Estimation of Airline Itinerary Choice Models Using Disaggregate Ticket Data

Laurie Garrow

with Matthew Higgins, GA Tech

Virginie Lurkin* Michael Schyns, University of Liege

Northwestern University

Evanston, IL

October, 2015

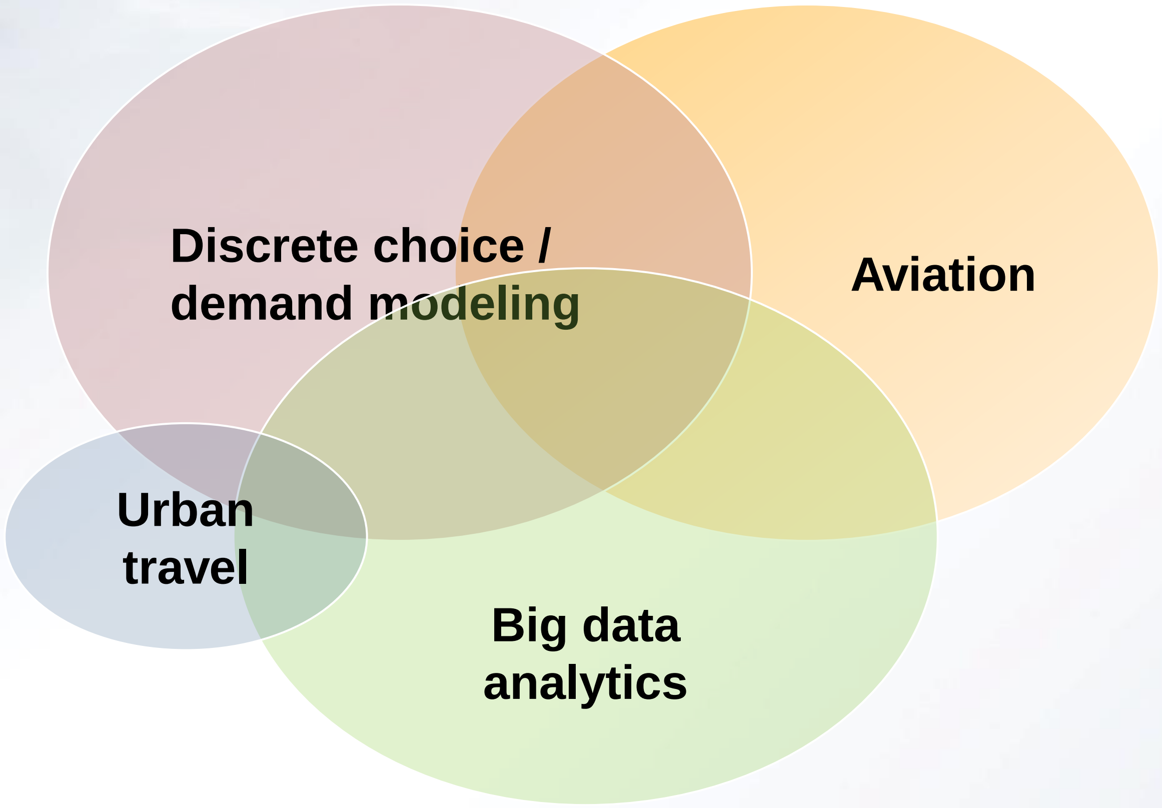
HEC ULg

Ecole de Gestion de l'Université de Liège

Georgia Institute of Technology

fnrs LA LIBERTÉ DE CHERCHER

My Research


Discrete choice /
demand modeling
Aviation
Urban
travel
Big data
analytics

My Background

Leadership:

President, AGIFORS
Former President, INFORMS Transportation Science and Logistics
Former Board Member, INFORMS Revenue Management and Pricing
Former Chair, INFORMS Aviation Applications Section, 2011-12
Former Co-Chair, Emerging Methods, TRB Travel Demand, 2007-12

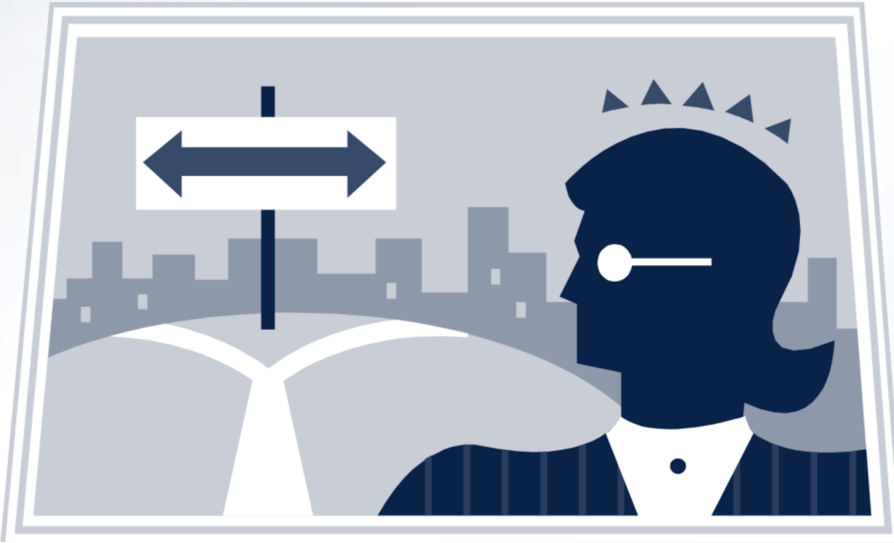
Teaching:

Discrete choice analysis, demand modeling (CEE graduate)
Advanced statistical programing (CEE graduate)
Revenue management and pricing (MBA)
Civil engineering systems, probability (CEE undergraduate)

Ongoing Industry and Government Collaborations:

Boeing, American, Sabre, Airline Reporting Company, …
Parsons Brinkerhoff, AirSage, Epsilon, Georgia DOT

Research Philosophy



Georgia Institute of Technology

Research Portfolio

Travel Demand Modeling

Georgia Tech | School of Civil and Environmental Engineering

HOMEPAGE | RESEARCH | 2013 INFORMS RMP | PUBS | BIO | STUDENTS | TEACHING | OUTREACH | SCHOLARSHIPS

CURRENT RESEARCH 1 2 3 4 5

Have you ever chosen a flight based on whether an aisle or middle seat was available? GT researchers are developing itinerary choice models based on online pricing and seat map information. Learn more

Welcome to Dr. Laurie Garrow's Research Group

Understanding demand for products and services is an integral part of many fields. Within the airline industry, there has been increased interest in modeling demand as the collection of individuals' decisions using discrete choice models. Within the urban travel demand field, there has been increasing interest in using non-traditional data sources, such as credit reporting data, to model transportation decisions. Dr. Garrow and her research group are working on ways to use online data and non-traditional data sources to enhance our understanding of traveler behavior, particularly within the airline industry.

http://garrowlab.ce.gatech.edu/

Georgia Institute of Technology

Outline

1. Review of network planning models and problem motivation
2. Research objectives
3. Data
4. Methodology
5. Results
6. Future research

Georgia Institute of Technology

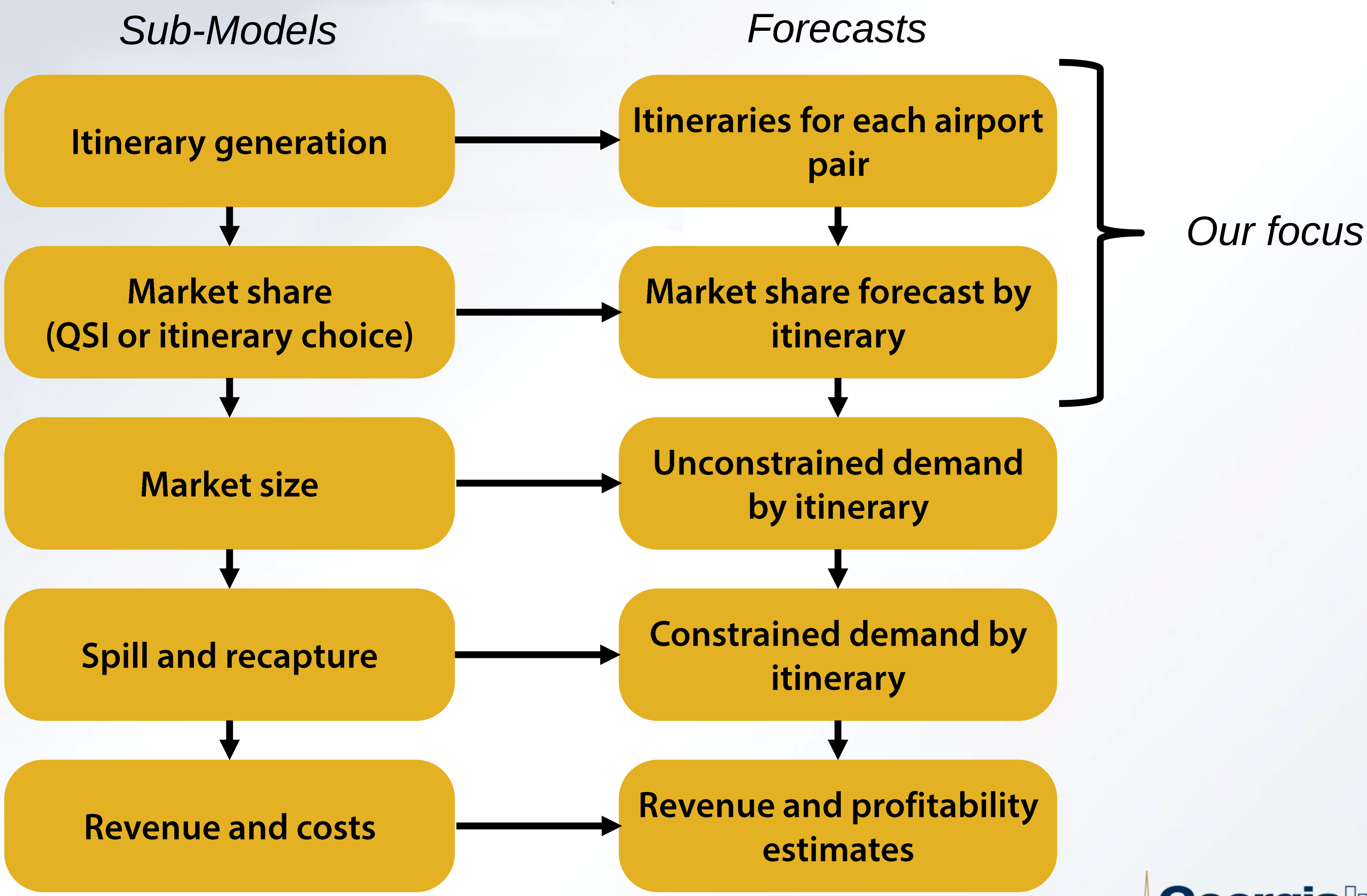
Network Planning Models

1. Are used to forecast **schedule profitability**

2. Support many decisions such as **where** to fly, **when** to fly, **what equipment** to use/purchase, which airlines/flights to **codeshare** with, etc.

3. Contain multiple **sub-modules**

Network Planning Sub-Models



Reference: Garrow 2010, Figure 7.1.

Quality of Service Index (QSI)

QSI models developed in 1957 and can be thought of in terms of **ratios**

$$QSI_i = (\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4), \text{ or}$$

$$QSI_i = (\beta_1 X_1)(\beta_2 X_2)(\beta_3 X_3)(\beta_4 X_4).$$

$$S_i = \frac{QSI_i}{\sum_{j \in J} QSI_j}$$

where

β are preference weights

X are quality measures (*e.g.*, # stops, fare, carrier, equipment type)

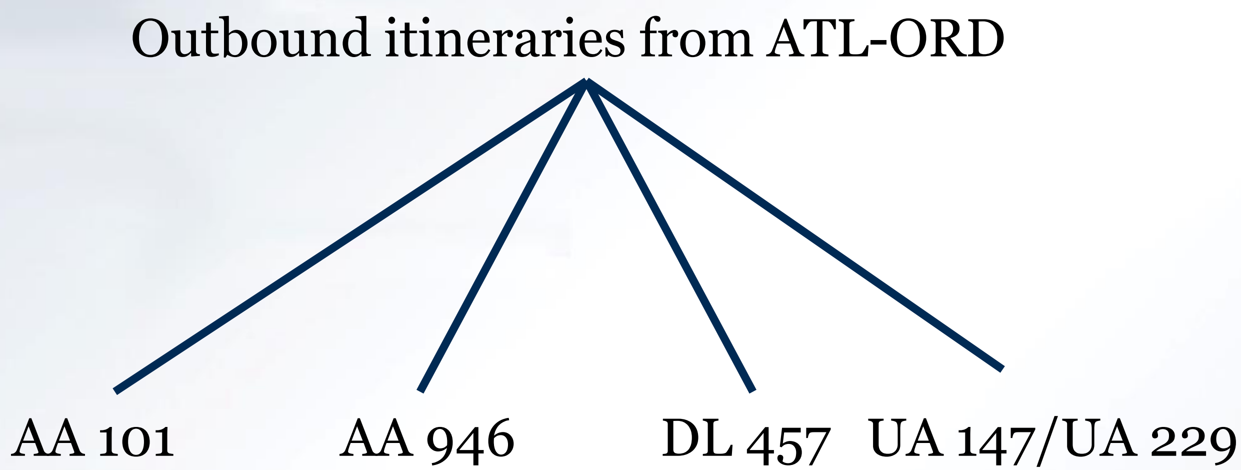
i,j are indices for itineraries

Limitations

- β are usually **not estimated**
- QSI models don't incorporate **competitive factors**

Georgia Institute of Technology

Itinerary Choice Model



$$U_i = V_i + \varepsilon_i$$

$$V_i = \alpha_i + \beta_1 \text{cost} + \beta_2 \text{time} + \ldots$$

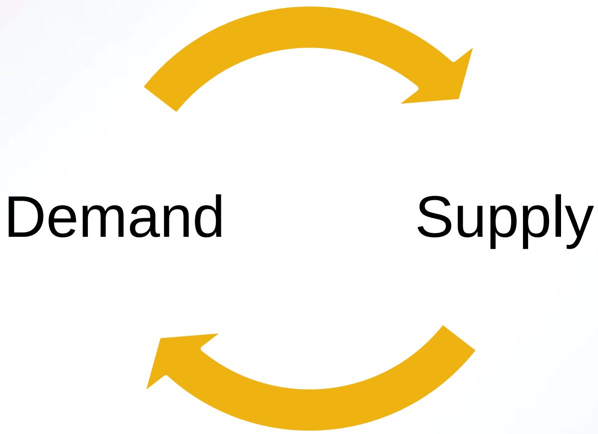
$$P_i = \frac{e^{V_i}}{\sum_j e^{V_j}}$$

Factors Influencing Itinerary Choice

The Fundamental Problem

100 pax	40 pax	120 pax
$500	**$120**	**$700**

$$\text{demand} = \boldsymbol{\beta} \times \text{price} + \ \ldots\ + \boldsymbol{\varepsilon}$$



$$\boldsymbol{\beta} = \mathbf{+0.14}$$

Outline

Georgia Institute of Technology

Research Objectives

1. Use ticketing data from Airlines Reporting Corporation (ARC) to **generate itineraries** and **estimate** choice models

2. Estimate models that account for **price endogeneity**

Georgia Institute of Technology

Outline

Georgia Institute of Technology

Data

1. ARC ticketing data for May 2013 departures
2. Restrict analysis to Continental U.S. markets
3. Include simple one-way and round-trip tickets with at most 2 connections
4. Eliminated tickets with fares < $50 (employee and frequent flyers) or in top 0.1% (charter flights)
5. More than 9.6 million tickets meet these criteria

Georgia Institute of Technology

Explanatory Variables

Carrier characteristics

- Carrier preferences
- **Marketing relationships**
- **Airport share**

Itinerary characteristics

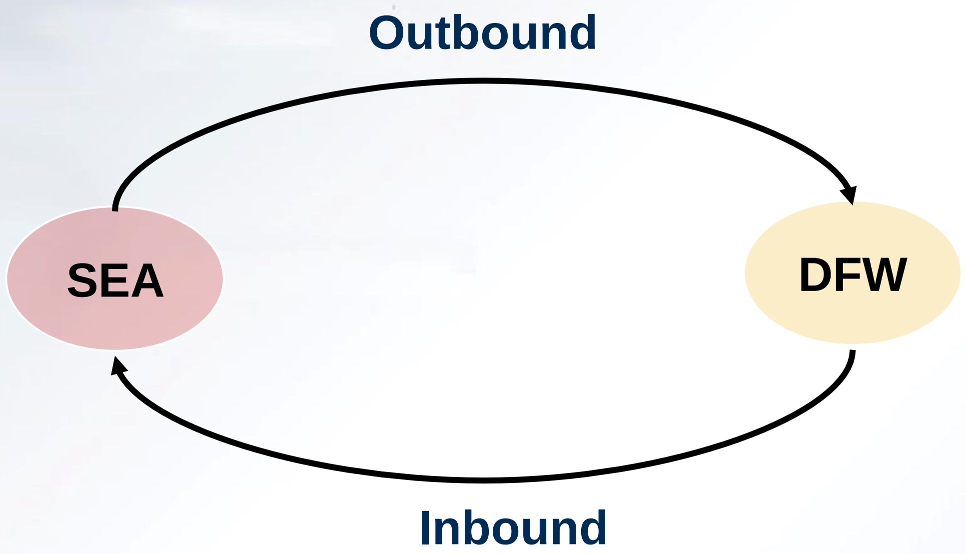
- **Price**
- **Departure time of day preferences**
- Elapsed time
- Number of connections
- Short connection (<60 minutes) indicator
- Direct flight indicator

Georgia Institute of Technology

Marketing Relationships

- **Online** = Same marketing and operating carrier all legs
- **Codeshare** = Same marketing carrier, different operating carrier
- **Interline** = Different marketing carriers, different operating carrier

Airport Share



$$Share_k^{OB} = \frac{\#\ weekly\ flights_k^{ORG}}{\sum_{k=1}^{K} \#\ weekly\ flights_k^{ORG}}, k = operating\ carrier$$

$$Share_k^{IB} = \frac{\#\ weekly\ flights_k^{DST}}{\sum_{k=1}^{K} \#\ weekly\ flights_k^{DST}}$$

Georgia Institute of Technology

Price

"Business" Prices	"Leisure" Prices
Average price for **First, Business,** and **Unrestricted Coach** fares	Average price for **Restricted Coach** and **Other** fares

- Average is taken by **origin, destination, carrier and level of service (NS, 1 CNX, 2 CNX)**
- Assume outbound (or inbound) price = total price/2
- Exclude taxes

Georgia Institute of Technology

Departure Time of Day

1

Departure time preferences vary by

- ✓ **Length of haul**
- ✓ **Direction of travel**
- ✓ **Number of time zones**
- ✓ **Day of week**
- ✓ **Itinerary type (OW, OB, IB)**

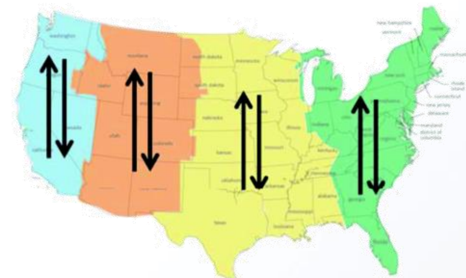
2

Continuous time of day preference formulation is preferred over discrete formulation to avoid counter-intuitive forecasts

10 Time of Day Classifications

Same time zone, < 600 miles

Same time zone, ≥ 600 miles



1 time zone westbound, < 600 miles

1 time zone westbound, ≥ 600 miles

For each classification, estimate separate time of day preferences for **outbound**, **inbound** and **one-way** itineraries and **day of week**

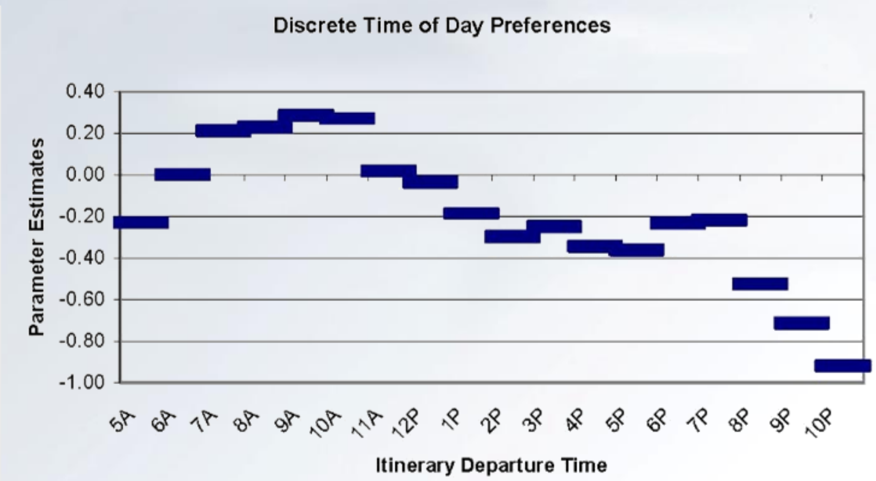
Georgia Institute of Technology

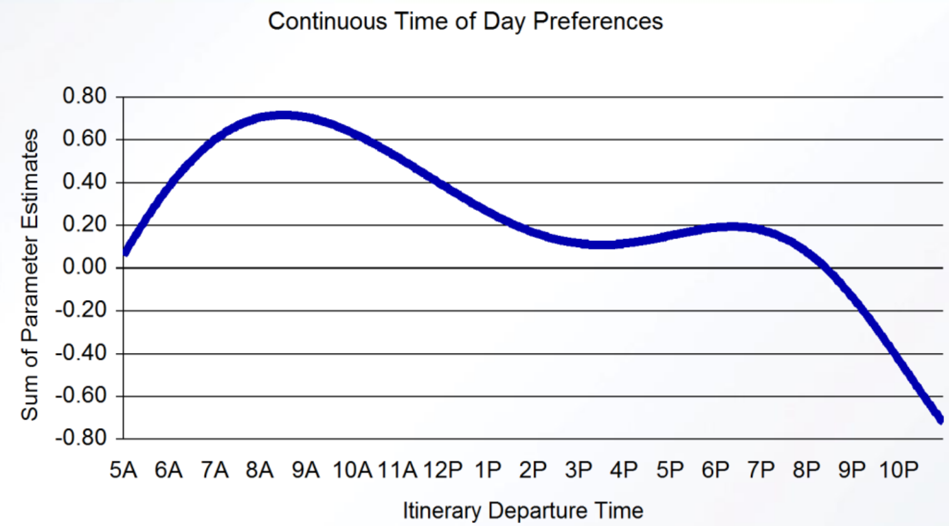
Descriptive Statistics

Segment	Distance			Choice Sets				
	Min	Mean	Max	# OD	Min Alts	Avg Alts	Max Alts	# Pax
Same TZ ≤ 600	67	419	600	3923	2	19	81	1,995,096
Same TZ > 600	601	855	1534	3034	2	25	107	1,599,528
1 TZ EB ≤ 600	118	463	600	766	2	18	69	284,983
1 TZ EB > 600	601	995	1925	3223	2	25	123	1,283,187
1 TZ WB ≤ 600	118	463	600	755	2	18	66	286,818
1 TZ WB > 600	601	994	1925	3251	2	24	132	1,296,951
2 TZ EB	643	1596	2451	1573	2	30	115	641,831
2 TZ WB	643	1597	2451	1541	2	28	109	642,802
3 TZ EB	1578	2229	2774	1074	2	43	172	653,091
3 TZ WB	1575	2227	2774	1059	2	41	164	650,062

Georgia Institute of Technology

Continuous Time of Day





$$Continuous\ time_{cmd} = \beta_{1cmd} sin\left(\frac{2\pi t}{1440}\right) + \beta_{2cmd} cos\left(\frac{2\pi t}{1440}\right) + \beta_{3cmd} sin\left(\frac{4\pi t}{1440}\right) + \beta_{4cmd} cos\left(\frac{4\pi t}{1440}\right) + \beta_{5cmd} sin\left(\frac{6\pi t}{1440}\right) + \beta_{6cmd} cos\left(\frac{6\pi t}{1440}\right)$$

where

$c = time\ of\ day\ classification\ 1,\ldots 10$

$m = outbound, inbound, oneway$

$d = day\ of\ week\ 1, \ldots 7$

$t = departure\ time\ in\ minutes\ past\ midnight$

$1440 = number\ of\ minutes\ in\ a\ day$

Reference: Koppelman, Coldren, and Parker (2008).

Data Representativeness

Carrier	ARC Data	DB1B Market Data
DL	29.5%	23.4%
UA	22.9%	17.1%
US	18.4%	10.0%
AA	17.5%	19.0%
AS	3.3%	4.2%
B6	3.2%	3.0%
F9	2.2%	1.7%
FL	1.4%	2.8%
VX	1.3%	0.9%
SY	0.3%	0.2%
WN	0.0%	17.7%
Total	**100%**	**100%**

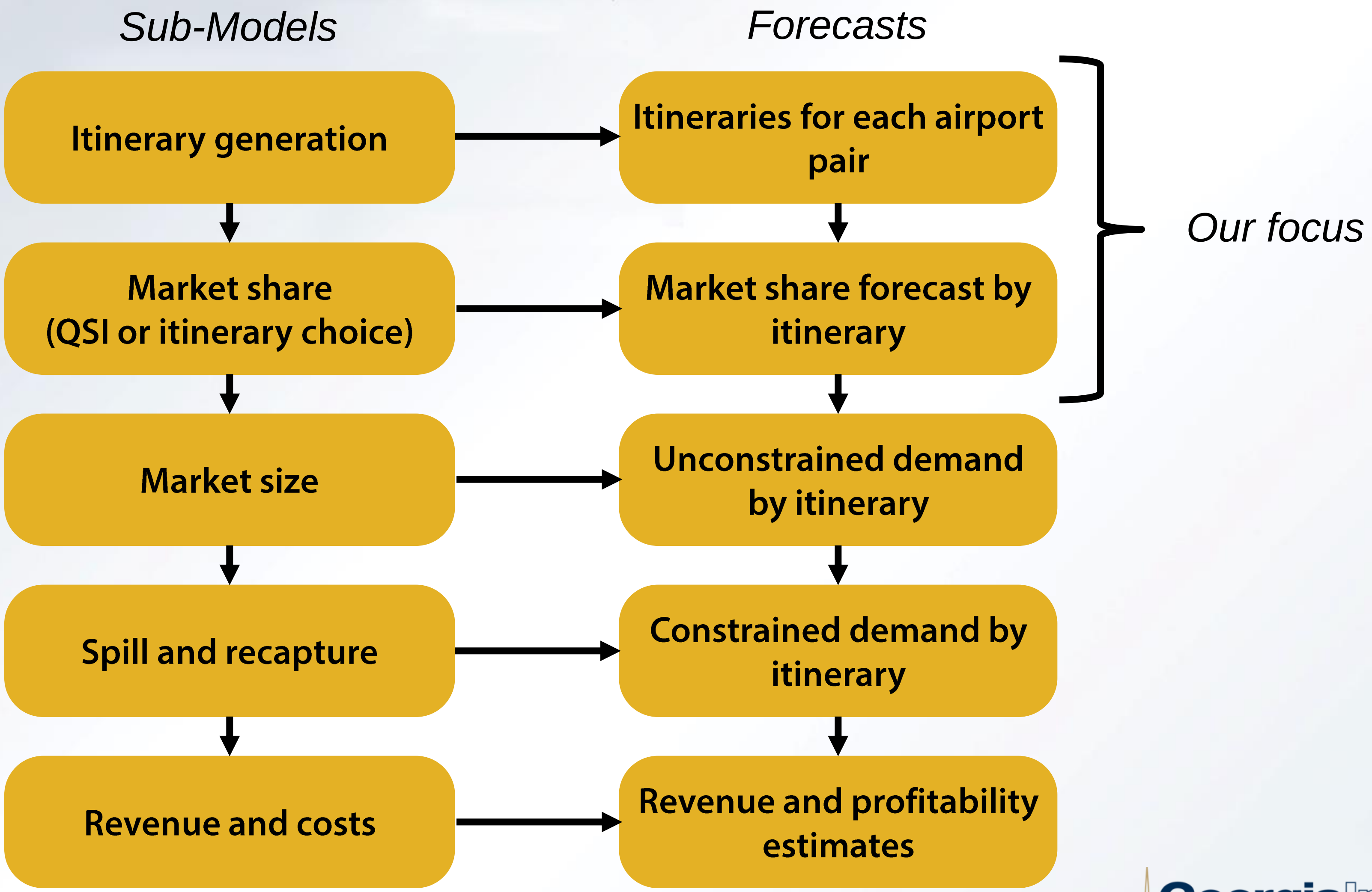
Georgia Institute of Technology

Outline

1. Review of network planning models and problem motivation
2. Research objectives
3. Data
4. Methodology
5. Results
6. Future research

Georgia Institute of Technology

Network Planning Sub-Models



Reference: Garrow 2010, Figure 7.1.

Define Choice Sets

1. Construct **choice sets** for **each OD city pair that departs on day of week *d***

2. Create a **representative weekly schedule** as the Monday after the 9th of the month [May 13 – May 19, 2013]

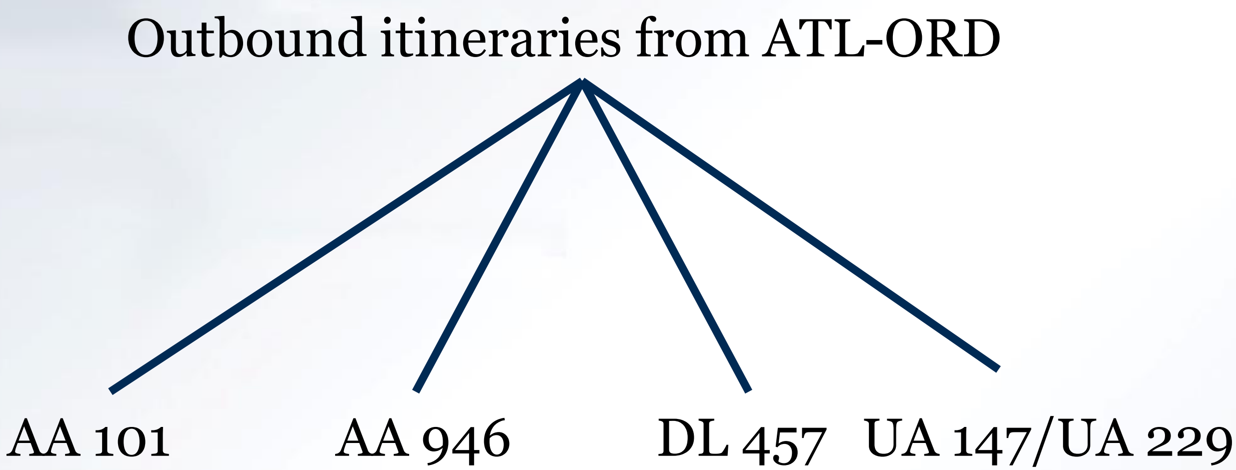
3. Define a unique itinerary by **org_l, dst_l, op $carr_l$, op flt num_l, dept dow_l** for legs $l=1,2,3$

4. Map all demand to representative schedule/unique itinerary

5. Eliminate choice sets with demand < 30 pax/month

Mapping process is **98%** accurate for all variables and screening rule changes MNL parameter estimates by **4.4%**

Itinerary Choice Model



$$U_i = V_i + \varepsilon_i$$

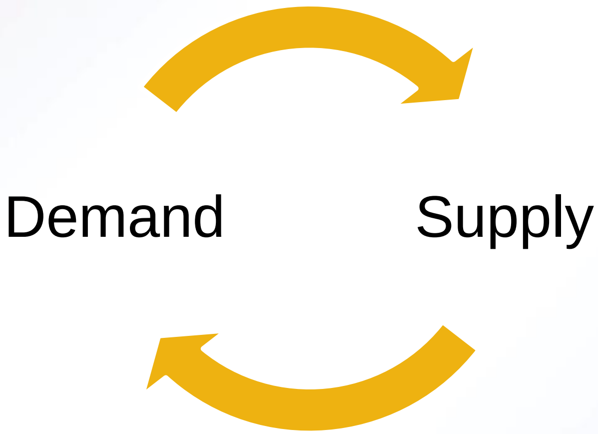
$$V_i = \alpha_i + \beta_1 \text{cost} + \beta_2 \text{time} + \ldots$$

$$P_i = \frac{e^{V_i}}{\sum_j e^{V_j}}$$

The Fundamental Problem

100 pax **$500** | 40 pax **$120** | 120 pax **$700**

$$\text{demand} = \boldsymbol{\beta} \times \text{price} + \ \dots \ + \boldsymbol{\varepsilon}$$



$$\boldsymbol{\beta} = \mathbf{+0.14}$$

Georgia Institute of Technology

The Fundamental Solution

1. Multiple approaches for correcting **price endogeneity**

2. We will focus on two-stage control function method that uses **instruments**

The Basic Idea of Control Function

Validity tests ("are instruments valid?")

1. Instruments should be **correlated with price**
2. Instruments should **not be correlated with choice**

Georgia Institute of Technology

Two-Stage Control Function Method

Stage 1: Linear Regression

$$price = \alpha_0 + \alpha_1 \text{ sin2pi\_MO\_OW\_S1} + \ldots + \alpha_{1260} \text{ cos6pi\_SU\_IB\_S10} + \ldots + \alpha_{1276} \text{ interline}$$

$$+ \alpha_{1277} \text{ IV1} + \alpha_{1278} \text{ IV2} + \mu$$

Endogenous variable

Instruments

Exogenous Variables

$$\hat{\gamma} = price - \widehat{price}$$

Save residuals

Stage 2: Discrete Choice Model

$$V = \alpha_1 \text{ sin2pi\_MO\_OW\_S1} + \ldots + \alpha_{1269} \text{ price} + \ldots + \alpha_{1278} \text{ interline}$$

$$+ \alpha_{1279} \hat{\gamma} + \varepsilon$$

Georgia Institute of Technology

Statistical Test 1 – Endogeneity?

Stage 2: Discrete Choice Model

$$V = \alpha_1 \text{ sin2pi\_MO\_OW\_S1} + \ldots. + \alpha_{1269} \text{ price } + \ldots + \alpha_{1278} \text{ interline}$$

$$+ \underline{\alpha_{1279}}\ \boxed{\hat{\gamma}} + \varepsilon$$

Use **t-test to see if α_{1279}** is significant (if significant, price endogeneity is present)

Statistical Test 2 – Instruments Valid?

Estimate Two Discrete Choice Models

$$V = \alpha_1 \text{ sin2pi\_MO\_OW\_S1} + \ldots . + \alpha_{1269} \text{ price } + \ldots + \alpha_{1278} \text{ interline}$$

$$+ \alpha_{1279} \hat{\gamma} + \varepsilon$$

$$V = \alpha_1 \text{ sin2pi\_MO\_OW\_S1} + \ldots . + \alpha_{1269} \text{ price } + \ldots + \alpha_{1278} \text{ interline}$$

$$+ \alpha_{1279} \hat{\gamma} + \alpha_{1280} \text{ IV1} + \upsilon$$

- Use Likelihood-ratio test to **compare difference in log likelihoods** between two models.
- If difference < χ^2_{NR}=3.84 for one instrument, instruments are valid

References for Direct Test: Guevera and Ben-Akiva (2006); Guevara-Cue (2010).

Which Instruments?

1. Cost-shifting variables
2. Price instruments (“Hausman”)
3. Measures of competition and market power (“Stern”)
4. Measures of non-price characteristics of other products (“BLP” for Berry, Levinsohn, and Pakes (1995))

Cost-Shifting Variables

Used for **aggregate-level** demand estimation

Description	Airline Examples
Variables that **impact a product's cost** but that are uncorrelated with demand shocks	Hsaio (2008) uses **route distance** and **unit jet fuel costs**
	Berry and Jia (2009) and Granados, et al. (2012) use a **hub indicator**
	Granados, et al. (2012) and Hotle et al. (2015)use **distance**
	Hotle, et al. (2015) use the portion of consumers arriving to a destination metropolitan area considered to be **business** and the **population** of the origin city

Hausman-Type Price Instruments

Based on economic theory that a firm's price in one city (market) is a function of the **average marginal costs** of a product + **markup** amount due to different willingness to pay across markets.

Description	Airline Examples
Price of the same brand in **other geographic contexts** are used as instruments of the brand in the market of interest.	Gayle (2004) uses airline's **average prices** in all other markets with **similar length of haul**
	Hotle et al. (2015) use the coefficient of variation of the **lowest offered nonstop fares** across competitors for a specific itinerary.

Georgia Institute of Technology

Stern: Competition and Market Power

Argues that the fact a firm sells multiple products is irrelevant to the value customers assign to a product, but is correlated with price and advertising.

Description	Airline Examples
Measures of the **level of market power** by multiproduct firms, and measures of the **level of competition**	Berry and Jia (2009) use **number of all carriers** offering service on a route
	Granados, et al. (2012) use the **Herfindahl index**
	Number of daily nonstop flights in the market operated by the airline of interest and competitor airlines
	Mumbower et al. (2014) use the number of **daily nonstop flights** in the market operated by competitor.
	Hotle et al. (2015) use the number of **monthly seats** flown in market interacted with **days from departure.**

BLP: Non-Price Char of Other Products

Use observed exogenous product characteristics, namely observed product characteristics for a **firm,** values of same product characteristics for **firm's other products**, values of same product characteristics for **competitors' products**.

Description	Airline Examples
Average non-price characteristics of the other products supplied by the same firm in the same market	**Average flight capacity** of other flights operated by the airline of interest in the same market
Average non-price characteristics of the other products supplied by the other firms in the same market	Berry and Jia (2009) use the **% of rival routes that offer direct flights**, the **average distance of rival routes**, and the **number of rival routes**

Our Instruments

1

Average price by carrier in other similar markets ("Hausman")

- Presence or not of an LCC carrier in the market
- Level of service: nonstop versus connection
- Number of major hubs at origin and destination (from 0 to 2)
- Minimum equipment type for itinerary: Widebody/Narrowbody Jet or other
- Business Area (based on Borenstein Business Index)

2

Number of seats by carrier and markets ("Stern")

Georgia Institute of Technology

Outline

1. Review of network planning models and problem motivation
2. Research objectives
3. Data
4. Methodology
5. Results
6. Future research

Georgia Institute of Technology

Results: 10% of the Data Set

# of markets (directional OD pairs)	19, 962
# choice sets (origin, destination, DOW)	93, 209
# passengers	277, 812
# alternatives in a choice set	941,220
- Min	2
- Max	172
- Mean	37.2

Model Fit Statistics	
LL at zero	-2,581,976.9
LL at convergence	-2,479,081.4
Rho-square w.r.t. zero	0.2202

Georgia Institute of Technology

Results: Value of Time

Variable	Before Correction	After Correction
High yield fare ($)	-0.0015	-0.0037
Low yield fare ($)	-0.0043	-0.0065
Elapsed time (min)	-0.0047	-0.0044

VOT	DCA No correction	DCA Control Function
Leisure	$65	$44
Business	$192	$77

A **business** traveler would pay **$77** to save **1 hour of travel**
A **leisure** traveler would pay **$44** to save **1 hour of travel**

Georgia Institute of Technology

Results: Price Elasticities

Model	Mean Elasticity – Business –	Mean Elasticity – Leisure –
DCA, no correction	-0.43	-0.82
DCA, control function	**-1.09**	**-1.22**

An elasticity of -1.22 means that a **10% increase** in leisure fares leads to a **22% decrease in demand**

An elasticity of -1.09 means that a **10% increase** in business fares leads to a **9% decrease in demand**

DCA with no correction is an **inelastic model** while DCA with control function is an **elastic model**

Georgia Institute of Technology

Relation to Other Airline Studies

Study	Level of Aggregation	Elasticity Estimate	Data Source
Gillen et al. (2002)	Market	-0.79 to -1.43 (long-haul)	Meta study
InterVistas (2007)	Route/Market National Pan-National	-1.40 to -1.54 -0.80 to -0.88 -0.60 to -0.66	DB1B
Hsiao (2008)	Market Route	-1.05 to -2.66 -1.76 to -2.97	DB1B
Granados et al. (2012)	Booking channel: Leisure travel Business travel	-1.33 to -2.28 -0.34 to -1.29	Booking data
Mumbower et al. (2015)	Flight	-1.32 to -1.96	Daily online prices and seat maps
This study	**Route/Market**	**Biz: -1.09** **Leis: -1.22**	**Ticketing data**

Georgia Institute of Technology

Results: Itinerary Characteristics

Variable	Before Correction	After Correction
High yield fare ($)	-0.001	-0.004
Low yield fare ($)	-0.004	-0.007
Elapsed time (min)	-0.005	-0.004
Number of connections	-2.049	-2.202
Number of directs	-1.163	-1.199
ORG outbound freq share	0.981	0.971
DST inbound freq share	0.860	0.862
Short connection	-0.091	-0.068
Codeshare	0.486	0.500
Interline	-0.289	-0.216

Strong preference for nonstop itineraries

Directs are preferred over connections

Georgia Institute of Technology

Results: Itinerary Characteristics

Variable	Before Correction	After Correction
High yield fare ($)	-0.001	-0.004
Low yield fare ($)	-0.004	-0.007
Elapsed time (min)	-0.005	-0.004
Number of connections	-2.049	-2.202
Number of directs	-1.163	-1.199
ORG outbound freq share	0.981	0.971
DST inbound freq share	0.860	0.862
Short connection	-0.091	-0.068
Codeshare	0.486	0.500
Interline	-0.289	-0.216

Effect of flight frequency in "home" location

Slightly stronger effect for outbound passengers

Results: Itinerary Characteristics

Variable	Before Correction	After Correction
High yield fare ($)	-0.001	-0.004
Low yield fare ($)	-0.004	-0.007
Elapsed time (min)	-0.005	-0.004
Number of connections	-2.049	-2.202
Number of directs	-1.163	-1.199
ORG outbound freq share	0.981	0.971
DST inbound freq share	0.860	0.862
Short connection	-0.091	-0.068
Codeshare	0.486	0.500
Interline	-0.289	-0.216

Customers avoid short connections

But effect is not strong – for domestic connections

Georgia Institute of Technology

Results: Itinerary Characteristics

Variable	Before Correction	After Correction
High yield fare ($)	-0.001	-0.004
Low yield fare ($)	-0.004	-0.007
Elapsed time (min)	-0.005	-0.004
Number of connections	-2.049	-2.202
Number of directs	-1.163	-1.199
ORG outbound freq share	0.981	0.971
DST inbound freq share	0.860	0.862
Short connection	-0.091	-0.068
Codeshare	0.486	0.500
Interline	-0.289	-0.216

Code share itineraries **selected more often** than online itineraries

Online and codeshare itineraries are **preferred to interline** itineraries

Georgia Institute of Technology

Time of Day Results: Same TZ > 600 miles

Georgia Institute of Technology

Outline

1. Review of network planning models and problem motivation
2. Research objectives
3. Data
4. Methodology
5. Results
6. Future research

Georgia Institute of Technology

Future Work

1. Estimate **advanced discrete choice models** that incorporate competitive characteristics

2. Extend analysis to **BLP methods** to account for missing data and customer characteristics

3. Apply BLP methods to **merger and acquisition analysis** to isolate "how much" price increase post-merger is due to better product offerings

4. Ideally, work with an airline to implement discrete choice model and **evaluate forecasting benefits** of price formulation

Contributions

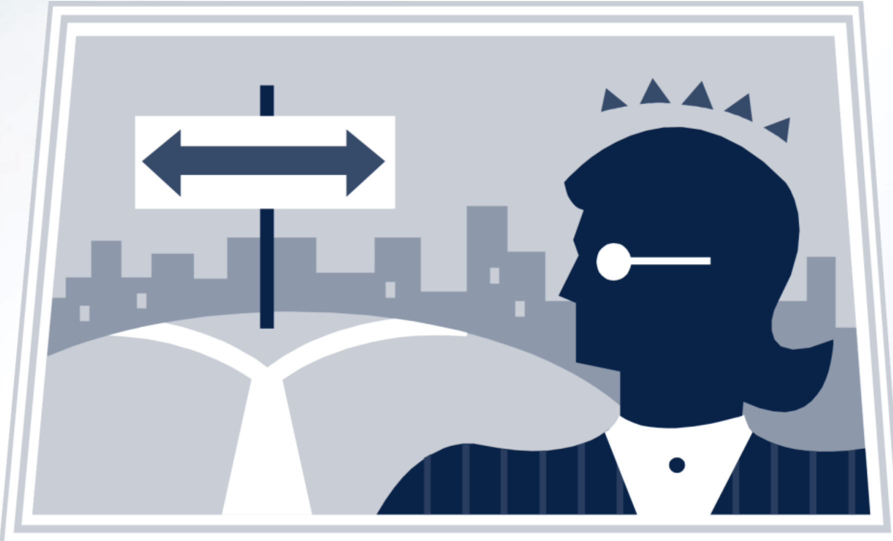
1 First estimates of **itinerary-level** price elasticities based on **detailed ticketing data**

2 Offer a set of **valid instruments** that can be used in future studies of air travel demand

3 Estimate detailed **time of day preferences** that vary as a function of distance, direction of travel (e.g., EW, WE, NS), number of time zones travelled, and itinerary segment (outbound, inbound, one-way)

4 Developed a framework that can be extended to **BLP methods** to correct for missing data and add customer characteristics

Research Philosophy



Acknowledgements

References

1. Berry, S. and Jia, P. 2010. Tracing the woes: An empirical analysis of the airline industry. *American Economic Journal: Microeconomics*, 2 (3), 1-43.
2. Berry, S., Levinsohn, J. and Pakes, A. 1995. Automobile prices in market equilibrium. *Econometrica*, 63 (4), 841-890.
3. Garrow, L.A. 2010. *Discrete Choice Modelling and Air Travel Demand: Theory and Applications*. Ashgate Publishing: Aldershot, United Kingdom. pp. 286.
4. Gayle, P.G. 2008. An empirical analysis of the competitive effects of the Delta/Continental/Northwest code-share alliance. *Journal of Law and Economics*, 51(4):743–766.
5. Gillen, D.W., Morrison, W.G. and Stewart, C. 2002. *Air Travel Demand Elasticities: Concepts, Issues and Measurement*. Department of Finance, Canada. www.fin.gc.ca/consultresp/airtravel/airtravstdy_-eng.asp
6. Granados, N., Gupta, A. and Kauffman, R.J. 2012 Online and offline demand and price elasticities: Evidence from the air travel industry. *Information Systems Research,* 23 (1), 164-181.
7. Guevara-Cue, C.A. 2010. *Endogeneity and Sampling of Alternatives in Spatial Choice Models*. Doctoral Dissertation, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA.
8. Guevara, C.A. and Ben-Akiva, M. 2006. Endogeneity in residential location choice models. *Transportation Research Record: Journal of the Transportation Research Board*, 1977, 60-66.
9. Hotle, S., Castillo, M., Garrow, L.A. and Higgins, M.J. The impact of advance purchase deadlines on airline customers' search and purchase behaviors. *Transportation Research Part A* (submitted in April, 2014; under third round review as of May, 2015).
10. Hsiao, C-.Y. 2008. *Passenger Demand for Air Transportation in a Hub-and-Spoke Network*. Ph.D. Dissertation, Civil and Environmental Engineering, University of California, Berkeley.
11. InterVISTAS. 2007. *Estimating Air Travel Demand Elasticities: Final Report*. www.iata.org/whatwedo/Documents/economics/Intervistas_Elasticity_Study_2007.pdf.
12. Koppelman, F.S., Coldren, G.C. and Parker, R.A. 2008. Schedule delay impacts on air-travel itinerary demand. *Transportation Research Part B*, 42(3), 263-73.
13. Mumbower, S., Garrow, L.A. and Higgins, M.J. (2014). Estimating flight-level price elasticities using online airline data: A first step towards integrating pricing, demand, and revenue optimization. *Transportation Research Part A,* 66: 196–212.

Questions