A Preliminary Analysis of the Evolution of US Air Transportation Network

Dipasis Bhadra and Brendan Hogan*
Center for Advanced System Development (CAASD)
The MITRE Corporation
7515 Colshire Avenue
McLean, VA 22102

ABSTRACT

How does the air traffic network evolve over time? Is there any pattern to how the air traffic network evolves with respect to the size of the markets, type of competition, geospatial features, and structural changes following the events of September 11, 2001 (9/11)? Are the changes transitory or permanent in nature? Can we lay out the trajectory possibilities of the network and determine factors influencing them?

The National Airspace System (NAS) in the United States US) is structured primarily around a web of air transportation markets linked to each other through a network of 465 commercial airports located in and around 363 metropolitan statistical areas (MSAs). The total number of origin-destination (O&D) markets in the NAS ranges somewhere between 36,000 – 40,000 pairs depending upon seasons and economic cycles. In its present structure, these markets are hierarchical; a small number of markets account for the largest number of passengers and, hence, air traffic flows. For example, there were approximately 105 markets (0.3% of the total) which had 1,000 or more passengers a day (i.e., thick markets), but these accounted for almost 17% of the total passengers. On the other hand, there were almost 28,000 markets (78% of the total) with 10 or fewer passengers a day that accounted for only 6% of total passengers in 2003. These O&D market pairs have been served, generally speaking, by 52,000 – 56,000 flight segments (i.e., routings passengers took to travel the markets) depending upon the extent and intensity of network. In recent years, however, the network segments have increased sizably to an average of 67,000 – 72,000 segments leading to increased fragmentation.[1]

Understanding the evolutionary nature of the airline network is extremely important. Investment decisions with consequences stretching far into the future are being made to serve the airline network. A proper understanding of the dynamic nature of the airline network, therefore, is essential to minimize costly mistakes.

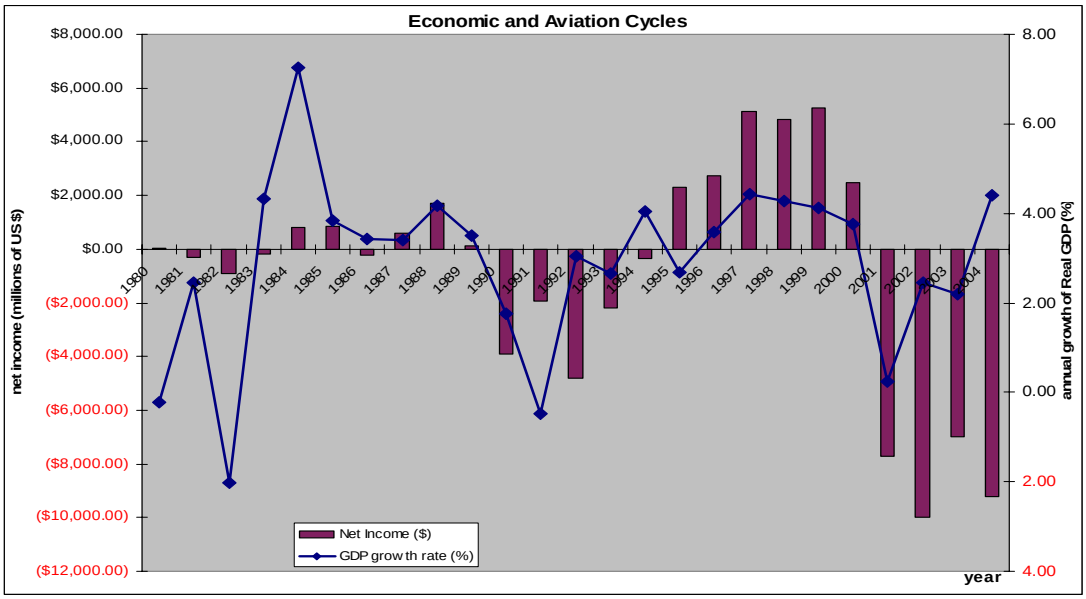
We used a multinomial logit model to analyze and determine itinerary choices in the US scheduled airline industry. Using 10% ticket sample data for the second quarter of 2003, we find that passengers, weighted average fare, average distance and types of air carriers empirically determine the itinerary choices. This simple model captures lower-order itinerary choices (i.e., those with less than three stops) fairly well for the sample of almost 360,000 itineraries.

---

* Paper to be presented at the 5th Annual ATIO/AIAA conference to be held at the Crystal City, VA during September 26 – 28, 2005. Authors are an economist and a senior simulation engineer, respectively, with the MITRE Corporation's Center for Advanced Aviation System Development (CAASD). An earlier version of this paper was presented at the 46th annual conference of the Transportation Research Forum in Washington DC. We thank all who participated and provided valuable suggestions that improve this version of the paper. All remaining errors are ours only. Corresponding author: dbhadra@mitre.org

I. Introduction

Events unfolding with the recession in the spring of 2001 plunged the US aviation industry into turmoil. The accumulated aviation net income from the last quarter century ($ 25.2 billion during 1980 – 2004) will have been completely wiped out by the losses accumulated during the 15 quarters (2001:Q2 – 2004:Q4) following the recession that began in March 2001, a staggering total of $33.92 billion.[2†] This tremendous loss occurred despite the fact that the US government had provided a cash grant of $5 billion and opened a loan grant program totaling $10 billion soon after the events of 9/11. Recently, the Congress has approved a bill providing further grants amounting to a little over $2 billion to airlines[‡] in addition to covering other security costs.



Source: Air Transport Association (2003); BEA (2003) and Industry Reports; 2004: Preliminary estimates

Figure 1. Economic and Aviation Cycles

The aviation business has been, for the most part, cyclical in nature. Generally speaking, economic growth preceded the expansion in air transportation [see Figure 2] and performance of both providers and manufacturers within the industry.

---

† For the entire length of the quarter century following the deregulation of the industry in 1978, the accumulated losses ($48.8 billions) exceeded accumulated profit ($25.2 billions) by a 2:1 margin. In other words, the industry has lost two dollars for every dollar it made.

‡ In the first week of April 2003, the United States Congress approved a $3.2 billion (H. R. 1467) proposal offering to reimburse struggling airlines security fees paid to the government since the hijack attacks of September 11, 2001, while the Senate approved (S. 728) a more complicated plan worth $3.5 billion. The compromise bill, worth $3.9 billion, was approved in April 2003.

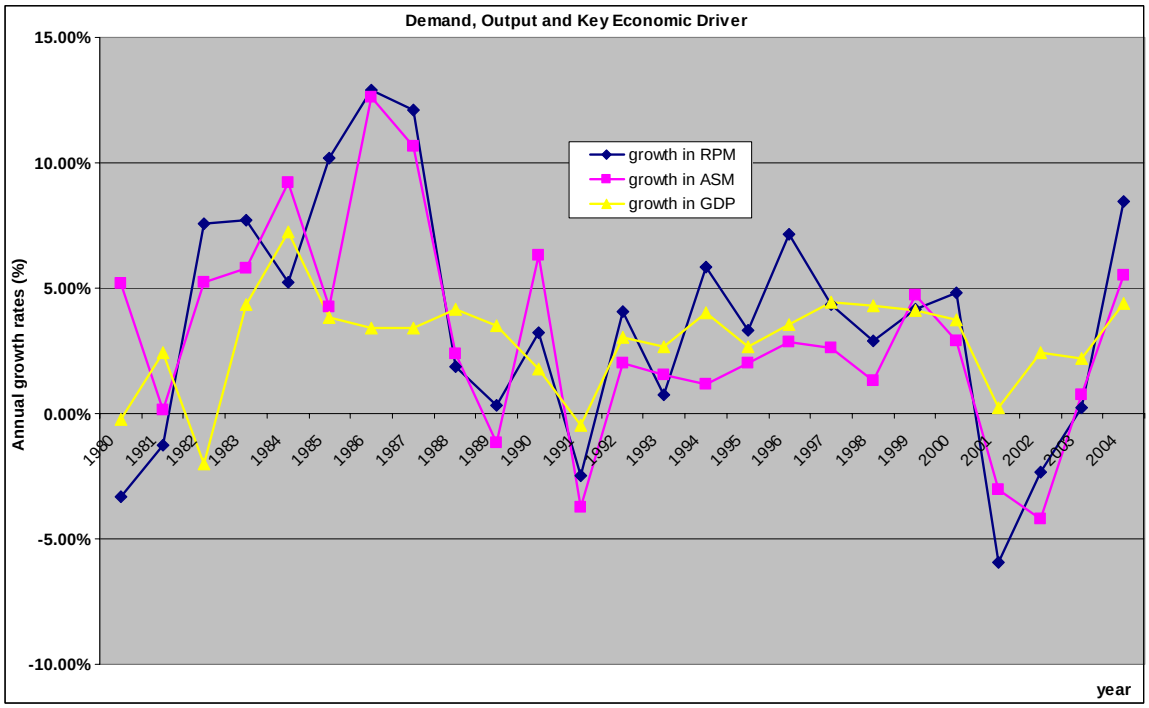


Figure 2. Demand, Capacity and Key Economic Driver

In the past, a large part of this adjustment, has been attained through inventory control (i.e., control and expansion of available seat miles (ASM)) resulting in an ever expanding network of hub-and-spokes [see Figure 3]. Network carriers have controlled the growth in ASM in response to demand (i.e., revenue passenger miles, or RPM) which have been largely influenced by the economic growth (i.e., growth in gross domestic product or GDP). Hence, a strong positive correlation, sometimes with a lag, has generally been observed throughout the last quarter century between GDP and RPM, and RPM and ASM. When growth in ASM lagged growth in RPM either the load factor or the average size of the aircraft (i.e., aircraft gauge) was adjusted upward in the short run and in the long run, respectively.

These changes have occurred concurrently with the overall expansion of the hub-and-spoke networks.[§] As Figure 3 below demonstrates, a large part of passenger growth had been absorbed via the expansion of hub-and-spoke networks.

[§] Airport hubs are defined in two ways. First is in terms of total enplanements (i.e., physical counts), as defined by Department of Transportation (DOT)/Federal Aviation Administration (FAA). Under this definition, there are four kinds of airports: large ($\geq$ 1% of total enplanements), medium (0.25 – 0.999), small hubs (0.05 – 0.249); and nonhub (< 0.05). The second definition categorizes an airport where a major commercial air carrier has more than one bank structure as a hub, i.e., operational or functional definition. Under this definition, an airport is defined as a hub when inbound flights are scheduled to arrive from multiple origins within a short space of time thus creating a bank of passengers. The coordinated arrival and departure banks together form a wave of activities and leads to peakiness in airline schedules. At present, some of the physical hubs are functional hubs as well. However, an airport can be an operational hub without being a physical hub (e.g., airports primarily serving connecting passengers) while a physical hub may exist without being an operational hub, e.g., airports primarily serving origin and destination passengers.

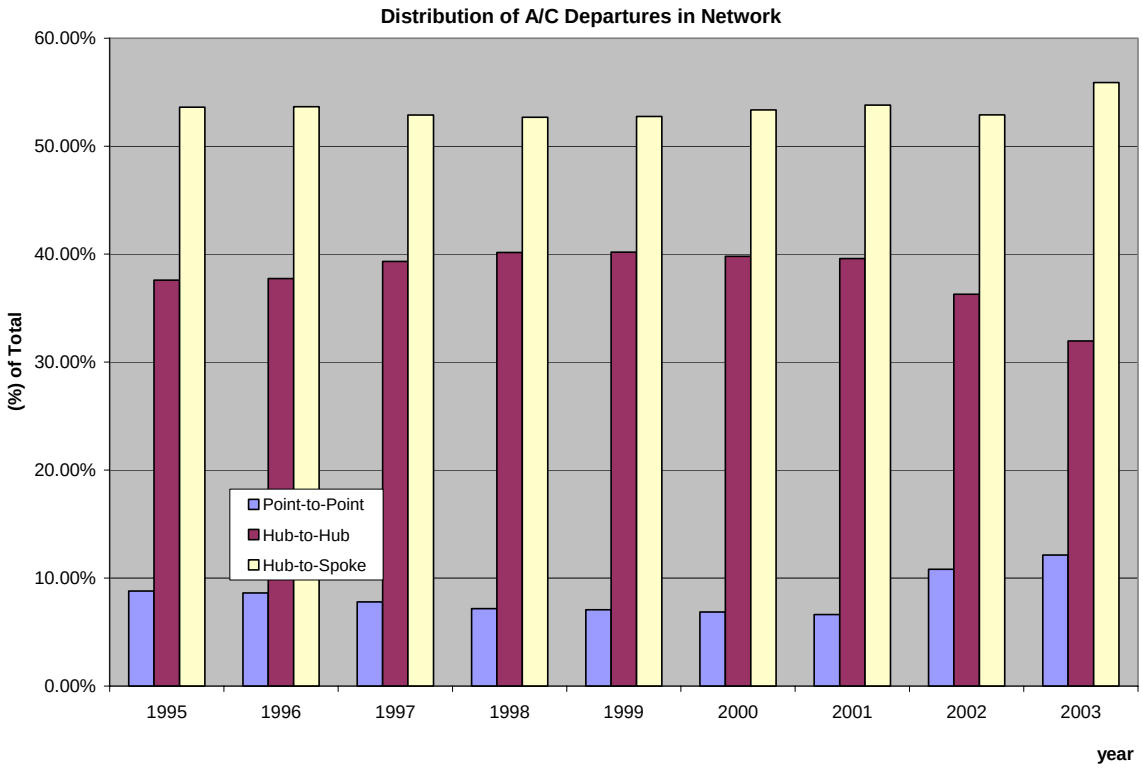


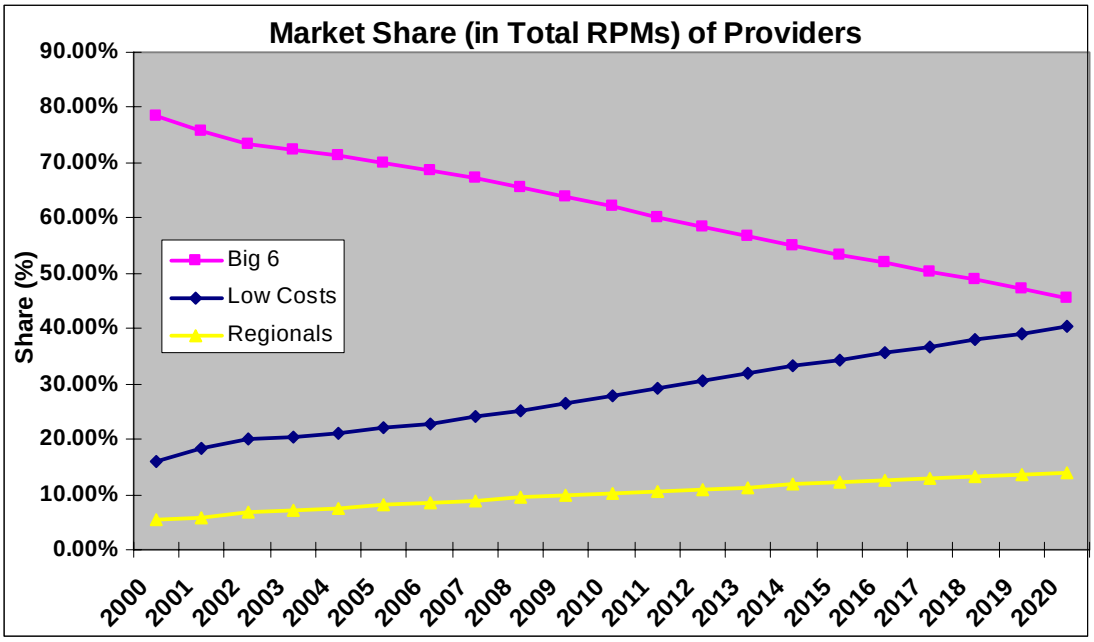
Figure 3. Distribution of Aircraft Departures in the Network

The current economic cycle that began with the mild economic slowdown of the Spring of 2001 is different in nature from all other past downturns. The impact of economic recession was exacerbated by the tragic events of 9/11. These were accompanied with longer-term structural changes such as use of internet technology in airline booking and accelerated market share of low-cost airlines in the network.

The mainline network carriers[**] consisting of six airlines[††] have shrunk their operations from that observed in 2000. In comparison, carriers providing direct services to origins and destinations (i.e., point-to-point carriers), regional carriers, and some other low cost network carriers have been observed to expand services during this period [see Figure 3]. Like previous recessions, 1990 – 1991 in particular, the low-cost and regional carriers have gained significant market share. This gain appears to have come exclusively from the network carriers losing market shares. Many analysts predict[3] a continued expansion of this trend, thus giving a very different outlook for the US airlines network in the future [see Figure 4 below].

[**] Network carriers run their operations primarily through a system of hub-and-spoke airports.

[††] American, United, Delta, Continental, US Airways, and Northwest In 2002, the big six together accounted for around 73% of total revenue passenger miles (RPM).



Source: Airline Monitor (2003)

Figure 4. Market Share by Provider Type

Understanding the evolutionary nature of the airline network is extremely important. Investment decisions with consequences stretching far into the future are being made to serve the airline network. A proper understanding of the dynamic nature of the airline network, therefore, is essential to minimize costly mistakes.

This paper is an attempt to understand the evolutionary nature of the US airline network. In particular, we address two empirical issues: What are the fundamental factors that determine and drive evolutions in US airline network? Second, if these changes in network are empirically meaningful, how can they be explained? Answering these issues may provide us with some important insights leading to improved policy-making in an environment that appears to be ever changing. The paper is organized as follows: Section II gives a quick background on the US airline network, followed by the recent transformation arising from industry responses to the recent downturn. Section III introduces the empirical framework, together with data and results. Section IV concludes the paper with policy suggestions and ideas for further research.

II. Airline Network in the United States

Air transportation in this country has a hierarchical structure. Much of the scheduled air transportation passes through the large hubs [see Figure 5], a feature that is consistent with the population distribution.[4] The purpose of airports, especially ones with heavy use, is to meet people's air travel needs. Thus, large physical hub airports appear to coincide with concentrated demographic and economic activities, a feature of large metropolitan areas as well.[5]

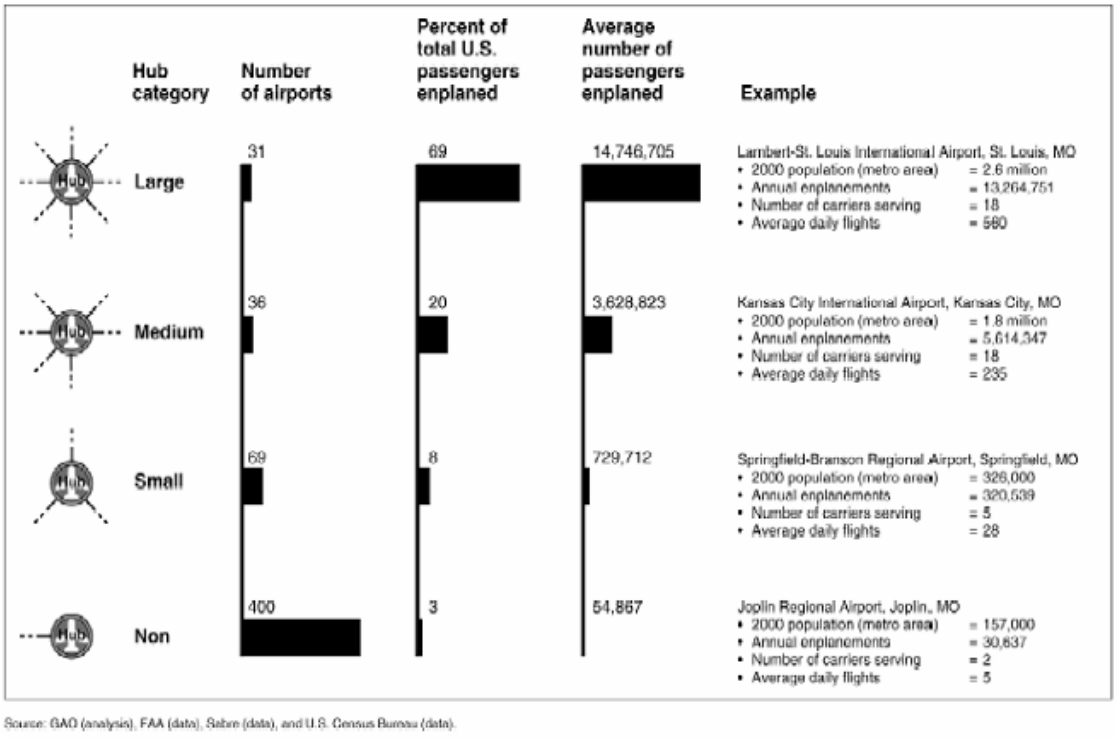


Figure 5. Passenger Flows Through US Airline Network

It is evident from Figure 3 that airline networks supporting hub-and-spoke activities are predominate. In a hub-and-spoke (HS) network flights are scheduled to arrive from multiple origins within a short space of time, thus creating a bank of passengers who then depart to multiple destinations within a short space of time through a well-coordinated schedule structure. The coordinated arrival and departure banks together form a wave. Formed efficiently, hubs act as switching centers intermediating flows between multiple origins and multiple destinations as well as contributing origin and destination traffic of their own. Figure 6 provides a stylized picture of how this works.

Hub-and-Spoke
Directed, Scheduled, Aggregated

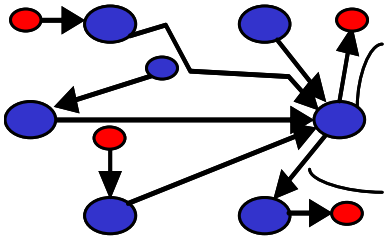
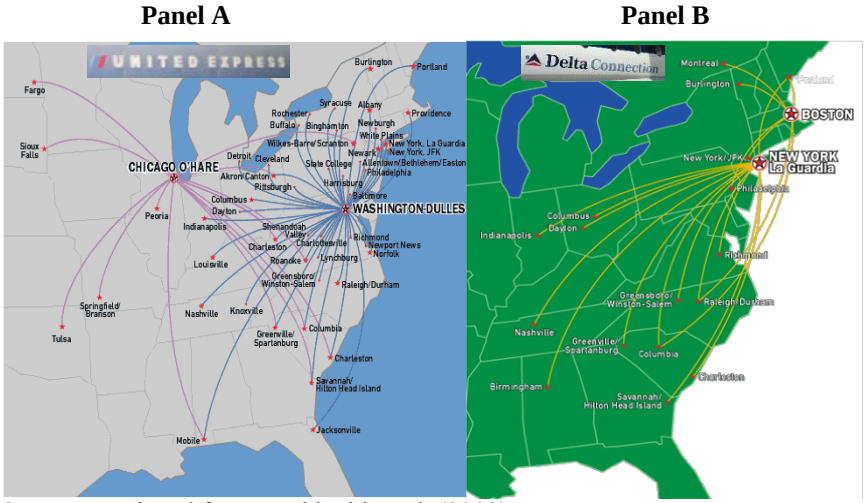


Figure 6. Hub-and-Spoke (HS) Network: Conceptual Framework

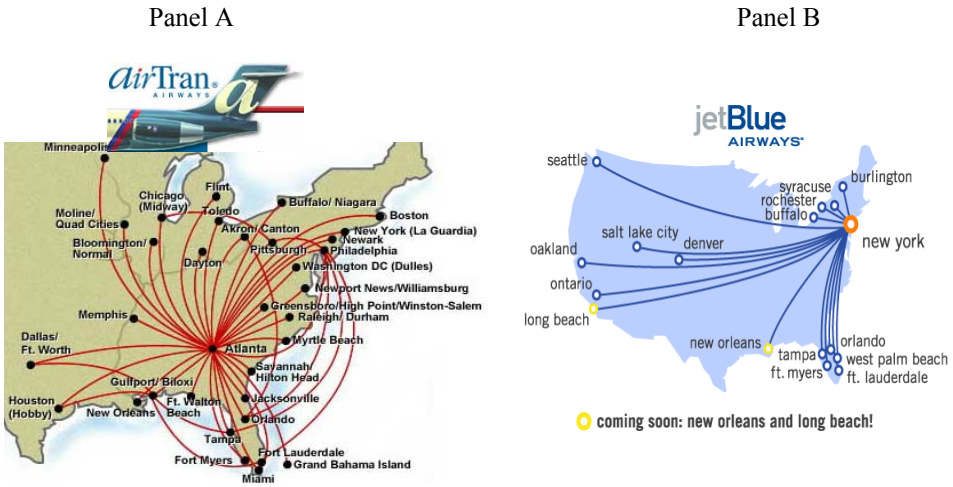


Source: Retrieved from world wide web (2003)

Figure 7. Hub-and-Spoke (HS) Network: Code Sharing Partners for Network Carriers

Most of the aggregation, however, takes place through mainline carriers' arrangements with regional carriers, also known as code sharing. While many mainline carriers fly to other hubs and/or spoke airports directly, they often enter into these contract arrangements with subsidiaries, directly owned or contracted with regional airlines to distribute passengers into spoke airports.[4] As the Panel A in Figure 7 demonstrates, large aggregation points for United's hub-and-spoke network are at Washington's Dulles International Airport (IAD) and Chicago's O'Hare International Airport (ORD) airports, via United Express. Similarly, Delta aggregates its northeast passengers at LaGuardia Airport(LGA) and General Edward Lawrence Logan International Airport (BOS).

Whether hubs connect directly to destination airports or via other hubs depends on market features such as size and composition of the market, fare, connection possibilities.[6] Hubbing is not limited to network carriers. Some of the low-cost carriers (LCCs) also aggregate passengers at their major hub airports. Figure 8 demonstrates two such situations where two major LCCs are distributing passengers directly to spoke airports (case of AirTran) and directly to destinations (case of jetBlue).



Source: Retrieved from world wide web (2003)

Figure 8. Hub-and-Spoke (HS) Network: Low-Cost Carriers

For a point-to-point (PP) network, on the other hand, airlines may fly directly, ideally speaking, from one airport to any other within the NAS. Figure 9 gives a stylized example of such a network.

Point-to-Point
Directed, Scheduled
Aggregated

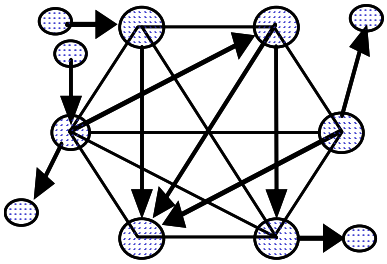
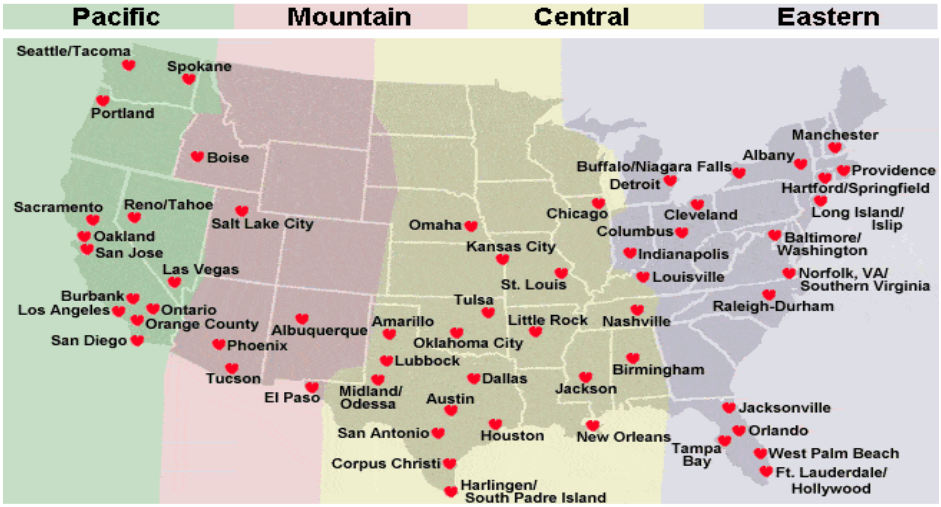


Figure 9. Point-to-Point (PP) Network: Conceptual Framework

Southwest Airlines is the nation's largest low-cost carrier, also carrying the largest number of domestic passengers within the country. After the events following the recession in 2001, Southwest has become a major force in the US air travel through aggressively gaining market share in areas where other mainline carriers have retreated. Given an apparently distributed network structure [see Figure 10] and less emphasis on banking, it is often assumed that Southwest indeed has an PP network. Thus, an expanding Southwest may have a distributory impact, as opposed to the presently observed hierarchical structure, if the airline begins to capture more market share. This process may be further enhanced if other low-cost carriers entering into the market follow Southwest's network structure.



Source: Retrieved from world wide web (2003)

Figure 10. Point-to-Point (PP) Network: Southwest Example

Last, but not least, a changing network may have an immense impact on manufacturers. For example, Airbus has been aggressively pursuing the very large aircraft (VLA) A380 model over the last few years. As it stands currently (second quarter, 2005), the A380 now has over 150 firm orders. In order for it to be economically viable, Airbus may require around 250 orders. Boeing, on the other hand, recently abandoned the *Sonic Cruiser* program in favor of more traditional fuel-efficient jet carriers (B787) to counter Airbus' VLA development. If one evaluates the size, speed, and types of markets these two models are to serve, the underlying assumptions relating to networks appear to be opposite; the A380 appears to be continuing with the assumption that an HS network will continue to dominate in size and proportion, while B787 is primarily designed to serve more point-to-point traffic with speed and fuel efficiency.

Network theory offers a framework for system-level analysis of air transportation architectures as networks.[7, 8] The underlying analytical framework of such theories provides useful tools for quantitative analysis of network behaviors with respect to critical determinants such as services to the customers, cost, performance and competitive structure of the industry, robustness and vulnerability of the network.

Applying these theories to the air transportation network allows us to incorporate topological characteristics along with other characteristics in an analytical framework. When the topological flows (e.g., Baltimore-Washington International Airport (BWI) to Los Angeles International Airport (LAX) flow as east-west flows) are combined with market characteristics (e.g., BWI being the largest operation point for Southwest and LAX being one of the major gateway international airports), they can provide important insights into understanding how airline networks develop and function. Indeed, networks are the result of interactions between topology and economic and demographic factors.[5] Thus, understanding the scalability (i.e., achievement of small-world behavior) of the Southwest network and the functional characteristics, for example, can be a primal factor in understanding emerging networks at all layers in the topologies within which Southwest Airlines operates or may operate in the future. The absence of air transportation topologies underlying earlier analytical framework has confined much of our focus on the infrastructure and transport layers in the architecture.[8] The presence of a topology combined with market and economic factors allows for better mental models of the linkages between mobility, operations, transport, and infrastructure layers.

This paper is an attempt to understand the nature of airline networks using itinerary choices as the primary focus variable that incorporates both the topology and market characteristics. In particular, we examine how the itinerary choices are determined using market topology (i.e., directional flows) and market characteristics (e.g., passenger volume, fare, and types of carriers). The goal is to better understand the determinants of the itineraries which, in turn, determine the flows of passengers and operations in the US NAS.

III. Determining Choice of Routes in the NAS: An Econometric Framework

A. Binary Logit

In situations where airlines have only two choices of routes (e.g., direct flight versus one-stop flight) to assign to flight segments, there is essentially a binary qualitative choice.[‡‡] Since a linear probability model does not guarantee the predicted values of that choice to lie between (0, 1), it requires a process of translating the values of the attribute X (i.e., vector containing explanatory variables for the choice) to a probability which ranges in value from 0 to 1. We would also like to maintain the property that such a transformation would allow increases in X to be associated with an increase (or decrease) in the dependent choice variable for all values of X. Together, these requirements suggest the use of the cumulative probability function (F). F is defined as having its value equal to the probability that an observed value of a variable (for every X) will be less than or equal to a particular X. The range of F is then (0, 1) since all probabilities lie between 0 and 1. The resulting probability distribution may be expressed as follows:

$$P_i = F(\alpha + \beta X_i) = F(Z_i) \tag{1}$$

where α and β are the parameters of the model and F represents the distribution for each observation i. Common models in this category include Probit (standard normal), Logit (logistic), and Gompit (extreme value) specifications for the F function. The two cumulative probability functions, the normal (Probit) and the logistic, have been widely used in the literature and among practitioners.[9, 10]

To understand the logit specification, let us assume that there exists a theoretical continuous index Z_i which is determined by an explanatory variable X_i, for each observation i. Thus, we can write,

$$Z_i = \alpha + \beta X_i \tag{2}$$

Observations on Z_i are not available unless we have data that distinguish whether individual observations are in one choice category (e.g, direct itinerary choice) or a second choice category (e.g., non-direct itinerary choice). Logit methodology allows us to solve the problem of how to obtain estimates for the parameters while at the same time obtaining information about the underlying index Z.

[‡‡] This section is developed for demonstration purposes. Two itinerary choices (i.e., direct versus non-direct) are rather simplistic. Nonetheless, binary choice logit provides a conceptual framework that is relatively easy to understand.

Let Y represent a dummy variable that equals 1 when the itinerary 1 is chosen and 0 when the other category is chosen.[§§] Then assume that each individual choice Z_i* represents the critical cutoff value which translates the underlying index into a choice decision, such as,

$$\begin{array}{ll} \text{Category 1} = 1 & \text{if } Z_i > Z_i^* \\ \text{Individual choice for} & \\ \text{Non-category 1} = 0 & \text{if } Z_i \leq Z_i^* \end{array} \tag{3}$$

In this case, the threshold is set to zero, but the choice of a threshold value is irrelevant as long as a constant term is included in X_i. The logit model assumes that Z_i* is a cumulative distribution function for the logistic distribution, so that the probability that Z_i* is less than (or equal to) Z_i can be computed from the probability distribution function. The standardized cumulative probability distribution function for the logistic distribution is written as:

$$P_i\left(y_i = 1 \middle| x_i, \beta\right) = \frac{e^{\mu\beta' x_{in}}}{e^{\mu\beta' x_{in}} + e^{\mu\beta' x_{jn}}} = \frac{1}{1 + e^{-\mu\beta'\left(x_{in} - x_{jn}\right)}} \tag{4}$$

where x_{in} and x_{jn} are vectors describing the attributes of alternatives i (Itinerary = 1) and j (Itinerary = 0); μ is a scale parameter that is positive in value. When the parameters of the Z_i are linear, the parameter μ cannot be distinguished from the overall scale of the β's.[11] Oftentimes, μ is assumed to be equal to 1. By construction, the variable P_i will lie between (0,1). P_i is the probability that an event occurs, (e.g. P_1 is the probability of the itinerary choice 1, direct travel).[***]

B. Standard Multinomial Logit Model (MNLM)

Oftentimes, choices are not restricted to a binary set. The choices of itinerary for assignment in flight segment/s are often numerous. The majority of the US airlines have multiple choices of itineraries. Under such circumstances, the choice set will have to be expanded into multinomial choices. Thus, when there is more than one itinerary to choose from, (e.g. category of itinerary choices = 1,…,J), the probabilities associated with all those choices are P_1, P_2, …, P_J. However, these probabilities will sum to 1: $P_1 + P_2 + \ldots + P_J = 1$.

For unordered qualitative variables (also known as polytomous variables) such as itinerary choices by the airlines, categories must be truly nominal and mutually exclusive. Furthermore, the ordering of the numerical values of the variables is also of no importance. Therefore, any category can be used as the baseline category. However, such choice is usually based on some *a priori* theoretical or operational motivation.

From equation (4), for j > 2, the probability mass function can be generalized as follows:

$$P_n(i) = \exp^{\mu\beta' x_{in}} / \Sigma_{j \in C_n} \exp^{\mu\beta' x_{jn}} \qquad \forall\ i \in C_n \tag{5}$$

where i = i-th choice belonging to the complete set of choices, C_n. when j = 2, equation (5) reduces to equation (4), i.e., binary logit (0 and 1 being special case). Furthermore, equation (5) defines a proper probability mass function since $\forall\ i \in C_n$,

$$0 \leq P_n(i) \leq 1 \tag{6}$$

and,

$$\Sigma_{j \in C_n} P_n(i) = 1 \tag{7}$$

[§§] Instead of strictly defining one category of itinerary, one can also put all others in one category. For example, choice of one category (direct flights) and all others (i.e., all non-direct flights) can be defined under this binary choice.

[***] Binary choices have been widely discussed in the literature, primarily to explain voting behavior[12] for a theoretical framework; and, http://www2.chass.ncsu.edu/garson/pa765/logit.htm for applications in voting behavior context).

That is, the probability of individual choices, $P_n(i)$ is positive (equation 6) and they sum to 1 (equation 7). Furthermore, all disturbances in the choices are assumed to be (i) independent, (ii) identically distributed, and (iii) logistically-distributed. (i) – (iii) together are also known as *iid property*.[11]

The maximum likelihood function (ℓ^*) for a generalized multinomial choice model is given by the following equation:

$$\ell^* = \prod_{n=1}^{N} \prod_{i \in Cn} P_n(i)^{y_{in}} \tag{8}$$

Equation (5) describes a logit model for which parameters are linear corresponding to equation (8). Taking the logarithm of equation (8), we seek to attain maximum of ℓ^* as follows:

$$\ln \ell^* = \sum_{n=1}^{N} \sum_{i \in Cn} y_{in} \left(\beta' x_{in} - \ln \sum_{i \in Cn} \exp^{\beta' x_{jn}} \right) \tag{9}$$

Taking the first-order derivative of $\ln \ell^*$ with respect to coefficients (β_k) and setting it equal to zero, we derive the first-order conditions as follows:

$$\sum_{n=1}^{N} \sum_{i \in Cn} \left[y_{in} - P_n(i) \right] x_{ink} = 0 \qquad \forall \ k = 1,\ldots, K \tag{10}$$

or,

$$1/N \sum_{n=1}^{N} \sum_{i \in Cn} y_{in} x_{ink} = 1/N \sum_{n=1}^{N} \sum_{i \in Cn} P_n(i) x_{ink} \qquad \forall \ k = 1,\ldots, K \tag{11}$$

The estimator of β_k that maximizes the above function ℓ is consistent, asymptotically normal, and asymptotically efficient.[9] Equation (11) states that average value of attributes for the chosen alternatives (lefthand side of the equation) is equal to the average value predicted by the estimated choice probabilities (righthand side of the equation).[11] If an alternative-specific constant, for example, is defined for a particular alternative i, then, at the maximum-likelihood estimates, ℓ^*, equation (11) reduces to the following:

$$\sum_{n=1}^{N} y_{in} = \sum_{n=1}^{N} P_n(i) \tag{12}$$

which implies that the sum of all choice probabilities for alternative i, taken over the sample, will equal the number in the sample that actually chose i. The estimated vector, β_k, is a vector consisting of slope parameters that will determine the effect of the X vector on the probabilities of i-th choices. The computational methods and processes for solving the system of K equations in equation (12) are identical to those used in the binary logit case described earlier.

Our empirical framework consists of six qualitative choices for airline routing: direct route; one-stop, two-stop, three-stop, four-stop, five-stop and above, which we refer to as Category 1 – Category 6, respectively. The breakdowns of these itineraries have been summarized in the following figure [Figure 11]. There were almost 360,000 distinct itinerary choices in the second quarter of 2003. Interestingly, only 3% of these distinct itineraries (i.e., a little over 9,000) had direct trips, but these contained the majority share of the passengers (64% of the total).

One- and two-stop itineraries appear to dominate with combined share of 89% of the total itineraries. Four- and five-stops itineraries are very few, totaling less than 1% [Figure 11].

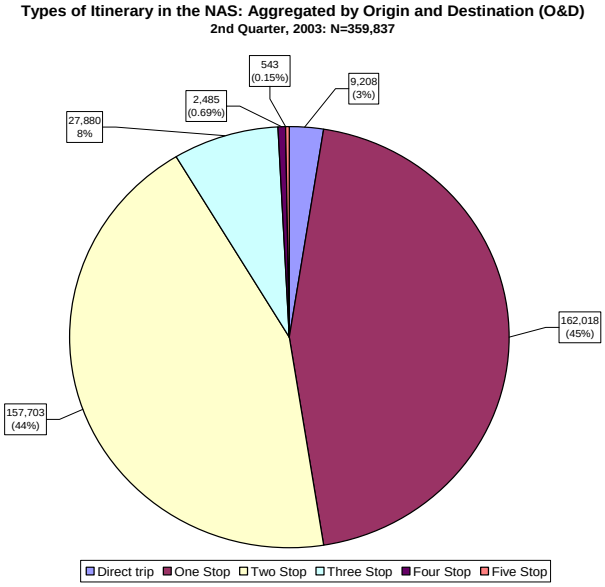


Figure 11: Categorization of Scheduled Hops or Market Categories in the US NAS: Case of Q2:2003

In terms of modeling the itinerary choice, the vector of explanatory variables (the X vector) consists of the following variables: inline passengers, average distance, passengers in the O&D market, weighted average fare, and types of carriers' present. We consider the levels of passengers as an exogenous variable although we acknowledge that it too is determined via a set of other exogenous variables (i.e., fares, income, population[4]) that presently falls outside the scope of this model. Finally, the US air transportation network is heavily dependent on types of carriers servicing different markets. At present, almost a quarter of passengers are served by low-cost carriers while the rest are being served by a combination of network and regional carriers. Many of the regional carriers enter into contracts with network carriers serving smaller markets. In order to identify the competitive forces in markets, we formulate two dummy variables, each representing the presence (=1) or absence (=0) of that type of carrier in the market.

Given these, our empirical framework can be specified as follows:

$$P_i(y_i = j \mid x_i, \beta) \quad (j = 0, 1, \ldots, 5) = \alpha_{ij} + \beta_1(\text{passengers\_Inline}) + \beta_2(\text{Average Distance}) + \beta_3(\text{Passengers\_O\&D Market}) + \beta_4(\text{Weighted Average Fare}) + \beta_5(\text{Presence of Network Carriers}) + \beta_6(\text{Presence of LCC Carriers}) + \varepsilon_i \quad (13)$$

where P_i represents the probability of six itinerary choices, from direct routing (=0) to five stops (=5); passengers_inline are the total number of passengers being served by a particular itinerary routing; average distance is the ultimate distance between origin and destination in miles; Passengers_O&D Market are the total number of passengers being served by the entire O&D market (i.e., aggregate of all passengers_inline); Weighted Average Fare is the average fare weighted by the share of passengers; and two dummy variables (i.e., 0 = absent; 1 = present) representing Presence of Network and LCC Carriers.

We use maximum likelihood (ML) estimation procedure for estimating (13). There are two reasons for which ML is often chosen as a general approach for estimating logistic regressions, especially for large samples. First, ML estimators are consistent, asymptotically efficient and asymptotically normal. Second, it is fairly straightforward to derive ML estimators. These are desirable properties given that we use large samples[13, 14] in our empirical analysis.

C. Data and Sources

Data for this exercise comes from the Bureau of Transportation Statistics/Department of Transportation's (BTS/DOT) Origin (O) and Destination (D) survey (which is called DB1B). DB1B is a 10% sample of airline tickets from reporting carriers. It includes such items as the reporting carrier, number of passengers, ticket fare, and total miles flown for each itinerary, as well as information about whether the itinerary was domestic or round-trip. This file is related to both the O&D segment and market files by the unique itinerary ID on each record.[15]

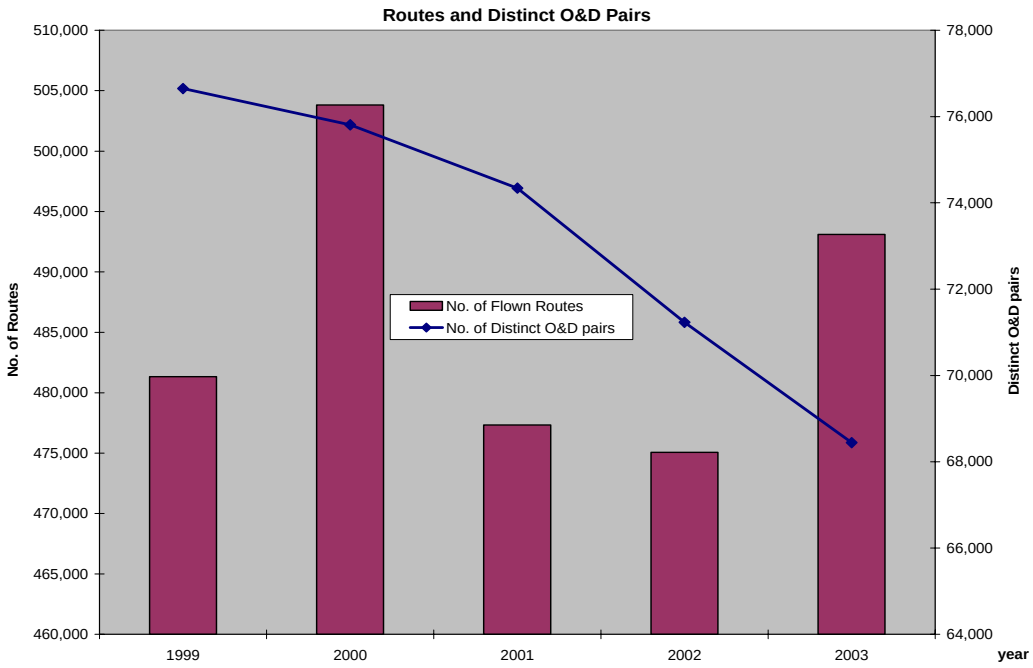


Figure 12. No. of Itineraries and Distinct O&D Market Pairs Served in the US NAS

For demonstrating the empirical analysis reported in this paper,[†††] we use data for the second quarter of 2003. During this period, there were 493,100 itineraries that were reported under 10% sample. This was reduced to 359,837 observations when aggregated by distinct passenger routings.[‡‡‡] On the other hand, when these itineraries were aggregated over distinct O&D markets, there were 68,442 observations. Clearly, not all the markets were served equally. It is obvious that while the top markets were served with more itinerary choices (e.g., BOS – LAX), relatively smaller markets (i.e., thin markets with less than 100 passengers a day) were given less choices.

[†††] The dataset is quite large. With more than 100 million records, quarterly 10% data is available for the period 1993 – 2004. In order to limit our analysis to a manageable magnitude within a reasonable computing time, we use the second quarter of 2003 as a sample case. However, the same empirical framework can be repeated for other quarters as well.

[‡‡‡] In other words, 133,263 itineraries were served by more than one reporting carrier.

It is interesting to note that while the number of itineraries appears to be somewhat declining [Figure 12] over time, the number of distinct O&D market pairs seem to have gone down definitively. In other words, although there are fewer destination choices there are more choices as to how to get there. This seems to corroborate, *albeit* indirectly, the flattening nature of the network.

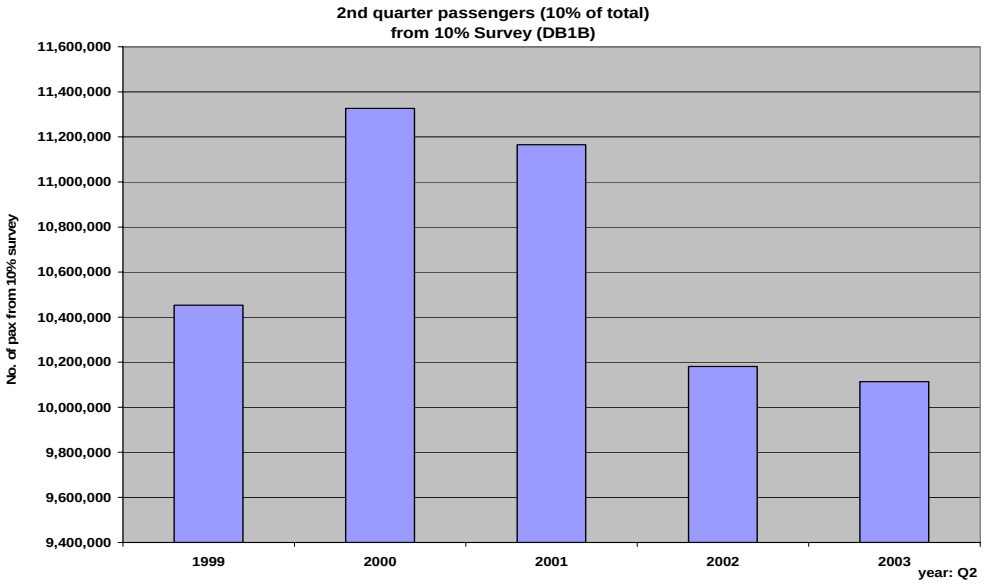


Figure 13. Number of Passengers (10%) from 10% Survey

On the other hand, the total number of passengers that are covered by the itineraries and O&D market pairs have remained stable over, on average, 106 million each second quarter. On average, the 10% ticket sample covers around 10.6 million average passengers every second quarter [see Figure 13]. With its peak passengers of around 11.3 million in second quarter of 2000 (therefore, in total will be around 113 million), the years following tragic events of 2001 saw a steady decline to the tune of 11.1 to 10.2 million in the second quarters of 2002 and 2003, respectively [see Figure 13 above].

The US air travel network has a primacy structure. Although there are 485 commercial airports in the country at present, the top 35 airports[§§§] (31 large hub airports and 4 medium hub airports[****]) known as the Operational Evolution Plan (OEP) airports account for the majority of the scheduled passenger flows in the country. For example, these airports together accounted for 73% of total scheduled enplanements and 69% of total scheduled aircraft operations in 2002.

---

§§§ These 35 airports are also known as Operational Evolution Plan (OEP) airports. OEP is a major FAA initiative to meet emerging air transportation needs for the next 10 years. For more details, see http://www.faa.gov/programs/oep/index.htm.

**** Large hubs are defined as those with ≥ 1% of national enplanements while medium hubs are defined are those with 0.5 – 0.999% share of national enplanements.

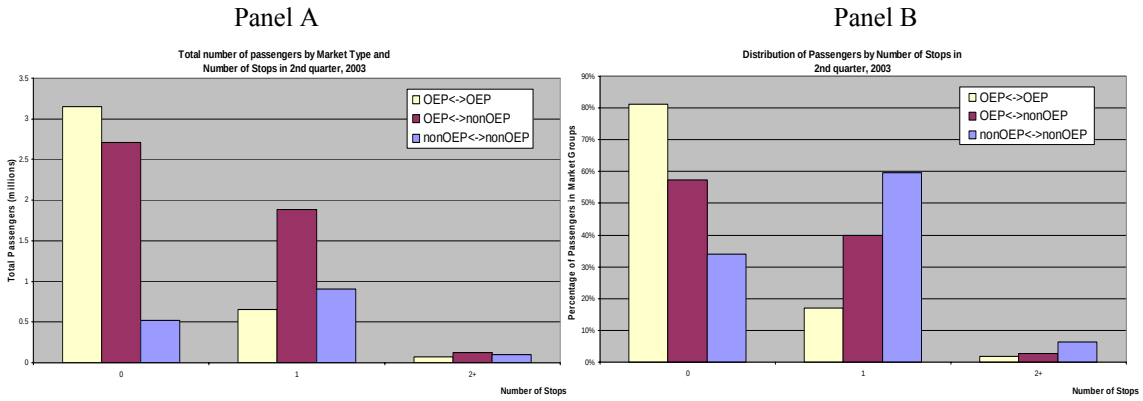


Figure 14. Passengers (total and %) by Market Hops

This primacy structure is also revealed in the itinerary choices. It is evident that a proportionately large part of the travel that takes place between OEP airports takes place using direct routing (i.e., zero stops). Travel between two OEP airports is conducted via direct trips a little over 80% of the time, followed by one-stop itineraries at a little over 15% [see Panel B in Figure 14]. Consequently, the largest numbers of passengers are making direct trips between OEP airports, with over 30 million in the second quarter of 2003 [see Panel A in Figure 14]. Itineraries between OEP airports and non-OEP airports via direct trips came in a close second with more than 25 million passengers served. A little over 5 million passengers traveling between non-OEP airports and non-OEP airports availed themselves of direct flights in the second quarter of 2003 [see Panel A in Figure 14]. A proportionately large share traveled between OEP airports and non-OEP airports via one-stop itineraries. One-stop itineraries accounted for almost 40% (or, almost 20 million total passengers) of this market group. These are clearly the spoke ends of the hub-and-spoke networks

IV. Results

The multinomial logit procedure was used to estimate Equation (13). Multinomial logit models (MNLMs) use maximum-likelihood estimation for polytomous dependent variables,[††††] and hence it is also known as polytomous logistic regression. Notice here that the groups formed by the categories of a polytomous dependent variables are not truly independent (i.e., choice of one itinerary may also depend on other itinerary choices as well) thus preventing one from simply doing as many separate logistic regressions as there are categories. Multinomial logit handles non-independence by estimating the models for all outcomes simultaneously except, as in the use of dummy variables in linear regression, one category is used as baseline. Since effects must sum to zero, the model for the reference group can be reproduced given the other parameters. For the estimation, direct routing (itinerary choice 0) is used as the baseline. This category is chosen as the baseline because it serves as the lowest category in cardinal ranking. Therefore, all other categories can be thought of as a cardinal upgrading over direct itinerary choice.

[††††] Polytomous variables are also known as unordered qualitative variables, such as itinerary choices in air travel. The ordering of the numerical values of the variables as such has no importance. Notice also that these categories must be truly nominal and mutually exclusive.

Table 1. Logistic Regression Results for Itinerary Choice (for second quarter of 2003)

Parameters*	One-Stop Versus Direct Route	Two-Stop Versus Direct Route	Three-Stop Versus Direct Route	Four-Stop Versus Direct Route	Five-Stop Versus Direct Route	Direct Route Versus All Non-Direct Routes**
Intercept	1.3093	1.7519	0.8474	-1.3892	-2.2153	-1.5764
Passengers_Inline	-0.0129	-0.4963	-1.8749	-2.9525	-2.6818	0.0154
Passengers_O&D Market	0.00177	0.00187	0.00192	0.00196	0.00199	-0.00182
Weighted Average Fare	0.00616	0.00496	0.00460	0.00538	0.00502	-0.00586
Average Distance	0.000282	0.00128	0.00161	0.00176	0.00181	-0.00062
Presence of Network Carriers	0.4609	0.4126	0.6635	1.0609	-0.00869	-0.4640
Presence of LCC Carriers	-0.7429	-1.4307	-1.4664	-1.5098	-2.7836	0.9311

Notes: '*': All parameters are statistically significant at greater than 99% level of significance;'**': There are two ways of deriving this. First, we can rerun logit program using a different base and derive the parameters; and/or use all non-direct routes (i.e., itinerary stops $\geq$ 1) as a choice against the alternative of direct route as a binary model. We run the latter to extract the model parameters for direct route.

Results from the estimation have been summarized in Table 1. It is important to note here that interpretation of the coefficient values is not the same under qualitative choice models as they are under linear and many non-linear models. It is complicated by the fact that estimated coefficients, i.e., *effect coefficients,* from an MNLM model cannot be interpreted as the marginal effect on the dependent variable. Nonetheless, their signs and magnitudes provide important information. Estimated effect coefficients, for example, represent the change in the *log odds* of the dependent variable, i.e., a particular type of itinerary choice due to changes in the explanatory variables. Despite the difficulties in explaining estimated coefficients directly, positive values of β_i would imply that increasing β_i will increase the probability of selecting a particular itinerary, and vice versa. As noted earlier, the estimated parameters (β_k) hold for individual choices, estimated over the entire sample maximizing log-likelihood function (ℓ). Like their counterparts in linear models, estimated β_k from multinomial logit models represent, on average, the sample as well.

Estimated parameters in the model (Table 1) indicate that passengers (both inline and market), average fare, distance, and types of carriers—all are statistically significant on the odd ratios of all itinerary choices. In particular, in line passengers (passengers_inline defined as the total number of passengers being served by a particular route) tend to affect itinerary choices negatively. Furthermore, the more in-line passengers there are, the less likely it is to have higher order or more stop itineraries from the defined set of choices (i.e., negative monotonic). Interestingly, however, passengers in an O&D market (defined as the total number of passengers in the entire O&D market) tend to exert a positive influence on the itinerary choices in a positive monotonic fashion (i.e., higher the number of passengers, the higher the number of itinerary choices). Together, these two results imply that larger markets would tend to have more itinerary choices but lesser stops will be availed. The weighted average fare (defined as the average fare weighted by share of passengers) tends to exert a positive influence on itinerary choices but the magnitude goes down with higher order of itinerary choices. In other words, travelers tend to pay less for more stops (this is more apparent for one, two, and three stops), although they value positively more itinerary choices.

Finally, two dummy variables (i.e., 0 = absent; 1 = present, representing the presence of network and LCCs, respectively) show the inherent qualitative differences between these two sets of carriers. While the presence of network carriers positively influences more itinerary choices, the presence of LCCs, on the other hand, does just the opposite. The signs and the magnitude of these two variables lend credence that while network carriers, in general, pursue hub-and-spoke types of air travel (via series of sequenced itineraries), the LCCs, on the other hand, pursue point-to-point network (via direct routing)[1] for empirical analysis of these aspects via passenger dimensions].

Although the model is estimated on the sample as a whole, and hence, estimated parameters are valid for the sample, oftentimes, results of the Logit choice model are evaluated at each observation point. Evaluating the above model at each observation point and comparing the observed probabilities with that of actual probabilities may, in fact, reveal further information regarding the structure of the model and hence the underlying choices. The estimated

parameters can be used to predict the itinerary choice responses at the individual observations in order to evaluate the model's performance. These predicted responses were compared to that of actual choices. This experiment has been reported in the following figure [Figure 15].

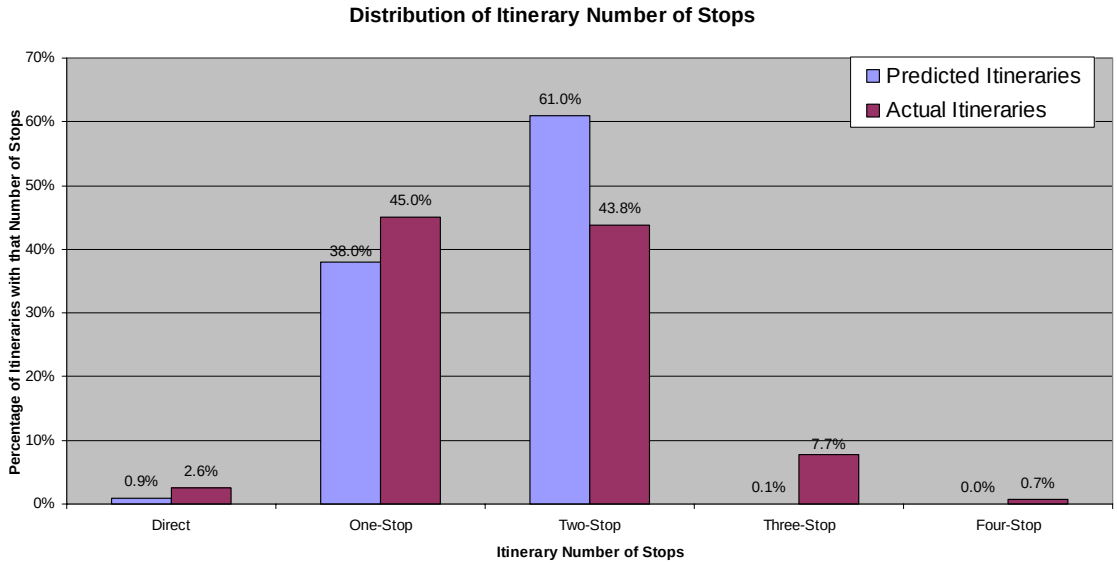


Figure 15. Validation and Verification of Results

Comparing the overall distribution of itineraries produced from the model with that of the actual data gives some insight into the predictive power of the logit model developed.[‡‡‡‡] Overall, the choice model seems to perform well for one and two stop itineraries and somewhat well for direct routing. The model does not perform well in picking itineraries with more than two stops. Many of these actual multi-stop itineraries end up as two-stop trips in the current formulation of the model, which accounts for some of that spike in the predicted itinerary distribution [see Figure 15].

Clearly, there are quite a few limitations in modeling itinerary choices in this particular way that can be expanded in future work. First, the proposed framework does not yield any information regarding where passengers stop over, i.e., choices are not location-specific. A reference network without carriers (i.e., absence of any airline behavior modeling) makes it difficult to point out where passengers will stop. This may be justified because, from an operational perspective, it does not really matter which airlines are flying, rather how many places they may be stopping is more important. Second, the model is good in predicting trips requiring fewer connections, (i.e., direct trips up to two hops), but does not seem to be that good for predicting more than two hops. Clustering itinerary choices into a small number of groups may improve model results.

[‡‡‡‡] It is also possible, although not reported graphically here, to analyze the model performance at the level of different itinerary types. That is, we can consider itineraries that were direct in the actual data and observe the distribution of number of stops of the predicted itineraries that resulted. By doing this as well for itineraries that were actually one-stop, two-stop, etc. we can get a sense for the biases that our model contains and therefore more ideas for improving the model in the future.

V. Conclusions

In this paper, we use a multinomial logistic regression model to determine the choice of itineraries in the US NAS. By categorizing itineraries into six categories, we found that that in-line and market passengers, distance, average fare, and types of air carriers are capable of estimating these choices fairly well. The empirical findings indicate that the estimated model is capable of predicting fairly well for one- and two-stop itineraries (in excess of 75%) and somewhat well for direct routing (slightly in excess of 35%). The model performs poorly for larger number of hops.

These findings have important implications. First, the estimated model enables mapping of passengers onto itinerary choices, and given types of markets, distance, and carriers, helps us to determine the air travel network. This provides another tool that can be now used to replace arbitrary assumptions relating to the air travel network. Second, and most importantly, the empirical relationship between itinerary choices and passengers, given assumptions regarding fleet choices, allows derivations of frequencies by market segments. This further allows generation of schedules or timetables specific to airports that are driven by, among other things, passenger forecasts which ultimately drive the itinerary choices. This ensures that we can model and simulate the operations of US NAS far more efficiently, corresponding to different passengers demand scenarios than was previously the case.[16]

There are quite a few areas of research that can be pursued in the near future. First, airlines behavior can be explicitly modeled in this framework because many of the itinerary choices result from airlines optimizing some objective functions. Second, we would also like to nest different itinerary choices according to their likely haul or distance of travel leading to likely improvements in results.

REFERENCES

[1]Bhadra, D. and P. A. Texter. "Airline Networks: An Econometric Framework to Analyze Domestic U.S. Air Travel," *Journal of Transportation and Statistics*, 7(1) (2004), 87-102.

[2]ATA, 2003 Air Transport Association. See http://www.airlines.org/public/home/default1.asp (2003) BEA 2003 Department of Commerce: Bureau of Economic Analysis, www.bea.gov.

[3]Airline Monitor 2003. Monthly issues, Ponte Vedra Beach, Florida (2003).

[4]Bhadra, D. "Demand for Air Travel in the United States: Bottom-Up Econometric Estimation and Implications for Forecasts by O&D Pairs," *Journal of Air Transportation*, September, 8(2) (2003), 19-56 .

[5]Bhadra, D. and D. Hechtman. "Determinants of Airport Hubbing in the US: An Econometric Framework," *Public Works Management and Policy*, Vol. 9, No. 1, July, 2004, 26-50.

[6]Shy. Oz. *The Economics of Network Industries*, Cambridge University Press, London. (2001).

[7]Barabasi, Albert-Laszlo. *Linked, The New Science of Networks*, Perseus Books, (2002).

[8]Holmes, Bruce. *Network Theory: A Primer and Questions for Air Transportation System Applications,* NASA internal document, (2004).

[9]McFadden, D. "Conditional Logit Analysis of Qualitative Choice Behavior," in Frontiers in Econometrics, ed. By P. Zarembka, New York: Academic Press, (1974), 105-142.

[10]Hosner, D.W. and S. Lemeshow. *Applied Logistic Regression*. New York: John Wiley & Sons, (1989).

[11]Ben-Akiva, M. and S. R. Lerman. *Discrete Choice Analysis: Theory and Applications to Travel Demand*, The MIT Press, Cambridge, MA, (1984).

[12]Pindyck, R.S. and D.L. Rubinfeld, 1991, *Econometric Models and Econometric Forecasts*, 3rd edition, New York: McGraw-Hill.

[13]SAS/ETS version 8 1993 SAS/ETS Software: Applications Guide 2: Econometric Modeling, Simulation, and Forecasting, Version 6, First Edition, SAS Institute, Cary, NC, (1993).

[14]Allison, Paul D. *Logistic Regression Using the SAS System: Theory and Application*, SAS Institute, (2001).

[15]BTS/DOT Origin and Destination survey (page 15)
U.S. Department of Transportation. O&D Survey Reporting Regulations for Large Air Carriers: Code of Federal Regulations Part 241 (1992), Section 19-7, Office of the Secretary, Washington, DC.

[16]Bhadra, D., J. Gentry, B. Hogan, and M. Wells. "Future Air Traffic Timetable Estimator," *Journal of Aircraft*, 42(2) (2005), 320-328.

Detailed Results: Available Upon Request

The SAS System
The LOGISTIC Procedure

Model Information	
Data Set	WORK.NETWORK2003_SUMAGGHOPS_REGDATA
Response Variable	ItineraryStop
Number of Response Levels	6
Number of Observations	359837
Model	generalized logit
Optimization Technique	Fisher's scoring

Response Profile		
Ordered Value	ItineraryStop	Total Frequency
1	0	9208
2	1	162018
3	2	157703
4	3	27880
5	4	2485
6	5	543

Logits modeled use ItineraryStop=0 as the reference category.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	760670.34	501976.07

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
SC	760724.31	502353.84
-2 Log L	760660.34	501906.07

R-Square	0.5128	Max-rescaled R-Square	0.5832

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	258754.270	30	<.0001
Score	96644.3222	30	<.0001
Wald	67241.1257	30	<.0001

Type III Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
pax_inline	5	45865.6000	<.0001
pax_marketODMkt	5	3859.1460	<.0001
weightedavfare	5	946.6304	<.0001
avgdistance	5	26746.3771	<.0001
networkdummy	5	90.4425	<.0001
LCCdummy	5	506.2525	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter	ItineraryStop	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1	1.3093	0.0701	348.5525	<.0001
Intercept	2	1	1.7519	0.0814	463.6816	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter	ItineraryStop	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	3	1	0.8474	0.1067	63.0220	<.0001
Intercept	4	1	-1.3892	0.2954	22.1168	<.0001
Intercept	5	1	-2.2153	0.4170	28.2161	<.0001
pax_inline	1	1	-0.0129	0.000179	5209.8586	<.0001
pax_inline	2	1	-0.4963	0.00257	37208.0029	<.0001
pax_inline	3	1	-1.8749	0.0215	7620.8869	<.0001
pax_inline	4	1	-2.9525	0.1315	504.2507	<.0001
pax_inline	5	1	-2.6818	0.2411	123.6929	<.0001
pax_marketODMkt	1	1	0.00177	0.000046	1453.6859	<.0001
pax_marketODMkt	2	1	0.00187	0.000046	1625.4476	<.0001
pax_marketODMkt	3	1	0.00192	0.000046	1704.0463	<.0001
pax_marketODMkt	4	1	0.00196	0.000047	1752.2751	<.0001
pax_marketODMkt	5	1	0.00199	0.000048	1699.7947	<.0001
weightedavfare	1	1	0.00616	0.000264	543.7471	<.0001
weightedavfare	2	1	0.00496	0.000267	346.4521	<.0001
weightedavfare	3	1	0.00460	0.000275	280.4416	<.0001
weightedavfare	4	1	0.00538	0.000310	301.8103	<.0001
weightedavfare	5	1	0.00502	0.000445	126.9723	<.0001
avgdistance	1	1	0.000282	0.000023	148.3234	<.0001
avgdistance	2	1	0.00128	0.000024	2904.3104	<.0001
avgdistance	3	1	0.00161	0.000025	4312.3608	<.0001
avgdistance	4	1	0.00176	0.000031	3310.1567	<.0001
avgdistance	5	1	0.00181	0.000048	1442.5482	<.0001
networkdummy	1	1	0.4609	0.0554	69.2106	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter	ItineraryStop	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
networkdummy	2	1	0.4126	0.0685	36.3255	<.0001
networkdummy	3	1	0.6635	0.0942	49.6479	<.0001
networkdummy	4	1	1.0609	0.2601	16.6367	<.0001
networkdummy	5	1	-0.00869	0.3264	0.0007	0.9788
LCCdummy	1	1	-0.7429	0.0516	207.3674	<.0001
LCCdummy	2	1	-1.4307	0.0656	475.9343	<.0001
LCCdummy	3	1	-1.4664	0.0949	238.5372	<.0001
LCCdummy	4	1	-1.5098	0.2783	29.4365	<.0001
LCCdummy	5	1	-2.7836	0.4081	46.5343	<.0001

Odds Ratio Estimates				
Effect	ItineraryStop	Point Estimate	95% Wald Confidence Limits	
pax_inline	1	0.987	0.987	0.988
pax_inline	2	0.609	0.606	0.612
pax_inline	3	0.153	0.147	0.160
pax_inline	4	0.052	0.040	0.068
pax_inline	5	0.068	0.043	0.110
pax_marketODMkt	1	1.002	1.002	1.002
pax_marketODMkt	2	1.002	1.002	1.002
pax_marketODMkt	3	1.002	1.002	1.002
pax_marketODMkt	4	1.002	1.002	1.002
pax_marketODMkt	5	1.002	1.002	1.002
weightedavfare	1	1.006	1.006	1.007
weightedavfare	2	1.005	1.004	1.006

Odds Ratio Estimates				
Effect	ItineraryStop	Point Estimate	95% Wald Confidence Limits	
weightedavfare	3	1.005	1.004	1.005
weightedavfare	4	1.005	1.005	1.006
weightedavfare	5	1.005	1.004	1.006
avgdistance	1	1.000	1.000	1.000
avgdistance	2	1.001	1.001	1.001
avgdistance	3	1.002	1.002	1.002
avgdistance	4	1.002	1.002	1.002
avgdistance	5	1.002	1.002	1.002
networkdummy	1	1.585	1.422	1.767
networkdummy	2	1.511	1.321	1.728
networkdummy	3	1.942	1.614	2.335
networkdummy	4	2.889	1.735	4.810
networkdummy	5	0.991	0.523	1.880
LCCdummy	1	0.476	0.430	0.526
LCCdummy	2	0.239	0.210	0.272
LCCdummy	3	0.231	0.192	0.278
LCCdummy	4	0.221	0.128	0.381
LCCdummy	5	0.062	0.028	0.138