Journal of Air Transportation
Vol. 24, No. 3, July 2016

Modeling Airlines' Decisions on City-Pair Route Selection Using Discrete Choice Models

Zhenghui Sha[*]
Northwestern University, Evanston, Illinois 60208
and
Kushal Moolchandani,[†] Jitesh H. Panchal,[‡] and Daniel A. DeLaurentis[§]
Purdue University, West Lafayette, Indiana 47907

DOI: 10.2514/1.D0015

An approach based on the discrete choice random-utility theory is presented to model airlines' decisions of strategically adding or deleting city-pair routes. The approach consists of methods for identification of air transportation networks, determination of choice sets, and comparison and validation of developed discrete choice models. The developed approach enables the quantification and estimation of airlines' decision-making preferences to the identified explanatory variables, including market demand, direct operating costs, distance, and whether terminal airports are hubs or not. It is observed that market demand more significantly affects the decisions on route deletion than their addition. Furthermore, the effect of direct operating costs is significant in the decision of route deletion, whereas it is not in route addition. Finally, airlines' decisions vary, depending on the airport hub status. These trends are observed consistently over time in the current analysis of historical data from 2004 to 2013. The developed models show better prediction as compared to other models in literature. With the developed models, an air transportation network generator is constructed that, in turn, is used for model validation. This approach benefits those who want to understand airlines' decision-making behaviors and those who need to understand the past and future evolution of an air transportation network.

Downloaded by PURDUE UNIVERSITY on September 5, 2017 | http://arc.aiaa.org | DOI: 10.2514/1.D0015

Nomenclature

C_{add}	=	choice set for route addition
C_N^2	=	all possible combinations of size two in N elements
$D(t)$	=	airlines' true decision model at time t
$D'(t)$	=	approximation of airlines' true decision model at time t
E_n	=	number of routes in the air network at time step n
n_{as}	=	number of routes that are actually selected
n_{ans}	=	number of routes that are actually not selected
n_{ps}	=	number of routes that are predicted to select
n_{pns}	=	number of routes that are predicted to not select
O	=	size of union set
P	=	probability
R^2	=	coefficient of determination
U	=	utility
V	=	observed utility
$\boldsymbol{x}$	=	airlines' decision criteria
$\boldsymbol{x}'$	=	airlines' potential decision criteria
$\boldsymbol{y}$	=	decision outcomes
$\boldsymbol{y}_{\text{hist}}$	=	airlines' historical decisions
$\boldsymbol{y}'$	=	prediction of decisions
α	=	level of significance
$\boldsymbol{\beta}$	=	vector of parameters quantifying airlines' decision preferences
$\boldsymbol{\epsilon}$	=	uncertainty

I. Introduction

ROUTE selection, fleet planning, and schedule development are the three important strategic decisions airlines make [1]. Route selection decisions are concerned with where to offer origin–destination service subject to fleet availability constraints, with the intent of maximizing an airline's profitability. Route selection is important because it directly impacts networkwide effects such as the propagation of delays, the robustness of the network to service disruptions, and the network's traffic flow capacity. Therefore, organizations, such as the Federal Aviation Administration (FAA), NASA, the Bureau of Transportation Statistics (BTS), as well as many others are interested in modeling airline decisions of route selection, with the intent of answering key questions such as how investments at individual airports will translate to network restructuring, leading to a reduction of networkwide flight schedule delays [1]. In this paper, a route is equivalent to a segment connecting a city pair. Route selection means the decision of an airline on whether to add or delete a city-pair route, i.e., airlines' planning of origin–destination route service between two cities.

Airlines' decision-making strategies are not publicly known and likely depend on many criteria. Consequently, researchers make several simplifying assumptions while modeling airline decision-making strategies. Many studies replicate airline decisions under the assumption that economic considerations, especially profit maximization, are the primary drivers of decision making for most airlines. However, in this paper, the strategic planning process of specifying origin–destination routes is modeled as a discrete choice problem where the airlines are collectively modeled as a single monopolistic benevolent entity. This is a reasonable assumption, given the nature of our model that is aimed at assisting decision makers, such as the FAA. The goal is to approximate airlines' decisions based on publicly available data, such as those from the BTS [2] and the FAA [3].

The two schematics shown in Fig. 1 help clarify our objective further. Figure 1 (left) shows the airlines' true decision-making model $D(t)$ as a function of time t. This model takes into account the airlines' decision criteria $\boldsymbol{x}$, as well as external factors that describe the uncertainty associated with the decision-making process, such as the projected economic conditions, and gives airline decisions as the output $\boldsymbol{y}$. Figure 1 (right) shows our approximation of the airlines' model as $D'(t)$. For this model, the researchers determine the

Presented as Paper 2015-2438 at the 15th AIAA Aviation Technology, Integration, and Operations Conference, Dallas, TX, 22–26 June 2015; received 23 November 2015; revision received 7 April 2016; accepted for publication 25 July 2016; published online 21 September 2016. Copyright © 2016 by Daniel A. DeLaurentis. Published by the American Institute of Aeronautics and Astronautics, Inc., with permission. Copies of this paper may be made for personal and internal use, on condition that the copier pay the per-copy fee to the Copyright Clearance Center (CCC). All requests for copying and permission to reprint should be submitted to CCC at www.copyright.com; employ the ISSN 2380-9450 (online) to initiate your request.

[*] Postdoctoral Fellow, Department of Mechanical Engineering.

[†] Research Assistant, School of Aeronautics and Astronautics. Member AIAA.

[‡] Associate Professor, School of Mechanical Engineering.

[§] Professor, School of Aeronautics and Astronautics. Associate Fellow AIAA (Corresponding Author).

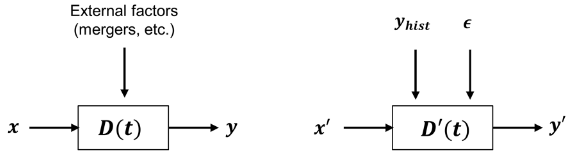


Fig. 1 Schematic of an actual airline decision model (left) vs an approximate decision model (right).

potential criteria $\boldsymbol{x}'$ that would influence airlines' decisions. Examples of these criteria include market demand, direct operating costs (DOCs), etc. Because researchers do not know the actual set of criteria used by the airlines, the modeled criteria may be incomplete. Therefore, the additional unknown criteria and external factors can be together modeled as uncertainty $\boldsymbol{\epsilon}$. This information along with airlines' historical decisions $\boldsymbol{y}_{\mathrm{hist}}$ can be used to statistically estimate the airlines' preference structures of the decision factors $\boldsymbol{x}'$. Those estimated preferences facilitate the prediction of route selection decisions $\boldsymbol{y}'$ through the model $D'(t)$. One of the benefits of approximate decision models is that they enable the creation of local mechanisms (e.g., route addition/deletion), which in turn help in forecasting the evolution of the air transportation network (ATN).

So, our research objective is on developing an approximate model [i.e., the $D'(t)$ model shown in Fig. 1] with acceptable performance at an aggregate level to understand the major decision-making factors of airlines and forecast the route selection and the evolution of the ATN. The objective of this paper is not to develop a model that best captures the decision-making interaction and dynamic from the market perspective, e.g., by using game theory to model the competition relationship of each airline. In this paper, the ATN is defined as a set of routes of passenger aircraft operations in the U.S. domestic market. Note that this single set is the union of the sets of several airlines that operate in the U.S. commercial airspace. In an ATN, nodes are airports (cities) and edges are the routes between two airports (a city pair). The structure (how nodes are connected with each other) of the ATN is also referred to as the topology. Therefore, the route selection in the context of the ATN refers to the addition and/or deletion of edges between two nodes. In general, the ATN will include all possible airports of the United States. However, in this study, we only consider the major airports. The network formed by these major airports is the ATN (or our network) being studied in this paper. The reason why only major airports are considered and our approach for selecting major airports is elaborated on in Sec. IV.A.

With the motivation of understanding the effects of airlines' decisions on the strategic planning process of origin–destination routes and the effects of such decisions on ATN evolution, we answer two primary research questions:

1) How can airlines' preference structure of route selection be estimated?

2) How can the estimated airlines' preference structure be used to forecast the evolution of the ATN?

In the following section, a review of literature on airline route selection is presented. In Sec. III, an introduction to discrete choice model (DCM) and our approach for estimating airlines' decision-making preferences for route selection are presented. The approach for data collection and processing is presented in Sec. IV. In Sec. V, the discrete choice analysis is performed. To verify and evaluate the models, in Sec. VI, a comparative study is performed. To validate the models, in Sec. VII, a framework is developed for regenerating the ATN in the past and forecasting it in the future. The resulting synthetic networks are compared with real networks for validation. Finally, conclusions are presented in Sec. VIII.

II. Literature Review

Approaches for airline origin–destination route selection can be classified into optimization based and network theory based. The optimization-based approaches formulate route selection as a mathematical programming problem with an objective function such as maximization of profit or traffic flow. For example, Jaillet et al. [4] presented three integer programming (IP) problems for airline network design, and they introduced heuristic schemes based on mathematical programming for designing capacitated networks and routing problems. A demand matrix for origin–destination cities was given, and the design resulted in a hub and spoke network. Lohatepanont and Barnhart [5] used a linear programming (LP) relaxation method to construct an integrated model for route selection and fleet assignment. They solved this model using the column and row generation approach with the objective of maximizing profit. Their inputs included a base list of routes, recapture rates, demand data, and fleet composition size, whereas the output was a recommended list of routes after addition and deletion, along with fleet assignment. Balakrishnan et al. [6] formulated the route selection problem as a mixed-integer program to select a set of profitable routes from a main base to one or more terminal bases. As inputs, they used traffic estimates for each origin–destination city pair, airline operating costs, and aircraft characteristics. Cordeau et al. [7] solved a simultaneous aircraft routing and crew scheduling problem with an objective function of minimizing costs. They modeled this using LP relaxation and solved it using Benders decomposition and column generation. Lederer and Nambimadom [8] studied the choices of different network designs and schedules using profit maximization as the objective function. Finally, Magnanti and Wong [9] reviewed some of the IP-based approaches to network design and described both discrete and continuous choice models and algorithms. We observe that, in models such as the ones discussed here, both demand and cost are the common variables.

These optimization-based approaches have their shortfalls. They only formulate the airlines' decision making into one single-objective problem, such as cost minimization, and thereby do not address the multiobjective decision-making nature of the airlines. Second, IP models usually have a high computational cost, and the complexity of these models increases rapidly as the number of nodes increases. Therefore, they are not applicable to large scale network design.

The second category of approaches is based on network theory. For example, Kotegawa et al. [10] developed a modified network growth model based on the Barabási–Albert (BA) model. They adopted network metrics such as the clustering coefficient (CC) and degree as the variables for establishing the model of link addition, whereas for the link removal mechanism, they included variables such as the DOCs per passenger, the number of flights per route, and the type of route (hub–hub, hub–spoke, spoke–spoke, and hub–spoke–hub). They also included a hub searching algorithm as an intermediate step in the link removal mechanism. Song et al. [11] presented a multitier model to estimate the network evolution. Such a model split the ATN into two tiers: primary network, and secondary network. Their model used the available demand data as well as the future demand estimation to predict the network evolution.

Kotegawa [12] used machine-learning algorithms to study network evolution using patterns derived from historical data. He compared algorithms including logistic regression (LR), random forests (RF), and support vector machines (SVMs) by calculating the prediction accuracy. In another study, Kotegawa et al. [13] compared three models, viz., LR, artificial neural network (ANN), and a fitness function (FF) model. All these models were trained using the historical data for the U.S. domestic network from 1990 to 2005, with a one-year step size. Network attributes, such as node degree, node weights, link weights, CC, eigenvector centrality, and nodal strength were used as the variables in the route selection models. Among these approaches, the ANN had a high computational cost. Thus, the ANN-based approach was only used to analyze a few subnetworks instead of the entire ATN. Furthermore, this study focused only on the route addition to the network due to reconfiguration. In another work, Kotegawa et al. [14] used SVMs and the RF method for route selection. The algorithm for link addition was based on type 1 and type 2 errors, route filtering was based on SVMs, and link deletion had a similar algorithm as addition but did not require any initial filtering. The model was trained using the U.S. domestic segment data from 1990 to 2008 with a network size of 304 nodes and a one-year step size. Common variables in network theory-based models included the degrees of nodes, the CC, eigenvector centrality, and the demand between two nodes.

Downloaded by PURDUE UNIVERSITY on September 5, 2017 | http://arc.aiaa.org | DOI: 10.2514/1.D0015

Downloaded by PURDUE UNIVERSITY on September 5, 2017 | http://arc.aiaa.org | DOI: 10.2514/1.D0015

Unfortunately, these machine-learning approaches resulted in low prediction accuracy (most of the approaches had accuracy less than 25%; see Sec. VI for details) and a lack of theoretical foundation for explaining airlines' decisions. Those models were driven by the purpose of replicating the phenomena and predicting the results but neglected the true mechanism underlining network restructuring, viz., the decision making of airlines. Also, these approaches adopted network metrics as the primary means of analysis, ignoring many airline-specific criteria in the process.

In summary, airline decisions are based on several proprietary criteria that are not completely known by other organizations who are interested in modeling these decisions. The major limitation of the existing models is that they do not capture the airlines' decision-making process well in route selection based on known decision factors. Thus, we have identified a need for developing a model of airlines' decision making that is responsive to changing market and economic conditions, availability of new technologies, etc., yet feasible for other interested organizations to use, such as the FAA, NASA, airlines, and university researchers. This identified need forms the motivation for our work. In other words, a decision-centric framework of modeling airline behaviors, hitherto lacking in literature, is the focus of this paper. The advantages of our 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. A tabular summarization of the literature survey is presented in the supplementary material [15].

III. Technical Approach

A. Overview of Discrete Choice Models

Our work is based on the central hypothesis that airlines' decisions of route selection can be mathematically modeled using discrete choice analysis (DCA) based on random-utility theory. DCA has been used in estimating individual entities' decision-making preferences in many complex networked systems such as the Internet [16,17] and road transportation systems [18]. In DCA, it is assumed that the decision maker has complete knowledge about his/her own utility U_i when choosing an alternative i. This utility consists of two parts: the observed utility V_i, and the unobserved utility ϵ_i, which describes the uncertainty in determining the true utility U; see Eq. (1):

$$U_i = V_i + \epsilon_i \tag{1}$$

The observed utility is usually modeled as a parameterized function of a set of explanatory variables that may affect the decision maker's decisions. Such variables are deterministic in nature from the researcher's point of view, and they can be identified through different ways, such as from surveys, subject matter experts, and the literature. The unobserved utility captures the randomness due to unobserved attributes, unobserved variations among decision makers, measurement errors, functional misspecification, and the bounded rationality of decision makers [19]. The approach uses a common formulation [19] that constructs the observed utility V_i in a linear form:

$$V_i = \boldsymbol{x}^T\boldsymbol{\beta}_i = \beta_{i1}x_{i1} + \beta_{i2}x_{i2} + \ldots + \beta_{ik}x_{ik} \tag{2}$$

where $\boldsymbol{x} = (x_i, x_2, \ldots, x_n)^T$ is a vector that contains n explanatory variables for the utility, and $\boldsymbol{\beta}_i = (\beta_{i1}, \beta_{i2}, \ldots, \beta_{ki})^T$ is the set of weights that quantifies decision maker's preferences. With the assumption of 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 a 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 $\epsilon_j - \epsilon_i$, and thus can be determined once the density function of ϵ is specified. Given different types of distribution $f(\epsilon_i)$, different DCMs can be obtained. For example, if we assume the distribution follows Gaussian, the resulting DCM is called a probit model [19]. In this paper, a multinomial logit model [19] is adopted. It is based on the assumption that ϵ_i is independent and identically distributed following Gumbel distribution [18]. There are two reasons for this choice. First, a closed form of the resulting choice probability can be obtained, which facilitates the computation. Second, the logit model is known to have good performance on predicting the decision-making preferences in different fields [16,17,19–21]. As a result, an analytic solution of Eq. (3) in the logit form is obtained:

$$P_i = \frac{e^{\boldsymbol{x}^T\boldsymbol{\beta}_i}}{\sum_{j=1}^{J} e^{\boldsymbol{x}^T\boldsymbol{\beta}_j}} \tag{4}$$

In the context of airline route selection, whether to establish or remove a route between a city pair can be modeled as a binary choice ($1 =$ yes or $0 =$ no). The probability that an airline chooses to add or delete a route is

$$P_1 = \frac{e^{\boldsymbol{x}^T\boldsymbol{\beta}_1}}{e^{\boldsymbol{x}^T\boldsymbol{\beta}_1} + e^{\boldsymbol{x}^T\boldsymbol{\beta}_0}} = \frac{e^{\boldsymbol{x}^T(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)}}{1 + e^{\boldsymbol{x}^T(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)}} = \frac{e^{\boldsymbol{x}^T\boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}^T\boldsymbol{\beta}}} \tag{5}$$

where $\boldsymbol{\beta} = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_0$ captures the difference between preferences for choosing yes or no. The parameters $\boldsymbol{\beta}$ can be readily estimated from the choice data through statistical estimation techniques. In this paper, the maximum likelihood estimation is adopted. The key is then to construct the preference structure in the form of the utility function V [in Eq. (2)] through the identification of the explanatory variables $\boldsymbol{x}^T$.

Referring back to the approximate decision model schematically depicted in Fig. 1, the identified decision factors $\boldsymbol{x}'$ correspond to the explanatory variables of the utility model, whereas $\boldsymbol{\epsilon}$ corresponds to the unobserved component of the airlines' utility. The data that reflects the airlines' actual decisions on the explanatory variables are $\boldsymbol{y}_{\text{hist}}$. Together, these inputs contribute to the decision model $D'(t)$, which in turn is used to determine the airlines' decisions $\boldsymbol{y}'$.

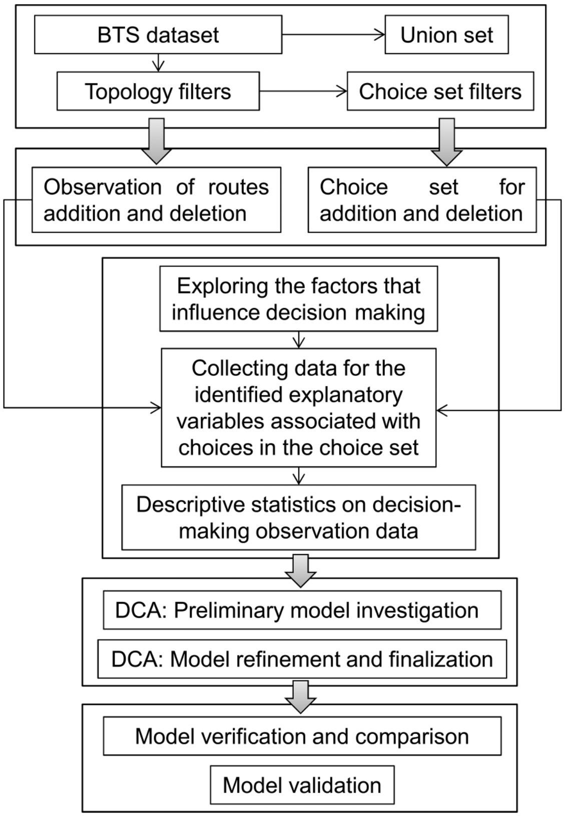

BTS dataset
Union set
Topology filters
Choice set filters
Observation of routes addition and deletion
Choice set for addition and deletion
Exploring the factors that influence decision making
Collecting data for the identified explanatory variables associated with choices in the choice set
Descriptive statistics on decision-making observation data
DCA: Preliminary model investigation
DCA: Model refinement and finalization
Model verification and comparison
Model validation

Fig. 2 Proposed approach to estimate airlines' preferences on route selection.

Downloaded by PURDUE UNIVERSITY on September 5, 2017 | http://arc.aiaa.org | DOI: 10.2514/1.D0015

B. Approach

In this section, we introduce our DCM-based approach of modeling the airline route selection, as illustrated in Fig. 2. We first determine the set of airports that form the ATN to be studied. We then use a set of unique filters based on heuristics for identifying the set of routes that connect these airports. Next, we determine the "choice set" that includes all the routes that the airline considers during decision making. In other words, during each year, the airline selects routes for adding or deleting from this choice set. By analyzing the available data over two consecutive years, we can observe the airlines' historical choices of route addition and deletion. In Secs. IV. A and IV.B, these filters are elaborated in detail.

Once the choice set and the airlines' historical choices are determined, the possible factors that affect airlines' decisions are explored and the corresponding data are collected. As we have highlighted in the literature review (Sec. II), the demand and cost are commonly used as explanatory variables in many airline route selection problem formulations [4–6]. These two, along with the distance between cities and the hub/nonhub nature of airports, form the set of explanatory variables in the decision model we present in this paper. The hub/nonhub nature of airports has been used in route selection problems: for example, in [10], and Belobaba et al. presented a discussion of operational advantages of hubs along with a discussion of route selection problems [1]. We have also tested to see whether distance would be a suitable explanatory variable for our model; on finding a positive answer, we chose to retain it. In addition, we have intentionally adopted a modeling approach that stays simple and tests for a predictive owner so as to avoid the inefficiency of a "put it all in" approach, which suffers from complexity and computational challenges, etc.

The sources of these data and the corresponding data analysis are presented in Sec. IV.C. With the choice set, the observation of historical choices, and the associated data of explanatory variables, the dataset required to perform DCA is completed. Using the mathematical formulation of the utility function in Eq. (2), the final decision models for route selection are established and the parameters $\boldsymbol{\beta}$ that quantify the airlines' preference are estimated.

To verify the obtained decision models and to evaluate their performance on predictability, we compare our models with existing models. The comparative study is performed at the route level. The core idea is to use the estimated preference structure to calculate the predicted probability of route addition or deletion. The prediction accuracy can thus be obtained by calculating the percentage of routes that has been actually added or deleted. We compare different models using these accuracies. Details of the developed approach for comparison and the results are shown in Sec. VI.

The validation of models is realized from two aspects: network regeneration and forecast. Validation is performed at the network level. The core idea is to use the obtained decision models as the linking mechanisms to construct a network topology generator. With this generator, a synthetic network either in the past (regeneration) or in the future (forecast) can be obtained. Details of the developed approach for validation and the corresponding results are presented in Sec. VII. The rest of the paper is organized in the order of execution of the approach shown in Fig. 2.

IV. Data Collection and Processing

A. Air Transportation Network

To determine the topology of the ATN, we use the data of operations of the large certified carriers, including data on the number passengers they carried, the number of flights performed, the operating costs, and the airports served. We adopt BTS data from forms T-100 [2] and T-2 [22] that provide data on airline operations, as well as schedule P-5.2 [23] that provides the airlines' economics data. The classification of "large certified carriers" includes all 41 airlines that are included in schedule P-5.2 [23].

The FAA provides a list of the largest commercial service airports in the United States based on the number of passenger boardings [3]. From this list, we select all airports that have "primary commercial operations" and are categorized as either large, medium, or small hubs. Together, this makes a set of 134 airports that handle over 96% of all enplanements in the United States in the calendar year 2013. Of these, we removed the Phoenix-Mesa Gateway Airport and Saipan International Airport from consideration, as there were no segment data (see [2] for details), leaving a final set of 132 airports that formed the nodes of the network in our study.

With the 132 selected airports, we determined the network topology from 2004 to 2013 based on the BTS T-100 segment dataset [2]. We included a route if it passed the following topology filters:

1) It had at least eight scheduled departures during any two consecutive months in a given year.

2) It had nonzero passenger demand.

3) It was classified in the "domestic" category in the BTS table.

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

As an example, when we apply these filters to the 2009 network data, we get a final set of 2015 routes (edges) out of the total 8646 routes required for the network to be fully connected.

B. Determine the Choice Set and Observations of Choices

We defined the choice set for route selection as the set of candidate routes that were available for the airline to add or delete, respectively. If a route was in operation for any number of years in the range of years previously under consideration but not present in the current year, then it formed part of the choice set for addition for the next year. We reasoned that if any of the airlines found a route feasible for operation in the past, then any airline was likely to consider operating on it in the future, provided sufficient demand existed. The choice set for route deletion was the set of routes being used in the current year. For example, the years analyzed in this paper are from 2005 to 2013. Then, all the routes present in these nine years formed our union set O of choices for addition and deletion. Using the network evolution from 2005 to 2006 as an instance, the choice set for addition C_{add} was to subtract the edge set E_5 of 2005 from the union set (i.e., $C_{\mathrm{add}} = O - E_5$), and the choice set for deletion was E_5. Because some values of segment demand were zero and some DOC values were not available, these routes were filtered out from our choice set.

To analyze airlines' decision-making preferences on route selection, we needed observations of which routes were actually added or deleted in an evolution, i.e., the choices made by airlines. To obtain the observations, we performed the edge dynamic analysis on network data from two consecutive years. For instance, from 2005 to 2006, with the filtered networks, the choice set for adding had 809 routes in which 170 routes were selected. The choice set for deleting had 2062 routes (i.e., the number of routes in 2005), and there were 138 routes selected to delete. The size of the choice set and number of observations of all nine evolution instances, from 2004 to 2013, can be referred to in the supplementary material [15]. According to the rule of thumb for a minimum sample size for regression analysis [24], our dataset was sufficient for fitting the models. In the following subsection, we discuss how the data regarding the demand, cost, and airport hub information were obtained from various data sources.

C. Explanatory Variables

1. Demand

We use the BTS T-100 dataset to obtain two kinds of demand: market demand, in which a passenger is counted only once as long as he/she remains on the same flight number; and segment demand, in which a passenger is counted for each leg of the trip. This is why segment demand is frequently higher than market demand. Using this dataset, we sum the demand reported by all airlines and make it symmetric by calculating the average of the bidirectional value between every airport pair. This assumption of symmetry is reasonable, especially over longer periods of time (such as one year), even though a given passenger may not fly a return trip on the same day. Finally, in cases where the BTS reports zero market demand, we use the corresponding value of segment demand as an estimate.

Downloaded by PURDUE UNIVERSITY on September 5, 2017 | http://arc.aiaa.org | DOI: 10.2514/1.D0015

Table 1 Results of estimated decision-making preference for route addition and deletion from 2005 to 2006

Variables	Parameter	Model for route addition			Model for route deletion		
		Mean	p value	Odds ratio	Mean	p value	Odds ratio
Intercept	β_0	−1.91	<0.001	0.15	1.49	<0.001	4.43
Hub level 1	β_{11}	0.96	0.018	2.62	−0.042	0.96	0.96
Hub level 2	β_{12}	−1.78	0.62	0.17	−3.80	0.0094	0.022
Market demand	β_2	0.087	<0.001	1.09	−0.23	<0.001	0.80
Unit cost	β_3	0.74	0.26	2.10	−1.97	0.026	0.14
Distance	β_4	−0.044	0.83	0.96	−0.17	0.59	0.84
Demand at hub level 1	β_{51}	0.14	<0.001	1.15	−1.39	<0.001	0.25
Demand at hub level 2	β_{52}	0.47	0.43	1.60	0.21	0.30	1.24
Cost at hub level 1	β_{61}	−0.65	0.34	0.52	0.49	0.75	1.63
Cost at hub level 2	β_{62}	11.50	0.24	98,964.85	2.75	0.008	15.59
Distance at hub level 1	β_{71}	−0.83	0.016	0.43	1.01	0.13	2.74
Distance at hub level 2	β_{72}	−0.16	0.94	0.85	−0.25	0.79	1.28
				Overall model fit			
Log likelihood at zero		−358.62				−435.94	
Log likelihood at convergence		−278.82				−76.68	
McFadden's adjusted R^2		0.19				0.80	

2. Cost

The DOCs serve as the second variable in our decision model. We calculate the DOCs over a route as the sum of the DOCs of all airlines that operated on that route, weighted by their number of operations. If the route is not active in a given year, then the cost data are also unavailable. In such cases, we estimate the DOCs by scaling the latest year's data using an "airline cost index" from the *Statistical Abstract of the United States* [25].

3. Hub and Nonhub

As mentioned earlier, we obtained the set of airports in our network from the FAA dataset. From this, we compared the list of 30 large hubs with the hubs listed by the three largest full-service airlines in the United States, viz., United Airlines [26], American Airlines [27], and Delta Airlines [28]; and we selected 21 airports that were fixed as hubs in this work. Finally, we added Chicago Midway to this list because of its large volume of operations [2], resulting in a set of 22 hub airports. Accordingly, we divided the routes into three categories based on whether their terminal airports were hubs or nonhubs. When both terminal airports were hubs, the route was identified by the label "2"; when only one airport was a hub, the route was labeled "1"; and when both airports were nonhubs, it was labeled "0." For example, in the network of year 2005 that had 2061 routes, there were 644 routes at hub level 0, 1198 routes at hub level 1, and 220 routes at hub level 2.

4. Distance

Although not explicitly a part of our preference structure, distance is an important factor in airline decisions on fleet allocation. Since the airlines take their aircraft inventory into account in their decision making, this factor indirectly affects their route selection decisions. The data on distances ares obtained from form T-100 [2].

V. Estimating Airlines' Preferences of Route Selection

A. Preference Structure Formulation

In this section, we formulate the preference structure of airlines' decisions on route selection. The four explanatory variables are the hub indicator x_1, where both airports are nonhubs (0), one airport is a hub (1), or both airports are hubs (2); 2) potential market demand x_2 (unit: 1000 passengers); 3) potential unit cost x_3 (unit: cent/nautical mile/seat); and 4) distance x_4 (unit: 1000 n miles).

We assume that an airline makes use of the utility maximization approach for its decision on route selection, and no factors outside of the model are used by the airline for its choice. Another important assumption is that route selection is driven by demand pull but not by supply push. With Eq. (2), the preference structure (i.e., the utility models) is constructed in Eq. (6). Consequently, the decision model is obtained from Eq. (5) which describes the binary choice of airlines on route selection:

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

Because of the existence of both categorical (e.g., the level of the hub) and continuous variables (e.g., the demand) in the utility function, we investigate the interaction effect of explanatory variables by including the cross-product terms in the model, as shown in Eq. (6). The purpose is to assess whether the effect of continuous variables, such as demand, on the airlines' decisions would be different according to the hub status of the routes.

B. Discrete Choice Analysis and Preferences Estimation

With the observation data of route selection, the DCA is performed to analyze whether the identified explanatory variables significantly affect airlines' decisions and which variables are most influential. In this section, we use the network evolution from 2005 to 2006 as an example to interpret the obtained results. A similar analysis is performed for other evolution instances from 2004 to 2013. The results of each evolution reflect the dynamics of decisions over time, which is discussed in Sec. V.C.

Table 1 shows the results of estimation of the decision-making preferences of airlines on route addition and deletion from 2005 to 2006.[¶] (Since hub level x_1 is a categorical variable that has three levels, when performing the regression, we must specify a reference level with the parameter set to zero; the parameters of the other two levels are then estimated by comparing with the reference level. Therefore, the parameters β_1, β_5, β_6, and β_7 are two-dimensional vectors in which the elements represent the parameters for each of the other two levels. In the analysis, the routes at hub level 0 are set as the reference level. For example, the regression will provide $\beta_1 = (\beta_{11}, \beta_{12})^T$, where β_{11} is the parameter for hub indicator variable, x_1 equals to 1, i.e., $x_1 = 1$ and β_{12} is the parameter for $x_1 = 2$.) As to the overall model fit, the log likelihood at convergence is −278.82 for the model of route addition and −76.68 for the model of route deletion; these values are both higher than the corresponding values of the null model. This is validated with the likelihood ratio test [29] in which the p value is less than 0.0001, indicating that the improvement is statistically significant. The overall model fit is quantified using the MacFadden's adjusted R^2 [31], which shows that the model for route deletion ($R^2 = 0.80$) performs better than that for route addition ($R^2 = 0.19$).

[¶] The correlation analysis is performed before the regression analysis. There is no significant collinearity between each two of the continuous explanatory variables.

Downloaded by PURDUE UNIVERSITY on September 5, 2017 | http://arc.aiaa.org | DOI: 10.2514/1.D0015

In the model of route addition, the p-value of the parameter of hub level 1 β_{11} is 0.018, which indicates that the hub status of a route has a nontrivial impact on the decision of route addition. When not specifying, the level of significance is $\alpha = 0.05$. The estimated mean of β_{11} is 0.96. 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 [30]. The odds ratio of hub level 1 is 2.62. This means the probability of adding a route between two airports that have at least one hub is 2.62 times greater than the probability of adding a route connecting nonhub airports. This reflects the hub-to-spoke expansion in the ATN in the evolution of 2005–2006.

The effect of demand on route addition is also statistically significant. The odds ratio of demand is 1.09, which means that, with an increase in 1000 passengers between city pairs, the probability of adding a route is 1.09 times greater than not adding that route. However, the effect of demand at different hub levels varies. In terms of the odds ratio, the effect of a 1000-passenger increase on routes at hub level 1 on adding a route is 1.27 times greater than the effect of demand on routes at hub level 0.

Both cost and distance do not significantly affect route addition. Furthermore, there is no significant interaction effect between cost and hub levels. However, the negative interaction effect exists between distance and hub level 1, indicating the increase of the routes' distance at hub level 1 results in the decrease of the probability of those routes being added. This reflects that the decision on adding hub-spoke type of routes is sensitive to distance.

In the model of route deletion, as shown in Table 1, demand plays as an important role. This is reflected in the corresponding p-values of β_2 and β_{51}. Similarly, whether routes connect hubs or not is a significant factor. However, the results show that the difference between hub levels 0 and 1 is not significant, whereas a difference exists between hub levels 0 and 2. In other words, the airlines are more likely to drop routes at hub level 0 than hub level 2 for the same amount of demand drop. The reason for this can be that the airline can fill seats on its routes at hub level 2 by diverting passengers such that they fly through these hubs, whereas the same is not feasible for routes between non-hub cities. However, in the model of route addition, we do not observe such a phenomenon. This is probably because the routes that connect both hub airports are already saturated. Thus, not too many routes between hub cities were added as compared with the routes added between non-hub cities. This is reasonable because the change of routes among hubs will not be so frequent because airlines want to maintain their national reach and obvious cost consideration.

In the analysis of evolution from 2005 to 2006, both the DOCs and distance are not as dominant as the demand on decisions of route selection, as reflected by the corresponding p values. For example, the DOC is significant in the model of route deletion but not in the case of route addition. A possible explanation could be as follows. The airline knows the economics of a route that it is currently operating and, given the present state of demand, if the route does not generate profit, then it is likely to be deleted from the network. Thus, in case of deletion, the DOC plays a significant role. In case of addition, however, an airline gives more importance to the potential demand as much as it does to the profits; hence, in this case, the DOC plays a relatively minor role. In the next subsection, these explanatory variables are analyzed for different network evolution instances to see whether their effects are consistently significant and whether the airline's preferences on specific variables are consistent.

C. Decisions over Time

To analyze decisions over time, the DCA with the model in Eq. (6) is performed on nine evolution instances of the ATN from 2004 to 2013. The purpose of this study is threefold: 1) to evaluate the model consistency, 2) to investigate how airlines' preferences on route selection change over time, and 3) to help construct the regression of preference parameters $\boldsymbol{\beta}$ as a function of time for estimating the preferences in a given year (see Sec. VII.A for details). To check whether the effect of a specific variable on route selection is consistently significant, we count the number of times that the p values are less than the level of significance at 0.1. To check if the preferences are consistent, we count the number of times that the estimated parameters have the same sign.

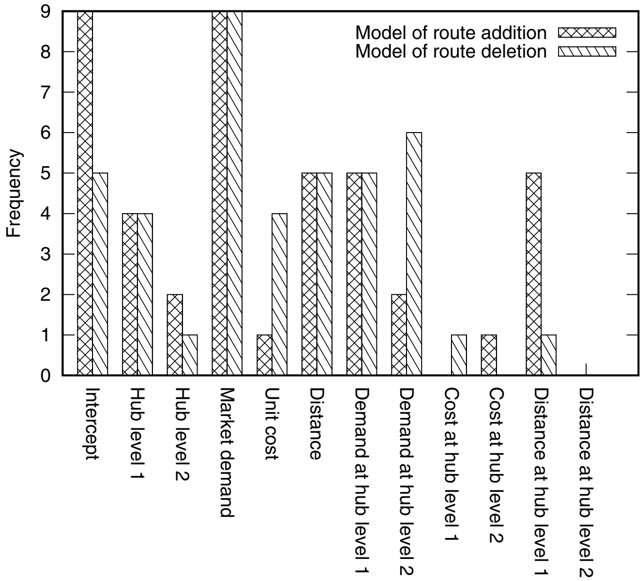


Fig. 3 Frequency of decision variables being statistically significant in nine evolution instances.

Figure 3 shows the bar plot of the frequency of decision variables being statistically significant in nine evolution instances. The results indicate that the effect of demand has always been significant in all nine evolution instances for both route addition and deletion. Since airlines' preferences in real scenarios do not change frequently over time, to be conservative, we only present the estimated preferences that are always positive or always negative, and meanwhile, they are always statistically significant. Thus, only the market demand is presented.

Figure 4 shows the estimated odds ratio of market demand in route addition/deletion, which is calculated based on the increase/decrease in every 1000 passengers on a route at each evolution from 2004 to 2013. The figure implies that the effect of demand on adding or deleting a route is fluctuating from 2004 to 2013; meanwhile, the route deletion decisions are more sensitive to the change in demand. For both route addition and deletion, it is observed that, in the evolution from 2006 to 2007, the change of demand has greater impacts on route selection than any other years. This behavior reflects the highly sensitive nature of the airline industry to economic

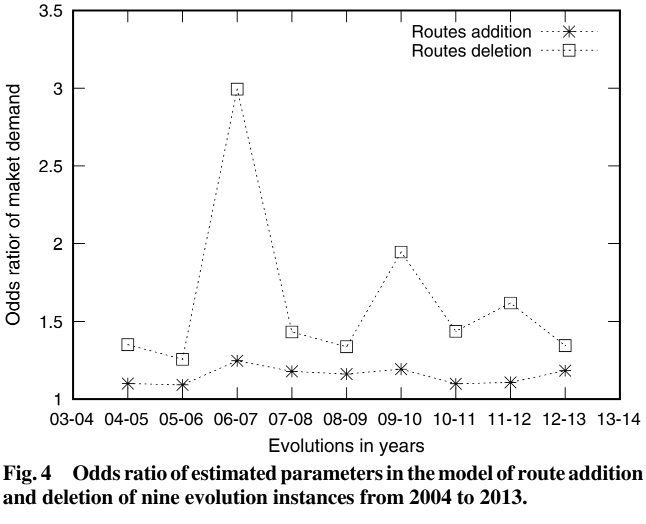


Fig. 4 Odds ratio of estimated parameters in the model of route addition and deletion of nine evolution instances from 2004 to 2013.

Downloaded by PURDUE UNIVERSITY on September 5, 2017 | http://arc.aiaa.org | DOI: 10.2514/1.D0015

changes. Specifically, the route selection was influenced starting from 2016 to 2017, which is actually one year before the economic downturn in 2008. This characteristic of the airline industry can be well captured by our models.

VI. Comparison of Airlines' Route Selection Models

To verify and evaluate our approach for modeling airlines' decisions on route selection, we compare the obtained models with those from Kotegawa et al. [12–14]. The DCMs are probabilistic as shown in Eq. (5), which are used to predict how likely a route between a city-pair will be added/deleted, with the estimated preference structure $\boldsymbol{\beta}$. These predicted routes are compared with the routes actually added/deleted to obtain the prediction accuracy, which is then compared with the prediction accuracy reported by existing models.

Prediction accuracy may be defined in several ways. In this paper, we choose sensitivity and specificity [32] for comparison because of their wide application in existing models. Equations (7) and (8) show how sensitivity and specificity are defined:

$$\text{sensitivity} = \frac{n_{\text{ps}}}{n_{\text{as}}} \tag{7}$$

$$\text{specificity} = \frac{n_{\text{pns}}}{n_{\text{ans}}} \tag{8}$$

where n_{ps} is the number of routes that are predicted to add or delete, and n_{as} is the number of routes that are actually added or deleted. Similarly, n_{pns} is the number of routes that are predicted to not select for adding or deleting, and n_{ans} is the number of routes that are actually not selected.

We apply this approach to nine evolution instances from 2004 to 2013. To account for the probabilistic nature of adding or deleting routes based on the predicted probability, we repeat the prediction process 10 times, as was done by Kotegawa et al. [13], and use the average value obtained. Figure 5 shows the prediction accuracy of each evolution instance of both route addition and deletion models. There are three prominent features of the models: first, in both measures, the accuracy of deletion is always higher, indicating that the model for route deletion has a better performance. This is because, from a practical perspective, it is harder to predict route addition than predicting deletion, as greater uncertainties are associated with the decision of establishing a city-pair route. For example, airlines are less risk tolerant and there are many regulations and policies (either at state or federal) that the airlines have to follow before adding a route. Airlines have to make a tradeoff about whether to enter or not by considering the associated cost, either monetary or time. On the other hand, the decision on deletion might be simpler: Is the route profitable? If yes, keep it; if not, delete it. Hence, the prediction accuracy of the model is higher. But from the computational perspective, it is actually harder to achieve a high accuracy in route deletion. This is because the choice set of deletion is normally larger than the choice set of addition, as reported in the supplementary material [15]. This, on the other hand, validates the appropriateness of our approach. Second, the model has consistently high specificity for both addition and deletion. The average values of specificity for nine years are 0.86 and 0.978 for the models of route addition and deletion, respectively. This means that the models are capable of accurately predicting the routes that are not selected. Third, the standard deviation of prediction accuracy at each run reflects that the model has a small variation in predictability.

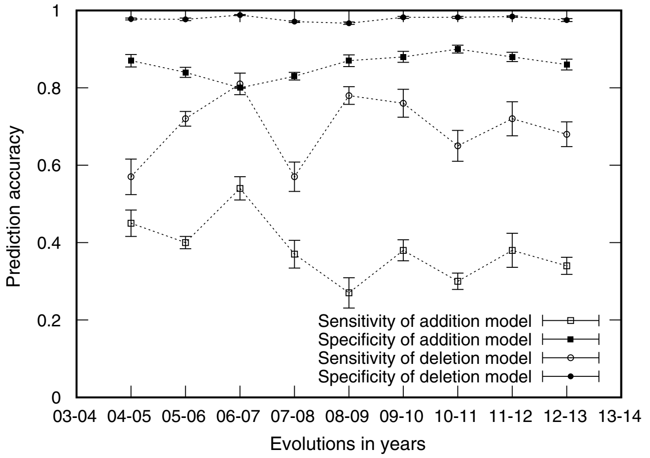


Fig. 5 Prediction accuracy of route selection in terms of sensitivity and specificity (the values are the average of 10 runs).

Existing models are developed mainly by three types of approaches: machine learning, modified network generation model, and heuristic. The models that are based on machine-learning techniques include support vector machines, artificial neural networks, logistic regression, and random forests. The modified network generation models include the fitness function model and the modified BA model. The heuristic-based approaches include the distance filtering model and other filtering principles, which are obtained by domain expertise. These models in literature [10,13,14] are compared with the random draw (RD) model of route addition and deletion.

In the following comparison, we only adopt sensitivity as the measure of prediction accuracy. Note that sensitivity is equivalent to accuracy 1 in [13]. We do not calculate accuracy 2, presented in [13], for comparison because accuracy 2 highly depends on the size of the choice set. As discussed in Sec. IV.B, our approach for generating the choice set is different from the approach in [13], which simply considers C_N^2 (all possible combinations of size 2) as the number of potential candidates for route addition. Our approach, by its nature, results in a high value of accuracy 2 because of the significant reduction in the size of the choice set.

In total, eight models of route addition and five models of route deletion are compared with our models based on the decision-centric approach. The model based on ANN is not compared because the ANN is computationally expensive and the authors of [12] only reported the prediction of two airline companies, Southwest and American Airlines, instead of the entire U.S. ATN. Also, the modified BA model is not compared because it only generates one data point, as it predicts network evolution from 1997 directly to 2007. Note that we are only able to perform the prediction for the years from 2004 to 2013 because of the lack of data. However, the number of evolution instances analyzed does not affect the prediction.** Table 2 shows the average sensitivity of our models for addition and deletion as compared with other existing models.

Our approach yields better results than all the models being studied for both route addition and deletion. The baseline approach that is based on RD produces low accuracy because it selects routes randomly. Among existing models of route addition, the RF at the beginning is capable of achieving similar accuracy performance as the DCM-based model. However, as reported in [14], a sharp decline of forecast accuracy happens outside its training dataset. This is a typical phenomenon for machine-learning algorithms that are overtrained. The same phenomenon is observed in the LR with network metrics variables in a route addition model. In both route addition and deletion models, the accuracy reached by LR with network metrics variables are closest to the results of our approach. However, the performance of the LR-based approach worsens when the size of the network shrinks, which is observed from the results of networks with 877 nodes and 244 nodes, as shown in Table 2. In contrast, our models are capable of generating high sensitivity, even if the analyzed ATN has just 134 nodes.

**We have performed a similar analysis by using the same approach for the network evolution instances from 2002 to 2009. The results support our argument.

Downloaded by PURDUE UNIVERSITY on September 5, 2017 | http://arc.aiaa.org | DOI: 10.2514/1.D0015

Table 2 Average sensitivity of the models of route addition and deletion

	Number of instances	Years	Average sensitivity
		Route addition	
DCM-based approach	9	04–13	0.38
SVM + LR	6	90–08 (not consecutive)	0.098
SVM + RD	6	90–08 (not consecutive)	0.016
RF	6	90–08 (not consecutive)	0.23
RD	6	90–08 (not consecutive)	0.0028
LR with 887 nodes	15	90–05	0.37
FF	15	90–05	0.17
LR with 244 nodes	14	90–04	0.087
FF with 244 nodes	14	90–04	0.06
		Route deletion	
DCM-based approach	9	04–13	0.70
LR + filtering	6	90–08 (not consecutive)	0.44
RD	6	90–08 (not consecutive)	0.054
LR	3	90–91; 98–99; 07–08	0.21
Filtering with distance	3	90–91; 98–99; 07–08	0.086
SVM	3	90–91; 98–99; 07–08	0.13

VII. Model Validation

Model validation is performed through network regeneration and forecast. The byproduct of this validation study is a topology generator for modeling the ATN and its evolution. In the network regeneration process, the network topology of the starting year is the seed for initializing the network evolution. With the predicted probability from DCA, routes from choice sets are selected to add and delete; thus, a synthetic network of the next year is obtained. This synthetic network is then used for regenerating a network of the following year. This process is repeated until the targeted year's network is obtained. The validation is achieved by comparing the synthetic networks with the real networks. Besides network regeneration, we also forecast the evolution of the ATN in the future. The forecast is performed to further validate the model when more uncertainty is imposed, as future data of explanatory variables are not accessible.

A. Network Regeneration

While regenerating networks for multiple years using a given seed network, the choice set of predicted networks may have predicted routes that had not been established in reality. Thus, these routes may lack of the data on market demand and DOC. To resolve this issue, for those routes, we use the data from the latest year during which that route existed. In addition, we cannot directly apply the estimated preference parameters ($\boldsymbol{\beta}$ values) from the real choice set to the predicted choice. To resolve this issue, we perform regression analysis of the estimated β parameters as a function of time to predict the preference structures for a given year. For example, this means doing a regression on the parameters of market demand, as shown in Fig. 4.

Because of the probabilistic nature of the network regeneration process, we regenerate the network from 2005 to 2013 five times, using the ATN of year 2004 as the seed network. In the nine evolution instances of regeneration, the average prediction accuracy in terms of sensitivity is 0.28 for route addition and 0.34 for route deletion. As compared with the average of prediction accuracy shown in Table 2, the decrease of sensitivity results from the linear regression of $\boldsymbol{\beta}$ values.

The validation is performed by comparing the topology of regenerated networks with real networks. Specifically, we calculate the associated metrics of networks in the year 2013, including the average degree, average CC, average path length (PL), etc. These network metrics are selected because they have been found to be effective in quantifying network topology, as reported in [33]. Also, in the context of the ATN, these metrics have been widely used [10–14]. For example, the degree indicates the total number of connections (city-pair routes) of a node. Thus, in the ATN, the average degree reflects the commercial route service provided by airlines. The CC is a measurement of local cohesiveness for a collection of nodes. It implies the local robustness and average CC reflects the system robustness. A higher average CC indicates that the ATN is more robust, as alternate connection paths may exist when a neighboring airport fails. Table 3 presents the detailed results of the metrics of the real network in 2013, as well as the average values and standard deviations of regenerated networks' metrics in five different runs. The largest percentage error among the five runs is less than 5%. Therefore, the results show that our model is capable of regenerating networks with structural properties close to the real networks.

Figure 6 shows a comparison of degree distribution and the distribution of CCs between the regenerated 2013 ATN and the real ATN. The difference is measured using the Kullback-Leibler (KL) divergence [34], which quantifies the information lost when the synthetic network is used to approximate the real network. The KL divergence of the degree distribution is 0.15, and it is 0.23 for the CC distribution. The closer this value is to zero, the better. The same analysis is done for all the other regenerated networks from 2005 to 2012. We observe that the models are capable of producing the edge dynamics and replicating the network evolution with small variation (see supplementary material [15] for details). These results validate the developed approach and the resulting models.

The differences between the real networks and the generated networks is possibly due to three reasons: 1) the probabilistic prediction error due to the statistical nature of the method applied, i. e., the DCA; 2) unavailability of data covering all the years; and 3) the linear regression analysis for estimating the preference structures in a given year.

B. Network Forecast

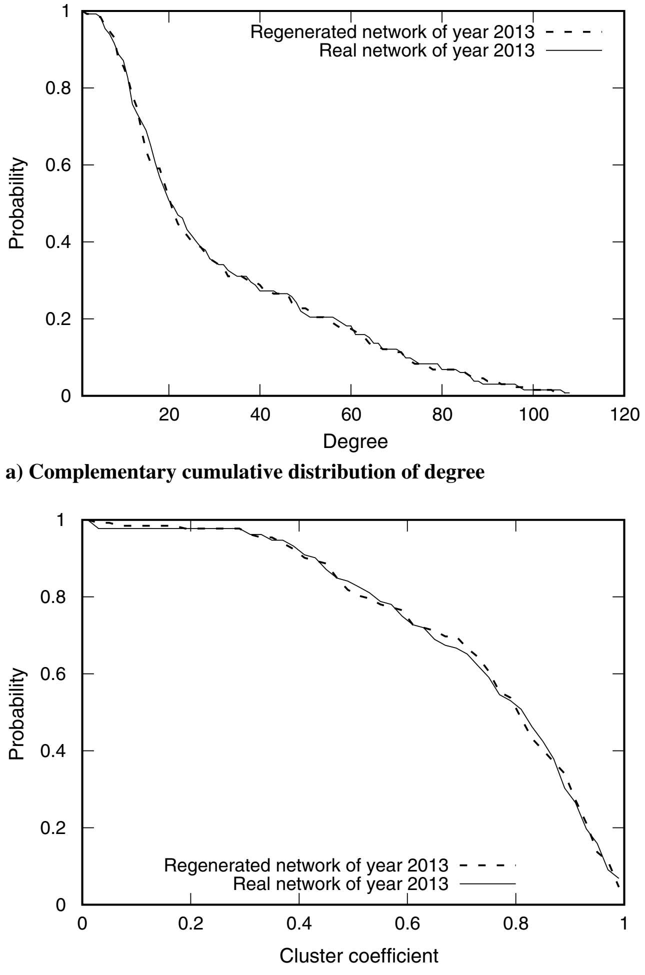
In a network forecast, not only are the preference structures unknown (and thus should be estimated) but the data of decision

Table 3 Comparison of network metrics of the real ATN and regenerated ATN in the year 2013 using the ATN in 2004 as the seed network[a]

	Average degree	Average CC	Average PL	Density	Hub = 0	Hub = 1	Hub = 2
Real network in 2013	31.045	0.731	1.823	0.237	588	1246	215
Regenerated network	30.664 (0.231)	0.738 (0.005)	1.828 (0.005)	0.234 (0.002)	567 (12)	1239 (8)	218 (3)
Largest % error	1.95	1.64	0.66	2.11	4.93	1.28	3.26

[a] The network is regenerated for five times. The mean and standard deviation are reported in the format of "mean (std)."

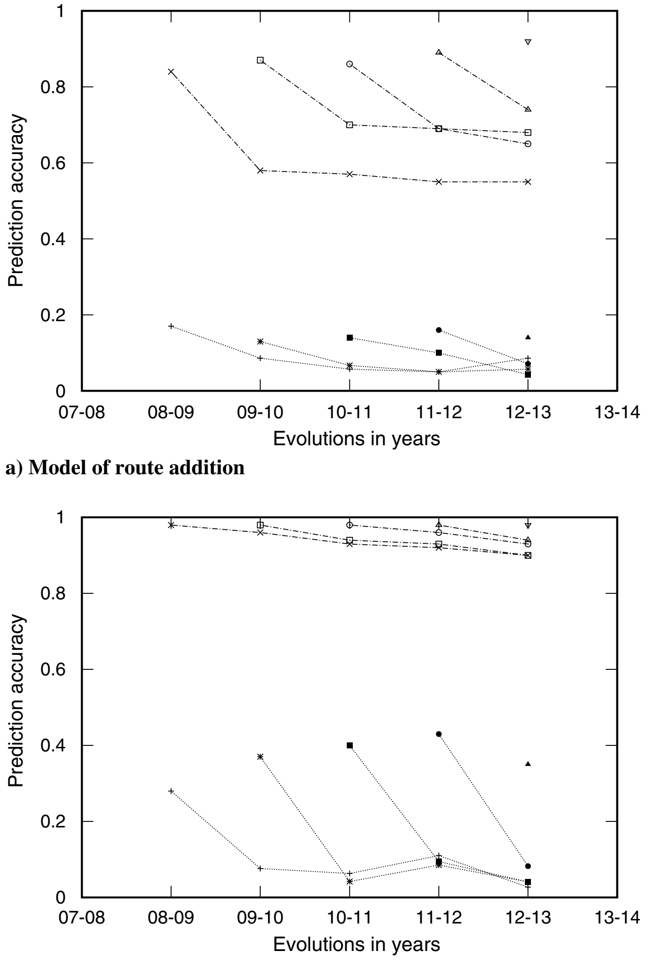
Downloaded by PURDUE UNIVERSITY on September 5, 2017 | http://arc.aiaa.org | DOI: 10.2514/1.D0015



a) Complementary cumulative distribution of degree

b) Complementary cumulative distribution of cluster coefficient

Fig. 6 Comparison of the distributions of the regenerated network with real network in 2013.



a) Model of route addition

b) Model of route deletion

Fig. 7 Prediction accuracy of the network forecast based on different starting years.

variables such as demand and DOC are also unknown. The latter aspect is fundamentally different from the network regeneration process. Therefore, the approach of the network forecast contains two parts: 1) the prediction of future preference structures, and 2) the prediction of future market demands and DOC. The prediction of preference structures follows the same method as discussed in the previous section. The prediction of market demand and DOC are performed by employing the FAA's forecast about future revenue passenger enplanements and fuel costs in a fiscal year [35]. Specifically, the predicted rate of growth in demand and cost are calculated, and then they are used to predict the demand and cost of each route in the next year.

To support the validation, instead of forecasting a network in the future, we assume that we are in 2012 and forecast the ATN in 2013. Furthermore, in order to evaluate the predictability of the developed models, we perform the forecast by assuming that we are in different years prior to 2013. In other words, we forecast the ATN in 2013, by assuming that we are in 2012, then assuming we are in 2011, and so on until the year 2008. The aim is to quantify the decrease in prediction accuracy of the model as we forecast further into the future, and consequently identify how far ahead can the model predict with an acceptable degree of accuracy.

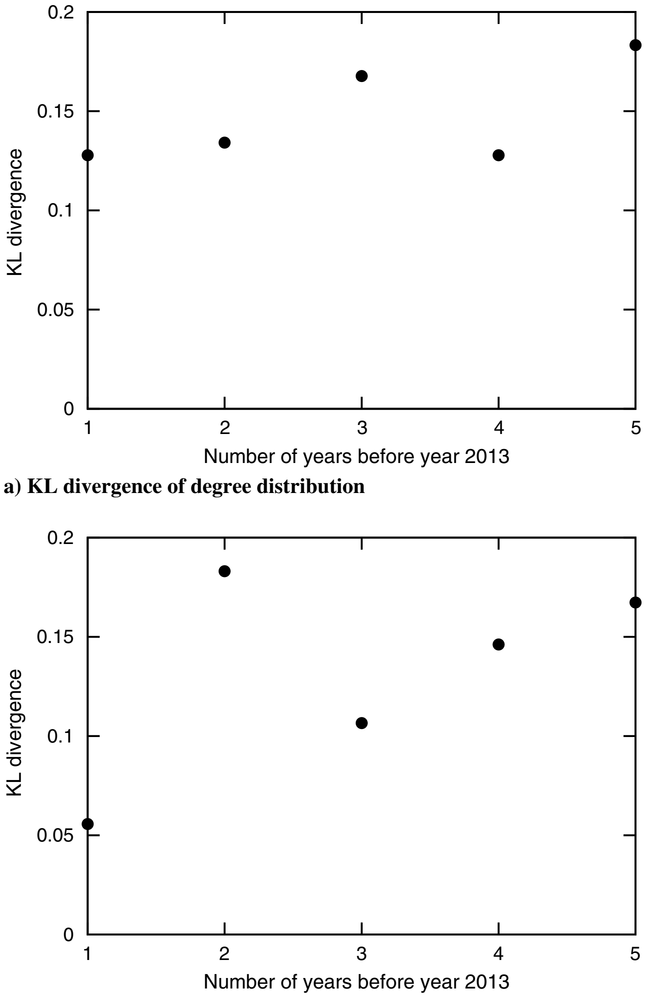
Figure 7 shows the prediction accuracy with different starting years. The dashed-dotted lines with hollow points at the top areas of figures represent specificity, whereas the dotted lines with solid points at the bottom area represent sensitivity. The lines, from left to right, indicate the results of models starting from different years before 2013. These results reveal the following insights:

1) The prediction accuracy is always higher in the forecast of first evolution than predicting multiple years ahead. For example, if forecasting the network of 2013 by assuming we are in 2008, the prediction accuracy of the first evolution (2008–2009) can reach above 20% on average but reduces to under 10% starting from the second evolution. The prediction accuracy is high in forecasting one year ahead because the seed network is the real network, which ensures that choice sets are the same as real ones.

2) After the first drop, the decrease in accuracy for each year ahead is not so significant; instead, the accuracy maintains at a certain level with fluctuation. For example, in the forecast starting from 2008, the sensitivity of both route addition model and route deletion model remains at 7% on average after the first drop, as shown in Fig. 7. Intuitively, the reduction should get more pronounced as the years increase. This is probably because of the approach for determining the choice set: as long as routes are not in the union set, they will not be deleted or added. So, for each evolution after the first one, the change in the network given the size of the union set will not be significant. This is also probably the reason why the prediction accuracy for deletion is always higher than that of addition.

To further reveal the impact of the number of years ahead to be predicted on the accuracy, we compare the metrics of the forecasted network of year 2013 with the real networks by calculating the percent error in metrics of the forecasted networks. We found the network metrics of the forecasted networks to be very close to the real network. For example, the percent errors of the average degree of the forecasted networks are 0.2, 0.7, 1.3, 0.8, and 4.7%, respectively, if forecasting with different starting years from 2012 to 2008. Besides, all the metrics listed in the first row of Table 3 have percent errors less than 6%, even if forecasting five years before 2013. In addition, we use KL divergence to quantify the difference between the distribution of forecasted networks and the real network. Figure 8 shows that 1) all

Downloaded by PURDUE UNIVERSITY on September 5, 2017 | http://arc.aiaa.org | DOI: 10.2514/1.D0015



a) KL divergence of degree distribution

b) KL divergence of cluster coefficient distribution

Fig. 8 Comparison of degree distribution and cluster coefficient distribution of the ATN in 2013 forecasted with different starting years.

forecasted networks have small KL divergence (less than 0.2) in degree distribution and CC distribution; and 2) the KL divergence increases, on average, as the number of years before 2013 increases. To conclude, the results shown in Figs. 7 and 8 help us draw the following conclusion: our approach is able to forecast the networks two years ahead with a prediction of about 10% in sensitivity and less than 1% error in network metrics.

In analyzing the error in the network forecast, the low prediction accuracy is probably due to the inaccurate prediction of data of market demand and DOCs. By comparing network regeneration and forecast, just because the use of real data in the network regeneration process, the synthetic network of 2013 has a smaller percent error, even if the starting year is 2004, i.e., nine years before 2013. The demand and DOCs as reported by the FAA [35] show that, as the number of years before the actual increases, the percent error of passenger enplanements (demand) increases. For five years before, such an error can change from 1% to above 10%. The propagation of error causes a significant reduction in prediction accuracy. Therefore, to have a better forecast, the approach for predicting the future demand and cost is crucial. Actually, we tried another approach to predict the demand and cost in a routewise way. Specifically, the demand and cost on a specific route are predicted based on historical data of that individual route using regression. However, because of the large fluctuation of the demand and cost in each year, the routewise approach turns out to be even worse. Even for forecasting one year ahead, some sensitivities can drop below 10%. The overall performance of the routewise approach is inferior to the aggregate approach. Thus, the results are not reported in this paper but can be referred to in [15].

VIII. Conclusions

A systematic approach is developed for estimating airlines' decision-making preferences on route selection based on discrete choice random-utility theory. The approach consists of methods for the identification of the air transportation network topology, the determination of choice set, and the comparison and validation of developed discrete choice models. Variables in the developed route selection models include market demand, direct operating costs, the distance between origin and destination airports, and whether a route connects hub airports or not. The main conclusions reached are the following:

1) The market demand is a significant decision factor for both route addition and deletion, and it is observed as always significant on a timescale.

2) The impact of the market demand varies, depending on whether a route connects hub or nonhub airports.

3) The market demand has greater impact on the decision of route deletion as compared with the route addition.

4) The direct operating cost is significant in the model of route deletion but not in the case of route addition. These revealed insights help people acquire deeper understanding about airlines' decisions on city-pair route selections.

As to the model's performance in prediction, the model of route deletion produces better results than that of route addition. Precisely, the model is capable of predicting routes to be added or deleted in the next year with an average accuracy of 38 and 70%, respectively. During validation, it is found that the regenerated networks match the topology of the real networks closely; and the percent error in network metrics, such as average degree, average clustering coefficient, and five others, is less than 5%. So, from the perspective of surrogate networks, the models' outputs are good enough to serve the purposes for which the model was constructed, as illustrated in Fig. 1. Furthermore, when forecasting the future evolution of a network (i.e., when both the data and model parameters are unknown), the model is capable of forecasting networks with 1% error in network metrics, if forecasting two years in the future. The prediction accuracy drops significantly if forecasting longer periods. In the future work, the results can be further analyzed to see if there are any patterns in those correct predictions versus those that were missed. Such an analysis will help in understanding the conditions under which the current model fails; thus, strategies of improvement can be developed.

This model can be further improved if more data are made available. For example, in the model of route addition, there are only seven observations of routes at hub level 2 in the choice set of the evolution instance of 2005–2006. With such a small number of observations, the model may not be able to get meaningful results. One way to solve this problem is to increase the time span in the analysis. In the future, as more datasets become available, the evolution instance can be set as a two-year period instead of a one-year period.

On the other hand, refer back to Fig. 1. The ultimate goal is to obtain an approximate model $D'(t)$ that can estimate airlines' route selection decisions given the explanatory variables $\boldsymbol{x}'$ as accurately as possible. In this paper, several insights on such decision making are drawn by using the demand, direct operating costs, hub/nonhub status, and the distance, as concluded in the first paragraph of this section. Since the developed DCA-based approach can be easily extended through the reconstruction of the utility function in Eq. (2), when data are available, other explanatory variables, such as seasonality, competition, and airline type (legacy versus low cost), can be tested and added to the model.

Another assumption in the current work is that only the routes that have been in operation in any one of the previous years are included in the choice set. This assumption is necessary for two reasons. First, because the approach is data based, a lack of data makes application of the approach difficult. Second, with 132 airports in the network, the decision model will have to be run over 8646 routes, which is the total number of all possible routes. This is unreasonable in certain instances, such as for cities that are very close with each other. Future

Downloaded by PURDUE UNIVERSITY on September 5, 2017 | http://arc.aiaa.org | DOI: 10.2514/1.D0015

work in this approach will relax these assumptions by considering additional criteria for identification of the choice set for route selection.

Finally, future work could also consider the development of a multilevel decision-making framework that incorporates the interactive decision making of passengers and agencies, such as the FAA, in addition to those of the airlines. Such a framework would help analyze the decisions of various stakeholders in the aviation industry and provide a holistic picture of the evolution and performance of the ATN.

Acknowledgments

The authors gratefully acknowledge the financial support from the National Science Foundation through the grant 1360361 from the system science program in the division of civil, mechanical and manufacturing innovation. We thank Joseph Thekinen from Purdue University for his help with some of the data collection. We thank Navindran Davendralingam and Apoorv Maheshwari from Purdue University for their help in reviewing the paper and providing domain-specific advice.

References

[1] Belobaba, P., Odoni, A., and Barnhart, C., "The Airline Planning Process," *The Global Airline Industry*, Wiley, West Sussex, U.K., 2009, pp. 153, 162–173.

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

[3] *Passenger Boarding (Enplanement) and All-Cargo Data for U.S. Airports* [online database], Federal Aviation Administration, https://www.faa.gov/airports/planning_capacity/passenger_allcargo_stats/passenger/ [retrieved Sept. 2016].

[4] Jaillet, P., Song, G., and Yu, G., "Airline Network Design and Hub Location Problems," *Location Science*, Vol. 4, No. 3, 1996, pp. 195–212. doi:10.1016/S0966-8349(96)00016-2

[5] Lohatepanont, M., and Barnhart, C., "Airline Schedule Planning: Integrated Models and Algorithms for Schedule Design and Fleet Assignment," *Transportation Science*, Vol. 38, No. 1, 2004, pp. 19–32. doi:10.1287/trsc.1030.0026

[6] Balakrishnan, A., Chien, T. W., and Wong, R. T., "Selecting Aircraft Routes for Long-Haul Operations: A Formulation and Solution Method," *Transportation Research Part B: Methodological*, Vol. 24, No. 1, 1990, pp. 57–72. doi:10.1016/0191-2615(90)90032-T

[7] Cordeau, J. F., Stojković, G., Soumis, F., and Desrosiers, J., "Benders Decomposition for Simultaneous Aircraft Routing and Crew Scheduling," *Transportation Science*, Vol. 35, No. 4, 2001, pp. 375–388. doi:10.1287/trsc.35.4.375.10432

[8] Lederer, P. J., and Nambimadom, R. S., "Airline Network Design," *Operations Research*, Vol. 46, No. 6, 1998, pp. 785–804. doi:10.1287/opre.46.6.785

[9] 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. doi:10.1287/trsc.18.1.1

[10] Kotegawa, T., Han, S. Y., and DeLaurentis, D. A., "Implementation of Enhanced Network Restructuring Algorithms for Improved Air Traffic Forecasts," *9th AIAA Aviation Technology, Integration, and Operations Conference (ATIO)*, AIAA Paper 2009-7043, 2009.

[11] Song, K., Lewe, J. H., and Mavris, D., "A Multi-Tier Evolution Model of Air Transportation Networks," *AIAA Aviation 2014*, AIAA Paper 2014-3267, 2014.

[12] Kotegawa, T., "Analyzing the Evolutionary Mechanisms of the Air Transportation System-of-Systems Using Network Theory and Machine Learning Algorithms," Ph.D. Dissertation, Purdue Univ., Lafayette, IN, Jan. 2012.

[13] Kotegawa, T., DeLaurentis, D. A., and Sengstacken, A., "Development of Network Restructuring Models for Improved Air Traffic Forecasts," *Transportation Research Part C: Emerging Technologies*, Vol. 18, No. 6, 2010, pp. 937–949. doi:10.1016/j.trc.2010.03.004

[14] Kotegawa, T., DeLaurentis, D., Noonan, K., and Post, J., "Impact of Commercial Airline Network Evolution on the US Air Transportation System," *Proceedings of the 9th USA/Europe Air Traffic Management Research and Development Seminar (ATM'11)*, 2011.

[15] Sha, Z., Moolchandani, K., Panchal, J., and Delaurentis, D., "Modeling Airlines Decisions on Route Selection Using Discrete Choice Models—Data and Supplementary Materials," Purdue Univ. Research Repository, Lafayette, IN, 2015. doi:10.4231/R74747TG

[16] Sha, Z., and Panchal, J., "Estimating the Node-Level Behaviors in Complex Evolutionary System," *Journal of Mechanial Design*, Vol. 136, No. 6, 2014, Paper 061003.

[17] Sha, Z., and Panchal, J., "Estimating Linking Preferences and Behaviors of Autonomous Systems in the Internet Using Discrete Choice Models," *IEEE International Conference on Systems, Man, and Cybernetics Conference*, IEEE Publ., Piscataway, NJ, 2014, pp. 1591–1597.

[18] Ben-Akiva, M., and Lerman, S. R., *Discrete Choice Analysis: Theory and Application to Travel Demand*, MIT Press, Cambridge, MA, 1985, pp. 100–129.

[19] Train, K., *Discrete Choice Methods with Simulation*, Cambridge Univ. Press, New York, 2009, pp. 41–85.

[20] Chen, W., Hoyle, C., and Wassenaar, H. J., *Decision-Based Design: Integrating Consumer Preferences into Engineering Design*, Springer, New York, 2013, pp. 79–92.

[21] Williams, H. C. W. L., "On the Formation of Travel Demand Models and Economic Evaluation Measures of User Benefit," *Environment and Planning*, Vol. 9, No. 3, 1997, pp. 285–344. doi:10.1068/a090285

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

[23] *Air Carrier Financial: Schedule P-5.2* [online database], U.S. Dept. of Transportation, Bureau of Transportation Statistics, http://www.transtats.bts.gov/Fields.asp?Table_ID=297 [retrieved Oct. 2014].

[24] Kelley, K., and Maxwell, S. E., "Sample Size for Multiple Regression: Obtaining Regression Coefficients That Are Accurate, Not Simply Significant," *Psychological Methods*, Vol. 8, No. 3, 2003, pp. 305–321. doi:10.1037/1082-989X.8.3.305

[25] *Statistical Abstract of the United States: 2012*, 131st ed., U.S. Census Bureau, Books Express Publ., Aug. 2011, pp. 677–683.

[26] *Corporate Fact Sheet* [online database], United Airlines, Chicago, IL, http://newsroom.unitedcontinentalholdings.com/corporate-fact-sheet [retrieved Sept. 2016].

[27] *Delta Air Lines Newsroom — Global Network* [online database], Delta Airlines, Atlanta, GA, http://news.delta.com/global-network [retrieved Sept. 2016].

[28] *American Airlines Group* [online database], American Airlines, Fort Worth, TX, http://phx.corporate-ir.net/phoenix.zhtml?c=117098&p=irol-IRHome [retrieved Sept. 2016].

[29] 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, Nos. 694–706, 1933, pp. 289–337. doi:10.1098/rsta.1933.0009

[30] Hailpern, S. M., and Visintainer, P. F., "Odds Ratios and Logistic Regression: Further Examples of Their Use and Interpretation," *Stata Journal*, Vol. 3, No. 3, 2003, pp. 213–225.

[31] Cameron, A. C., 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. doi:10.1016/S0304-4076(96)01818-0

[32] Peng, C. Y. J., Lee, K. L., and Ingersoll, G. M., "An Introduction to Logistic Regression Analysis and Reporting," *Journal of Educational Research*, Vol. 96, No. 1, 2002, pp. 3–14. doi:10.1080/00220670209598786

[33] Albert, R., and Barabasi, A. L., "Statistical Mechanics of Complex Networks," *Review of Modern Physics*, Vol. 74, No. 1, 2002, pp. 47–97. doi:10.1103/RevModPhys.74.47

[34] Kullback, S., and Leibler, R., "On Information and Sufficiency," *Annals of Mathematical Statistics*, Vol. 22, No. 1, 1951, pp. 79–86. doi:10.1214/aoms/1177729694

[35] "FAA Aerospace Forecasts Fiscal Years 2015–2035," Federal Aviation Administration [online database], 2015, https://www.faa.gov/data_research/aviation/aerospace_forecasts/ [retrieved Sept. 2016].

K. Bilimoria
Associate Editor