
Total 6,789 Words (= 4,289 Words + 5 Figures + 1 Tables)

Anand Seshadri
Via Department of Civil Engineering
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061
Phone: (540) 231-3363
Fax: (540) 231-7532
aseshadr@vt.edu

Hojong Baik (Corresponding Author)
Via Department of Civil Engineering
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061
Phone: (540) 231-2362
Fax: (540) 231-7352
hbaik@vt.edu

Antonio A. Trani
Via Department of Civil Engineering
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061
Phone: (540) 231-4418
Fax: (540) 231-7352
vuela@vt.edu
Abstract

This paper introduces an Origin-Transfer-Destination (O-T-D) model that predicts the route and segment passenger flows between airport pairs using origin-destination (O-D) passenger demand between airport pairs. The model is developed based on a multinomial logit model and covers 443 nodes (i.e., airports) and about 6000 links (flight segments). To create airline schedules between any airport pair, a network-building module has been developed using the Official Airline Guide (OAG) that provides flight segment information. Flight fares for routes are extracted from Airline Origin and Destination Survey (DB1B). The observed segment passenger flows between airport pairs are also obtained from DB1B. Two different types of logit models are calibrated and then validated by comparing estimated route and segment passengers between airport pairs with the observed route and segment passengers obtained from the DB1B and the Form 41 Segment Data (T-100), respectively.
INTRODUCTION

The National Airspace System is facing increasing congestion leading to large delays in certain hubs. The Federal Aviation Administration (FAA) and other government agencies, for example the Joint Program Development Office (JPDO) require models that are capable of predicting the passenger loads at major hubs. Since the events of September 11, 2001, the airlines have suffered large losses in revenues, passengers and market shares (1). Understanding the passenger response to changes in service variables, for example fare and frequency is of importance to airlines. To meet these practical needs, it is essential to develop a reliable model that estimates segment and route passenger flows between airport pairs.

This paper introduces a model, named as Origin-Transfer-Destination (O-T-D) model that predicts route and segment passengers by assigning Origin-Destination (O-D) passengers between airport pairs to the various routes that connect the airport pair. Consequently, O-T-D model provides air passenger demand at each airport that can be utilized in predicting airport congestions. It also provides load factors on each leg of the airline network that can be utilized in assessing route revenues and in determining route operating plans for aircraft types, frequencies, etc. The O-T-D problem can be stated as follow:

Given a matrix of Origin-Destination airport passenger demand, estimate the route, segment passenger flows for each airport pair and the enplanements at each airport.

The O-T-D mode requires O-D passenger demand between airport pairs that can be obtained from any socio-economic based forecasting model. One of the models is the Transportation System Analysis Model (TSAM) that is a nationwide intercity transportation demand model developed by the Air Transportation Systems Laboratory at Virginia Tech. The TSAM is based on four-step modeling transportation process that considers multimode including automobile and commercial airline, and consequently provides O-D air passenger demand between airport pairs. Figure 1 shows the modeling structure of TSAM together with input data sets. From the modeling viewpoint O-T-D analysis is a part of network assignment module that assigns air passengers to air routes. (See the accompanying paper (2) for more details.)

The O-T-D model introduced in this paper covers all the airports for which commercial service exists in the continental United States in the Official Airline Guide except following airports: 1) seaplane bases and heliports, and 2) airport pairs with low service frequency (less than 5 days per week). Year 2004 is selected as the base year for data acquisition since the events of September 11, 2001 had a considerable effect on airline network and fare structures and the most recent information available when the research was conducted was year 2004.

In this paper is organized as follow, the objective of the problem and the methodology adopted is described first, and the previous work done on the problem and their shortcomings are reviewed. We then explain the methodology adopted in detail with the results of analysis. Finally, we discuss possible application and directions for future research.

LITERATURE REVIEW

There have been many studies attempted on the problem of network assignment in ground transportation. However, most of the studies in ground transportation utilize the concept of user equilibrium, which is not applicable for airline networks since the service providers in ground networks (roadways, vehicles) are passive agents unlike the service providers (airlines) in airline networks. However, some studies have been attempted to solve the network assignment problem in airline networks. The main methodologies used are logit models, game theoretic models and fuzzy set based multi-agent models.

The logit model is based on the premise that every decision-maker maximizes his utility when presented with a range of choices. The utility is usually expressed as a linear function of several independent variables. The coefficients for the independent variables are estimated using a maximum
FIGURE 1: The Modeling Structure of Transportation System Analysis Model (TSAM).

likelihood technique. The logit model is computationally the simplest of all the three methodologies. Logit based formulations are employed by Weidner (3), Ghobrial and Kanafani (4) and Hsiao and Hansen (5). Weidner (3) used a nested logit model to solve the problem of passenger route choice. Supply costs, congestion, and other airport characteristics were used to model the route utility. Route distance and capacity saturation were used as surrogates for travel time and connecting delays. Ghobrial and Kanafani (4) used a logit based equilibrium model to determine segment flows for 25 airports in the Southeastern region of the United States. The independent variables considered were travel time, fare, frequency, number of legs and seats. The model initially assumed unit frequency for determining segment demand. A pre-set load factor was then assumed for computing frequency on each leg. The process was repeated until equilibrium was achieved. Hsiao and Hansen (5) used a logit-based equilibrium formulation to model segment demand. The variables used in the model are similar to Weidner (3). The authors used delays instead of capacity...
saturation. Assuming an initial delay, the process was iterated until equilibrium is reached. However, the logit model cannot take into account the entire spectrum of factors in the segment flow problem, for example, airline competition whose dynamics cannot be captured in the logit model and airline bankruptcies, which are difficult to model.

Game Theory models competition between airlines and passengers are passive agents in the game. Game Theory can address the issue of competition between airlines. Game-theoretic formulations were used by Hansen (6), Dobson and Lederer (7) and Adler (8). Hansen (6) used an n-player non-cooperative game to determine airline shares and segment flows. The variables used to model route utility were fare, route circuitry and the maximum and minimum flight frequency on each link of the route. Dobson and Lederer (7) used a game-theoretic framework to schedule flights and calculate segment flows. The route demand is determined by assigning a disutility to each itinerary for that route. The authors then maximize profit for each airline, which is a function of the airline schedule. The profit-maximizing schedule is obtained and therefore, the profit and segment flows. Adler (8) used a game-theoretic framework to model airline competition in Western Europe. The model scope is 20 nodes and four airlines. However, implementing game theory on a large network is quite difficult. Dobson and Lederer (7) state that a parallel computer with several thousand processors is required for implementing game theory on a large (50-node) network. All existing literature on the application of game theory only deal with networks of 10-30 cities which is about an order of magnitude less than the problem dealt with in this analysis.

Agent-based formulations were used by Teodorovic and Kalic (9). The authors used a set of rules based on fuzzy set to incorporate passenger choice in airline networks. However, precise formulation of the rules for a large-network can be quite involved because of the number of interacting phenomena. Teodorovic and Kalic implemented their model on a network consisting of 13 nodes in the Southeastern part of the United States. The methodology finally adopted was the logit model since it was not clear that the other methodologies could be implemented for a large-scale problem and the logit model offered computational simplicity, which was the quickest method to achieve an acceptable solution from the three methods.

**METHODOLOGY**

It should be noted that previous studies mainly focused on a sub-set of the entire commercial airport set. Game theoretic models in particular are restricted in scope. Most of the game-theoretic models deal with networks of fewer than 20 nodes and have a maximum of three players. However, the National Airspace System (NAS) has 443 airports receiving commercial service and about 10 mainline carriers. It is clear that applying the game-theoretic formulation to model the entire NAS would be computationally very difficult. Most of the studies using logit models use route distance as a surrogate for travel time. However, the relationship between travel time and route distance is not monotonic for commercial airline networks. In addition, most of the studies outlined above derive the airline network topology from the Federal Aviation Administration (FAA) the Airline Origin and Destination Survey (also know as DB1B), which makes the model unusable for future years since the DB1B is not available. The studies reviewed also did not account for seasonal variations in the airline schedule.

The methodology introduced in this paper considers the entire commercial airport set in the Continental United States. The airline network topology was constructed synthetically from the Official Airline Guide, which obviated the need for deriving the network from DB1B. A representative airline schedule was constructed for the entire year thus taking into account schedule variations across quarters. The model described also used real travel times instead of surrogates, thus removing the need for using route distances.

**Model Components**

The model consists of three sub-modules: 1) a flight information extraction module, which extracts a representative yearly schedule from Official Airline Guide (OAG); 2) a flight schedule generation module, which constructs a synthetic network using the output of the flight information extraction module; and 3) a route choice module, which derives the coefficient set for the explanatory variables. The three modules are briefly outlined in the next section.
The Official Airline Guide (OAG) was used to synthesize an effective yearly schedule. However, the airline schedule is not constant for the entire year as the air carriers vary the arrival and departure times to meet seasonal variations in demand. Therefore, an effective yearly schedule was extracted by taking the arrival and departure times of flights that were within 30 minutes of each other. It was assumed that an airline would not schedule flights between the same Origin-Destination pair within 30 minutes of each other.

The flight schedule generation module, as described in the next section was constructed to obviate the need to derive the network topology from the DB1B. This step is necessary since the DB1B yields poor samples for thin markets and a network topology based on the DB1B would not be reliable. Flight schedules between airport pairs are built-up in sequence: first the O-D pairs that can be connected by direct flights are connected, and followed by O-D pairs with two, three and more than three legs as described below.

**One Leg:** The records that had direct flights from an origin to a destination airport are extracted from the OAG. Any record with an unusually long flying time (multi-stop flights) was removed. The average of the remaining flight times were averaged and reported as the travel time from \( i \) (Origin) to \( j \) (Destination). However, for many O-D pairs that have direct service a significant proportion of the passengers used connecting itineraries. Therefore, the two-leg solution was appended with the one-leg network for all O-D pairs having direct service.

**Two Legs:** The list of all the airports that have service from origin airport \((i)\) was obtained. These O-D pairs could not be connected by a direct flight; therefore, the intermediate airports that connect them with two legs were determined. The travel time from Origin \((i)\) to Destination \((j)\) was the sum of the flight time from \(i\) to \(k\), the flight time from \(k\) to \(j\) and the connecting time at airport \(k\). The reported travel time was the average of the travel times of all feasible schedules along all the possible routes.

**Three Legs:** The list of all the large and medium hubs (FAA definition) \((H1, H2)\) to which the origin and destination airport have service were obtained. It was assumed that the large and medium hubs were connected by direct flights, as about 60\% of large and medium hubs are connected by direct flights in 2004. As in the previous cases, the travel time from \(i\) (Origin) to \(j\) (Destination) was the sum of the flight times from \(i\) to \(H1\), \(H1\) to \(H2\), \(H2\) to \(j\) and the connecting times at \(H1\) and \(H2\). Large connecting times (large travel times) were eliminated by a rolling average method.

**Multi (4 or more) Legs:** There were a few Origin-Destination pairs, which were unconnected even after using three legs. The potential number of combinations for any itinerary that consisted of more than three legs was too large to derive from first principles. For these airports, a recursive approach that utilized the previous solution was adopted. The set of intermediate airports that had service to destination airport were determined. If any of these intermediate airports had an itinerary to the origin airport, the final itinerary was the addition of another flight that connected the intermediate and destination airports to the initial itinerary. Itineraries with long travel times were eliminated by a rolling average method.

In the route choice module, a multinomial logit model is calibrated. The variables chosen for the logit model are:

1) **Relative Travel Time:** The relative travel time was the quotient of the actual travel time on the route and the average travel time for the O-D pair. Travel times were normalized to avoid scale biases.

2) **Relative Travel Cost:** The relative fare was the quotient of the actual fare on the route and the average fare for the O-D pair. As in the case of travel time, fares were normalized to avoid scale biases.
3) **Number of itineraries (Frequency):** The number of itineraries offered on a route was the number of connecting flight combinations that had reasonable connecting times. The maximum and minimum connecting times were taken as 2.5 hours and 40 minutes respectively. In case no O-D itineraries could be found with connecting times less than 2.5 hours, the maximum connecting time was taken as 3 hours greater than the minimum connecting time.

3) **Seats:** The number of seats for an itinerary was taken as the number of seats offered on the first leg of the itinerary.

4) **Legs:** The number of legs was included to account for additional disutility incurred for connecting itineraries (with potential for missed connections or lost baggage).

5) **Dummy Variables:** Dummy variables were introduced for airports, which had significant transfer disutility due to extraneous factors.

The resulting utility function used in the model can be expressed as:

\[ U_{ijk} = \alpha_1 * RTT_{ijk} + \alpha_2 * RTC_{ijk} + \alpha_3 * Freq_{ijk} + \alpha_4 * S_{ijk} + \alpha_5 * \eta_{ijk} + \beta_l \]  

where,

- \( i, j, k \): origin, destination and intermediate airport respectively,
- \( RTT_{ijk} \): relative route travel time,
- \( RTC_{ijk} \): relative route fare,
- \( Freq_{ijk} \): number of feasible itineraries on the route,
- \( S_{ijk} \): number of seats offered by the airline from airport \( i \) to airport \( k \),
- \( \eta_{ijk} \): number of legs on the itinerary, and
- \( \beta_l \): dummy variable for some airports \( l \).

The route passenger flow can be expressed as:

\[ D_{ijk} = D_j * \frac{\exp(U_{ijk})}{\sum_k \exp(U_{ijk})} \]  

where,

- \( D_j \): number of air passengers from airport \( i \) to \( j \), and
- \( D_{ijk} \): number of air passengers flying from airport \( i \) to \( j \) through airport \( k \).

Various versions of the logit model have been used in literature to model consumer choice behavior. The three most common logit formulations are the standard multinomial logit, nested logit and mixed logit. The logit used in this model was the multinomial logit, since the multinomial logit provided the best balance between computational effort and predictive power. The nested logit did not produce any appreciable increase in model accuracy. The mixed logit model is the most flexible version of the logit (10) which overcomes the problem of random taste variation across the sample set. However, the mixed logit could not be implemented in the present version of the model because of computational constraints.

Several issues had to be addressed in the model. The first issue was the lack of adequate fare observations for certain routes in the network. The DB1B is only a 10% sample of all the tickets sold by the airlines in the United States. The DB1B is therefore a weak data sample for smaller airports and thinner routes. However, since fare is a critical explanatory variable for the utility function in the logit model, a method to derive a credible representative route fare was needed. This was accomplished using a non-linear regression in which the dependent variable was fare-per-mile and the independent variable was the route distance. The equation was used in-lieu of the actual fare when adequate fare data did not exist for the route. The second issue was the overestimation and underestimation of enplaning passengers at certain large-hubs by more than 10%. Specifically, the resulting model overestimated the enplanements at ORD by about 16% and underestimated the enplanements at CLT by about 11%. To overcome this problem, the model incorporated dummy variables for the airports, to minimize prediction errors. The entire model calibration procedure is depicted as a flowchart in Figure 2.
RESULTS

Model Coefficients

The logit model was calibrated using Statistical Analysis Software (11) and resulting coefficients together with statistics for the explanatory variables are given in Table 1. (The t-statistics indicate that all variables are significant with a 95% level of confidence.) The coefficients for travel time and travel cost also enable the calculation of value-of-time for travelers. For example, for a flight from ROA, VA, to ORD, IL, the average (arithmetic mean) travel time is about 4.23 hours and the average fare paid is about $160. It can be inferred that, an increase in travel time of 20 minutes would cause a decrease in utility of \(-2.252\times\frac{1}{3}\times\frac{1}{4.23} = -0.1774\). To offset this decrease in utility, the fares need to be decreased by \frac{0.1774}{0.9588}\times160 = $29.60. The value of time for the traveler is therefore, 29.60\times3 or about $90 per hour.

Ghobrial and Kanafani (4) derived a value-of-time of about $51 per hour in 1995. Accounting for inflation, this is equivalent to about $65 in 2004 dollars. The difference could be explained by the fact that Ghobrial and Kanafani (4) considered a limited geographical extent and neglected regions with a high average income, for example New York and California. The authors also did not normalize the travel time and fare values. Normalization of travel times and fares ensures that value-of-time is a function of the total travel time. In general, shorter trips tend to have a higher value-of-time than longer ones. It can be concluded that the lack of normalization of travel times and fares would yield a bias in the estimation of the value-of-time.

The coefficient for frequency is positive, as increased frequency leads to decreased schedule delay for the traveler and therefore, an increase in utility. The coefficients for seats is also positive since a greater number of seats offered produce a corresponding decrease in the probability of the passenger being denied a seat. The magnitude of the coefficient for seats is small as the number of seats offered per year on a leg is large.

The coefficient for the number of legs is negative as passengers typically avoid connecting flights. This is supported by the fact that in the year 2004, 66% of air travelers took direct flights (5). From Table 1 it can be inferred from the R-square values that models are quite accurate in O-D airport pairs where both
direct and connecting itineraries are provided. On the other hand, the model does not produce very good results for O-D pairs which are linked by connecting itineraries alone. This phenomenon can be explained by the fact that a significant proportion of the first set consists of travelers between large and medium hubs where the flow magnitude is large enough for the model to produce good results. On the other hand, the second set consists of travelers between small airports, where the flow is small and local effects play a significant role.

As pointed out, the model leads to significant errors in the prediction of enplaning passengers for CLT and ORD (16% and 11% respectively). However, the introduction of dummy variables reduced the errors to 6% and 3% respectively. The model, which incorporates dummy variables, has a better log-likelihood ratio than the alternative model, which has no dummy variables. It can be inferred that the use of dummy variables increases the predictive power of the model.

### TABLE 1: Resulting Coefficients of Logit Models

<table>
<thead>
<tr>
<th>Variable</th>
<th>Logit model without of dummy variables</th>
<th>Logit model with dummy variables</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model for O-D pairs with direct and connecting itineraries</td>
<td>Model for O-D pairs with connecting itineraries</td>
</tr>
<tr>
<td>Relative Travel Time</td>
<td>-2.252 (-121.93)</td>
<td>-3.676 (-280.65)</td>
</tr>
<tr>
<td>Relative Fare</td>
<td>-0.9588 (-212.37)</td>
<td>-1.175 (-270.85)</td>
</tr>
<tr>
<td>Frequency</td>
<td>0.0873 (224.15)</td>
<td>0.1532 (361.01)</td>
</tr>
<tr>
<td>Seats</td>
<td>1.8743*10^-6 (109.92)</td>
<td>1.6114*10^-6 (106.78)</td>
</tr>
<tr>
<td>Number of Legs</td>
<td>-2.8730 (-315.28)</td>
<td>-</td>
</tr>
<tr>
<td>Dummy for ORD</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Dummy for CLT</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Log-likelihood ratio</td>
<td>0.681</td>
<td>0.153</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.9950</td>
<td>0.3856</td>
</tr>
</tbody>
</table>

(): t-statistics.

### Validation

Figure 3 compares estimated passenger flows with observed segment passenger flows obtained from T-100 data set. From the figure, it can be seen that the model reproduces the observed flows with good accuracy. From Figure 4, it can be inferred that the model is a good representation for passenger behavior for the large and medium hubs. From Figure 5 it can be seen that the introduction of dummy variables for CLT and ORD reduced the percentage error in enplaning passengers significantly. The two lines in Figure 5 represent points, which deviate from the ideal model (45-degree line) by more than 10 percent.
FIGURE 3: Observed vs. Estimated Segment Flows.
FIGURE 4: Observed vs. Estimated Route Passengers
A) Using the Model without Dummy Variables

B) Using the Model with Dummy Variables

FIGURE 5: Observed vs. Estimated enplanements from the Model.
CONCLUSIONS AND RECOMMENDATIONS

An O-T-D Model introduced in this paper uses Travel Time, Fares, Frequency, Seats and Number of Leg for 443 airports and a network based on the OAG has been developed. The model derives good results (Craig-Uhler R-sq = 0.9) for segment flows between large and medium hubs, which constitute about 75% of the segment flows in the NAS. However, the model does not yield very good results (Craig-Uhler R-sq = 0.4) from small and non-hubs. The DB1B is not a good source of information for the smaller airports, and this could be one of the reasons for the model's performance for the small hubs.

One of the shortcomings of any standard passenger choice dataset is that it has information only on the choice made by the passengers but has no information on the alternate choice set. In this model, a static alternate choice set (average route fares) was assumed. This assumption might not be realistic since fares fluctuate widely. Picking random fare values from the fare distribution might yield better results. The model could possibly be enhanced by attempting a mixed logit for the O-D pairs with having only connecting itineraries. Future versions of the model could include number of legs, ticket types (coach and first classes) and airport-specific variables such as hub type and airport delays which would obviate the need for dummy variables. Further methods to be investigated are game theory and mathematical programming.

Nevertheless, the introduced model can be used in predicting enplanements by origin and transfer passengers at each airport which is essential information for airport delay analysis. This model may also be imbedded a nationwide intercity transportation demand model as a network assignment module. The procedure has actually been implemented in Virginia Tech Transportation System Analysis Model (TSAM).

REFERENCES