choice set of adding routes are at hub level 0. However, in the choice set for deleting, most of routes are at hub level 1. Among 634 routes for adding, there are 419 routes at hub level 0 and 92 are added, 209 routes at hub level 1 and 75 are added, and 6 routes at hub level 2 and 3 routes are added. This indicates that as the level increases, the likelihood of being chosen increases. Among 2062 routes for deleting, there are 644 routes at hub level 0 and 87 routes are deleted, 1198 routes at hub level 1 and 49 routes are deleted, and 220 routes at hub level 2 and 2 are deleted. For the routes deleted, the impact of hub levels on whether a route is selected does not follow the trend as we observed in the routes for adding.

Figures 4 and 5 show the distribution of continuous variables – market demand, unit cost and distance – of the routes that are selected to add and delete. For the market demand, it is observed that most of the added and deleted routes have a market demand less than 10000 (see Figures 4a and 5a). However, for the routes with demand greater than 20000, 33 routes are added and 3 routes are deleted. This implies that even if most of the added or deleted routes



a) The distribution of level of hub of the choice set for adding routes and the routes are added

b) The distribution of level of hub of the choice set for deleting routes and the routes are deleted

Figure 3. The distribution of level of hub associated to the routes being added (3a) and deleted (3b)

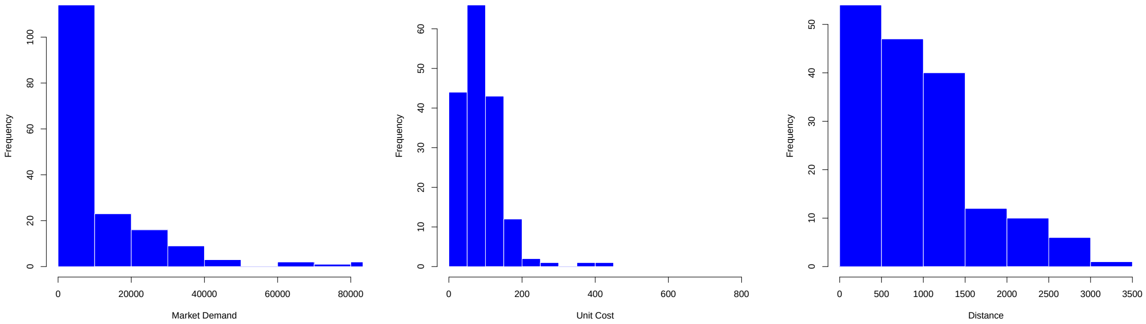
have small amount of demand, the routes with high demand have higher probability to be added than deleted. Table 2 shows that for routes addition, the added routes on average have demand of 10357.92, greater than 1224.46 average for the market demand of routes not added but in the choice set for addition. This indicates that routes with higher demand have a greater likelihood of being added. Similarly, the means of the routes being deleted and not deleted indicate that the smaller the demand on a route, the less likely a route is deleted. This conclusion is validated by the regression analysis in Section V.B, and the quantitative insights of the effect of the demand are obtained.

Table 2. Descriptive statistics of data associated to the chosen routes

Class of decision	Variables	Routes added		Routes not added	
		Mean	Std. Dev.	Mean	Std. Dev.
Routes addition	Market demand	10357.92	16268.78	1224.46	3779.49
	Unit Cost	88.18	60.32	97.55	109.06
	Distance	937.05	697.44	912.35	735.72
	Variables	Routes deleted		Routes not deleted	
		Mean	Std. Dev.	Mean	Std. Dev.
Routes deletion	Market demand	1499.99	5072.38	151664.35	179037.11
	Unit Cost	90.88	82.81	71.802	46.12
	Distance	738.48	591.99	920.48	681.02

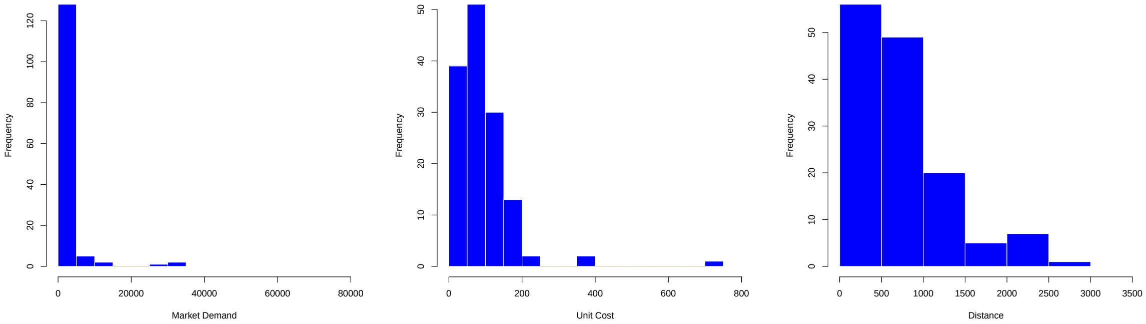
The distribution of unit cost of the routes being added has a very similar shape as the distribution of the routes being deleted. Most of the routes added or deleted have low cost. This is counter intuitive because it is assumed that routes with high cost would have high probability to be deleted. However, such phenomenon is not observed from the data. The mean of cost shown in Table 2 for the routes added and deleted reflects that the effect of cost on routes selection is not so significant. One possible reason for this is that most added or deleted routes belong to hub level 0 which both airports are not hubs. Such routes are most of time for flights that have a low running cost. These flights are added or deleted frequently based on many factors, such as seasonal variation in demand.

Similarly, it is observed from Figs. 4b and 5b that most of the added and deleted routes have short distances. Table 2 shows that the distance of the routes being added have the similar distance with the routes that are added. This indicates that distance may not be an important factor affecting the decision on whether a route will be added or not. This is intuitive since as long as the demand on a route is high and the cost is low, the airline could make profit and hence, such a route would be established. In the next section, the effect of distance on route addition and deletion will



a) Distribution of market demand **b) Distribution of unit cost** **c) Distribution of distance**

Figure 4. The distribution of continuous explanatory variables associated to the routes being added



a) Distribution of market demand **b) Distribution of unit cost** **c) Distribution of distance**

Figure 5. The distribution of continuous explanatory variables associated to the routes being deleted

be quantitatively analyzed.

V. The Airlines' Preference Structure

In this section, we describe the modeling of airlines' decisions on addition and deletion of routes using the explanatory variables described above. Table 3 shows the four variables identified to construct the preference structures. In this paper, the model established is a binary choice model in Eq. 5. Equation 6 shows the preference structure (i.e., the utility models) constructed, which follows from Eq. 1. In this model, we assume that airlines' decisions on route planning for the next year depend upon the market demand of a route in the current year, the distance of routes, the unit cost of running routes and whether the routes connect two hubs or not. An important assumption is that air routes planning is demand-pull but not supply-push.

Table 3. Summary of explanatory variables

Variables	Description
x_1	Hub indicator. 0 both airports are not hub; 1 at least one airport is hub; 2 both airports are hubs.
x_2	Potential market demand (unit: 1000 passengers)
x_3	Potential unit cost (unit: 1000 dollars/nautical mile/seat)
x_4	Distance (unit: 1000 miles)

This model consists of both qualitative and quantitative variables. The qualitative variable is the level of hub, and the quantitative variables are demand, cost, and distance. Because of the existence of both qualitative and quantitative variables in utility function, we investigate the interaction effect of explanatory variables by including the cross-product terms in the model, as shown in Equation 6. The purpose is to see if the effect of quantitative variables, such as demand, on the airlines' decisions would be different for routes at different level hub.

$$V = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_1 x_2 + \beta_6 x_1 x_3 + \beta_7 x_1 x_4 \tag{6}$$

Since x_1 is a qualitative variable that has three levels, the parameters β_1, β_5, β_6 and β_7 for each level will be estimated. However, when performing the regression, we must specify a reference level with parameter set to be 0, and the parameters of other two levels are then estimated by comparing with the reference level. In this paper, the parameters are estimated through the maximum likelihood estimation.

A. Discrete choice analysis on routes selection

With the data on how routes are actually added and deleted each year, the regression analysis is performed to investigate whether these variables significantly affect airlines' decisions and which variables are the most influential factors that airlines consider for planning routes in the next year. In this section, we interpret the regression results by using the results obtained from the network evolution from 2005 to 2006. Similar regression is performed for other evolution instances from 2002 to 2009. The results of each evolution reflects the dynamics of the decisions over time. Detailed discussion on the dynamics of decisions is presented in Section V.C.

Tables 4 and 5 show the results of estimation of the decision-making preferences of airlines on routes addition and deletion from 2005 to 2006, respectively.[a] As to the overall model fit of the two models, the log likelihood at convergence (i.e., -295 for the model of route addition and -420 for the model of route deletion) both increase as compared with the ones (i.e., -215 and -64.3) in the null model (the model without parameters). This improvement is validated with the likelihood ratio test [16]. The resulting p-value is less than 0.0001, indicating the improvement is statistically significant. The overall model fit is quantified using the MacFadden's adjusted R^2 [17]. The model of routes addition has $R^2 = 0.23$ and the model of routes deletion has $R^2 = 0.82$. This shows that the model for routes deletion has better performance on explaining and predicting routes planning.

Table 4. Results of estimated decision-making preference for routes addition from 2005 to 2006

Variables	Parameter	Mean	p-value	Odds ratio
Intercept	β_0	-1.42	<0.001	0.24
Hub level 1	β_{11}	0.90	0.031	2.46
Hub level 2	β_{12}	2.82	0.027	16.71
Market demand	β_2	0.094	<0.001	1.10
Unit cost	β_3	-0.56	0.70	0.57
Distance	β_4	-0.11	0.63	0.89
Demand at hub level 1	β_{51}	0.24	<0.001	1.27
Demand at hub level 2	β_{52}	1.19	0.18	3.28
Cost at hub level 1	β_{61}	0.814	0.66	2.26
Cost at hub level 2	β_{62}	-113	0.24	1.1e-49
Distance at hub level 1	β_{71}	-1.10	0.0056	0.33
Distance at hub level 2	β_{72}	4.89	0.35	132.52
	Overall Model fit			
Log likelihood at zero	-295			
Log likelihood at convergence	-215			
Likelihood ratio test	p-value <0.0001			
McFadden's adjusted R^2	0.23			

In the model of routes addition, route in which both airports are non-hubs (i.e., hub level 0) is set as the baseline level. The p-value for the parameter of hub level 1 is 0.031. This indicates that the mean of the estimated parameter is significantly different from 0. The value is 0.90, and the positive sign indicates that the probability of adding a route at hub level 1 is greater than the probability of adding a route at hub level 0. To better interpret this result, this coefficient is translated to the odds ratio which is the ratio of the probability of adding a route (P) to the probability of not adding a route ($1 - P$), i.e., $\frac{P}{1-P}$. The odds ratio of hub level 1 is 2.46. This means the probability of adding a route between

[a] The correlation analysis is performed before the regression analysis. There is no significant colinearity between the continuous explanatory variables.

Table 5. Results of estimated decision-making preference for routes deletion from 2005 to 2006

Variables	Parameter	Mean	p-value	Odds ratio
Intercept	β_0	1.18	<0.001	3.27
Hub level 1	β_{11}	-0.08	0.89	0.92
Hub level 2	β_{12}	-2.82	0.02	0.06
Market demand	β_2	-0.22	<0.001	0.80
Unit cost	β_3	-3.34	0.13	0.04
Distance	β_4	0.21	0.56	1.23
Demand at hub level 1	β_{51}	-1.39	<0.001	0.25
Demand at hub level 2	β_{52}	0.18	0.38	1.19
Cost at hub level 1	β_{61}	0.28	0.94	1.32
Cost at hub level 2	β_{62}	36.7	0.17	8.35e+15
Distance at hub level 1	β_{71}	1.09	0.16	2.97
Distance at hub level 2	β_{72}	-1.75	0.23	0.17
	Overall Model fit			
Log likelihood at zero	-420			
Log likelihood at convergence	-64.3			
Likelihood ratio test	p-value <0.0001			
McFadden's adjusted Rˆ2	0.82			

two airports that at least have one hub is 2.46 times greater than the probability of adding a route with both airports are non-hubs. Similarly, the probability of adding a route between two airports that are both hubs is 16.71 times greater than the probability of adding a route with both airports are non-hubs. These results are to be expected as evidenced by the large amount of literature that describes the merits of having hubs in the network.

The p-value for β_2, i.e., the preference parameter for demand, is less than 0.001. This indicates that the effect of demand on routes addition is statistically significant. The odds ratio of demand is 1.10, which means with 1000 passengers increase between two airports, the probability of adding a route is 1.10 times greater. The p-value of the estimated parameters of demand and cost are both greater than the level of significance 0.1. This result indicates that effects of demand and cost are not statistically significant. As to the interaction effect, we observe that the effect of demand at hub level 1 is significantly different from the effect of demand at hub level 0. In terms of the odds ratio, the effect of 1000 passengers increase of the routes at hub level 1 on routes addition is 1.27 times greater than the effect of demand on the routes at hub level 0. As shown in Table 4, there is no significant interaction effect between the cost and hub level indicating the effect of cost on routes addition has no difference at each level of hub.

Table 5 shows the results of the decision-making preference estimated for the routes deletion from 2005 to 2006. As the p-values shown in the table, besides the intercept, only the parameters of hub level 2, demand and the interaction between demand and hub level 1 are statistically significant. The estimated parameter for demand $\beta_2 = -0.22$. The negative sign indicates that the decrease of demand increases the probability of deleting a route. In terms of odds ratio, with 1000 passengers decreases on a route, the probability of deleting such route is 1.25 (=1/0.8) times greater than the probability of not deleting such route. The odds ratio for hub level 2 is -2.82. It means that the probability of deleting a route where both airports are non-hubs is 16.67(=1/0.06) times greater than the probability of deleting a route with both hubs. In addition, it is observed that the effect of demand on routes deletion is different at hub level 0 and level 1. Specifically, for deleting a route, the effect of 1000 passengers increase on the routes at hub level 0 is 4 times greater than the effect of demand on the routes at hub level 0.

In summary, with the regression analysis, it is found that 1) the demand is a significant variable that affects the probability of adding or deleting a route between two airports, 2) the decision on adding or deleting a route is different at different levels of hub. The introduction of hub levels into the decision models help understand the how airlines do the routes planning in more detail, 3) the effect of demand at different level of hub on routes planning is tested to be statistically different. However, in the analysis of evolution from 2005 to 2006, it is observed the cost and distance are not influential factors on the decisions of routes planning. In the Section V.C, the variables investigated in this section are analyzed in different network evolution instances to see if the effect of these variables are consistently significant (i.e., the resulting p-values are consistently high or not), and whether the airline's preferences on specific variables are consistent (i.e., the signs of parameters are consistently positive or negative).

B. Model validation

The validation of the models is performed by using the estimated parameters (i.e. the β values) to predict the probability of whether a route will be added or not in the evolution from 2005 to 2006. In terms of the performance of predictability, the Hosmer and Lemeshow test is adopted. The null hypotheses is that the predicted probabilities obtained from the model match the real observation. Thus, we expect p-value such that null hypothesis cannot be rejected. The test results are shown in Table 6. The results indicate that all the models have high p-value as compared with the level of significance of 0.1. For example, in model of route addition, the test statistic is 8.61 and the degree of freedom is 8. With the chi-square distribution, the p-value is 0.38. Therefore, the null hypothesis cannot rejected and we conclude that the predicted probabilities obtained from the model match the real observation. To get more insights on the models' predictability, the analysis on sensitivity and specificity [18] is performed. Table 7 shows the analysis on the model for routes addition, in 170 routes that are selected to add, the model predict correctly for 67 routes and fails to predict on 103 routes. The percentage of correctness is 39.4%. In the 464 routes that are not selected to add, the model falsely predict that there are 11 routes will be added, but successfully predict 453 routes. The percentage of correctness is 97.6%. Overall, the model has a 82% correctness on predicting the routes addition in the year 2006. Similarly, the analysis is performed on the model for routes deletion. The results show that the model has an overall percentage of correct prediction at 97.5%. This result shows the high performance of the model for predicting the decisions on adding and deleting routes. Especially, the model for routes deletion has a better performance in terms of the percentage of predicting correctness. It could be that when adding routes, airlines are more risky since the route is open for the taking, and the airline thinks long and hard about whether to enter or not. On the other hand, the method for deletion might be simpler: Is the route profitable? If yes, keep; if not, delete. Hence the much higher model success, which may reflect less uncertainty and thinking in routes deletion.

Table 6. Hosmer and Lemeshow goodness of fit test

Model for route addition	Chi-squared statistic: 8.61 Degree of freedom: 8 p-value = 0.38
Model for route deletion	Chi-squared statistic: 0.71 Degree of freedom: 8 p-value = 0.9995

Table 7. The observed and the predicted frequencies for routes addition by logistic regression with the cutoff probability of 0.50

Observed	Predicted		% Correct
	Added	Not added	
Chosen for adding	67	103	39.4%
Chosen for not adding	11	453	97.6%
Overall % correct			82%

Table 8. The observed and the predicted frequencies for routes deletion by logistic regression with the cutoff probability of 0.50

Observed	Predicted		% Correct
	Deleted	Not deleted	
Chosen for deleting	124	14	89.9%
Chosen for not deleting	38	1886	98.1%
Overall % correct			97.5%

C. Decisions over time

In this section, the regression analysis with the model in Eq. 6 is performed on 7 evolution instances of air transportation network from 2002 to 2009. The purpose of this study is to evaluate the model consistency and on the other hand to investigate how airlines' preferences on route planning would change overtime. The method is as follows. We use the proposed approach to estimate the airlines' preferences in each year. The regression analysis provides the estimated parameters for explanatory variables and the corresponding p-values. To check if the effect of a specific variable on routes planning is consistently significant, we count the number of times that the p-values are less than 0.1 level of significance. To check if the preferences are consistent, we count the number of times that the estimated parameters with the same sign. Table 9 shows the results on frequency of p-values that are less than 0.1 and the consistency of preferences for two models. The results show that the effect of demand has been always significant in the evolutions from 2002 to 2009. Also, in the model for route addition, the effect of demand at hub level 1 is always significantly higher than the one at hub level 0. This may reflect that airlines are more responsive to the demand dynamics on the routes with at least one hubs. This makes sense that most carriers have a hub-and-spoke network structure and route most passengers through their hub airports.

Table 9. Frequency that the estimated parameters are statistically significant in 7 evolution instances (level of significance = 0.1)

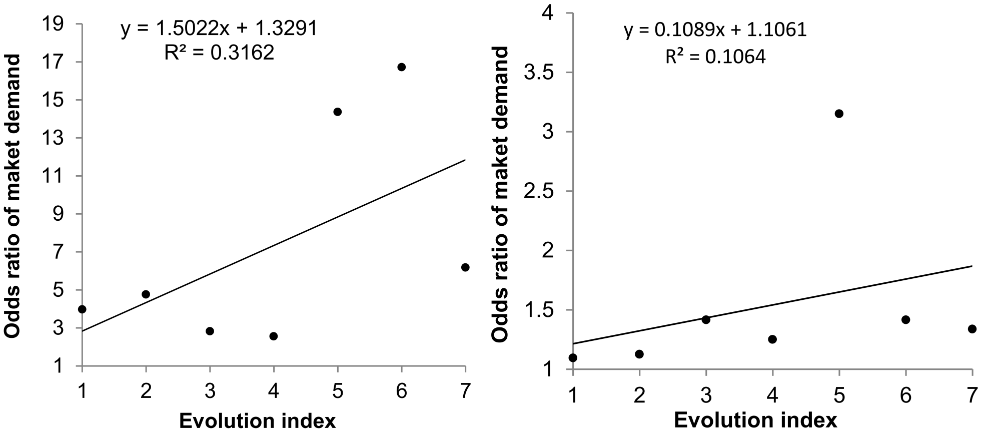
Variables	Model for adding routes		Model for deleting routes	
	Frequency of significance	Preferences consistency	Frequency of significance	Preferences consistency
Intercept	7	0+7-	3	5+2-
Hub level 1	4	5+2-	4	3+4-
Hub level 2	3	6+1-	2	2+5-
Market demand	7	7+0-	7	0+7-
Unit cost	2	6+1-	4	5+2-
Distance	4	1+6-	4	5+2-
Demand at hub level 1	7	7+0-	5	3+4-
Demand at hub level 2	3	1+6-	2	7+0-
Cost at hub level 1	1	3+4-	2	2+5-
Cost at hub level 2	0	2+5-	0	2+5-
Distance at hub level 1	3	3+4-	3	6+1-
Distance at hub level 2	0	5+2-	0	3+4-

From this analysis, it is also observed if a variable has been found to be consistently significant in affecting the airlines' decision, the airlines' preferences toward such factor likewise consistent; however, the converse is not necessarily true. For example, the estimated parameter of the interaction effect between demand and hub level 2 in the model for routes deletion is found to be always positive in the 7 evolutions. But the corresponding p-values are only twice less than 0.1. Since airlines' preferences in realistic do not change frequently overtime, to keep conservative in this paper, we only analyze the estimated preferences that have been perfectly consistent (preference consistence = (0+7-) or (7+0-)) and statistically significant (frequency of p-value $\leq 0.1 = 7$). Thus, only the estimated parameter for the demand is analyzed.

Figure 6 shows the mean of estimated parameter of demand at each evolution from 2002 to 2009. The Figure 6a shows the odds ratio of 1000 passengers increase on a route in the model for routes addition. Figure 6b shows the odds ratio of 10000 (i.e., 10 units) passengers decrease on a route in the model for routes deletion. Both figures imply that the effect of the demand variable on adding or deleting a route becomes more influential from 2002 to 2009. For example, in 2002, 10000 passengers increase on a route make the probability of being added increase 4 times as compared to not be added. However, in 2008, the ratio of adding to not adding become 6.18 if there is 10000 passengers rise on a route. Surprisingly, the odds ratio in 2007-2008 reaches 16.7 implies the big influence of demand on the decisions of adding routes in that year. The similar trend is observed in the route deletion. However, the ratio of deleting a route when the demand decreases rises slightly from 1.09 in 2002 to 1.34 in 2008 with an exception in the year from 2006 to 2007. For both route addition and deletion, it is observed that in the year 2007, the change of demand has greater impacts on routes planning than any other years. This behavior reflects the highly sensitive nature of the airline industry to economic changes, and coming in the year 2007, preceded the economic downturn in the following year. Further, over the period of years that we have considered in our study, the average load factors across the airline networks has consistently increased. This explains the above observed trend which a reflection of the airlines' strategy of restructuring their network to increase the number of passengers carried per flight.

VI. Closing Remarks and Future Work

In this paper, we have performed discrete choice analysis to identify a model of airlines' decisions on routes selection based on random-utility theory. The objective is to provide stakeholders, who may not have access to the airlines' decision-making processes, an approach to estimate airlines' decisions to make better response. In closing, we can make the following remarks based on our studies conducted so far:



a) **Odds ratio of demand for routes addition** b) **Odds ratio of demand for routes deletion**

Figure 6. The odds ratio of estimated parameters in the model of routes addition of 7 evolution instances from 2002 to 2009

1. In both models of route addition and deletion, the interaction effect between continuous variables and hub level 2 are not significant in most of the evolution instances. This may be due to the small number of observation of choices in the hub level 2. For example, in the evolution instance of 2005 – 2006, there are only 3 observations for addition and 2 observations for deletion. With such small number of observations, the model could not get significant statistical results.

2. Market demand has been always a significant affecting the decisions on route planning for both adding or deleting routes. Since increasing the number of passengers carried, by increasing the load factors on flights, is a proven approach to increasing revenues, this conclusion was to be expected.

3. The R^2 of the model for routes addition is 0.23, which indicates that this model can be improved further. Adding more explanatory variables could be one way to improve the method.

We noted that some studies have attempted to model airlines' decisions using a machine learning approach. In our further work, we will compare the machine learning approach with our approach to see the relative strengths and weaknesses. We will seek to identify if these two approaches can be made to work together so that they may complement each other. On the other hand, we have identified some more explanatory variables that may help model airlines' decisions, such as the season information. We will work to include these variables in our model including effects of airline's own operating model. Our planned future work also includes using the game theoretic models to model the interactive decision-making between the airlines and other stakeholders in the aviation industry. Such interaction can be used to study strategic decision-making among stakeholders at different levels of hierarchy in the industry. We will also attempt to further validate our model by comparing with other existing models of airlines' decision-making. As more data becomes available, we will use our model to predict the air transportation network topology in the future.

Acknowledgements

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

References

[1] Belobaba, P., O. A. and Barnhart, C., *The Global Airline Industry*, John Wiley & Sons, 2009.

[2] Phillip J. Lederer, R. S. N., "Airline Network Design," *Operations Research*, Vol. 46, No. 6, 1998, pp. 785–804, http://dx.doi.org/10.1287/opre.46.6.785.

[3] Jaillet, P., S. G. and Yu, G., "Airline Network Design and Hub Location Problems," *Location Science*, Vol. 4, No. 3, 1996, pp. 195–222.

[4] Magnanti, T. L. and Wong, R. T., "Network Design and Transportation Planning: Models and Algorithms," *Transportation Science*, Vol. 18, No. 1, 1984, pp. 1–55.

[5] Song, Kisun, L. J.-H. M. D., "A Multi-Tier Evolution Model of Air Transportation Networks," *AIAA Aviation 2014*, AIAA, Atlanta, GA, 2014.

[6] Kotegawa, T., *Analyzing the Evolutionary Mechanisms of the Air Transportation System-of-Systems using Network Theory and Machine Learning Algorithms*, Ph.D. thesis, Purdue University, Indiana, Jan. 2012.

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

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

[9] Bureau of Transportation Statistics, "Air Carrier Summary : T2: U.S. Air Carrier TRAFFIC And Capacity Statistics by Aircraft Type," online resource, Oct. 2014, http://www.transtats.bts.gov/Fields.asp?Table_ID=254.

[10] Bureau of Transportation Statistics, "Air Carrier Financial : Schedule P-5.2," online resource, Oct. 2014, http://www.transtats.bts.gov/Fields.asp?Table_ID=297.

[11] Federal Aviation Administration, "Passenger Boarding (Enplanement) and All-Cargo Data for U.S. Airports," online resource, Oct. 2014, http://www.faa.gov/

[12] United Airlines, "Corporate Fact Sheet," online resource, Nov. 2014, http://newsroom.unitedcontinentalholdings.com/corporate-fact-sheet.

[13] Delta Airlines, "Delta Air Lines Newsroom Global Network," online resource, Nov. 2014, http://news.delta.com/index.php?s=20309&cat=3191.

[14] American Airlines, "American Airlines Group," online resource, Nov. 2014, http://www.americanairlines.com/

[15] Census Bureau, *Statistical Abstract of the United States, 2012*, Books Express Pub., 2011.

[16] Neyman, J. and Pearson, E. S., "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, Vol. 231, 1933, pp. 289–337.

[17] Colin Cameron, A. and Windmeijer, F. A. G., "An R-squared measure of goodness of fit for some common nonlinear regression models," *Journal of Econometrics*, Vol. 77, No. 2, April 1997, pp. 329–342.

[18] Peng, C.-Y. J., Lee, K. L., and Ingersoll, G. M., "An Introduction to Logistic Regression Analysis and Reporting," *The Journal of Educational Research*, Vol. 96, No. 1, 2002, pp. 3–14.