GRA, Incorporated

Economic Counsel to the Transportation Industry

ACRP

Project Number 03-36

Using Disaggregated Socioeconomic Data in Air Passenger Demand Studies

Final Report

Appendix E

Background on Other Analytic Approaches

6/28/18

TRANSPORTATION RESEARCH BOARD
NAS-NRC

Prepared by:

GRA, Incorporated

With:

Atlanta Analytics, LLC

and

Aviation System Consulting, LLC

Table of Contents

List of Figures and Tables

E.1 Introduction

This appendix presents a detailed discussion of two analytical approaches addressed in the research that examine alternative sources of data or different ways in which more established disaggregated socioeconomic data can be used in air passenger demand studies:

1. The potential use of new data sources for air travel demand analysis, in particular household level financial data
2. Alternative ways to incorporate disaggregated socioeconomic data in air passenger demand studies

Section E-2 documents an exploratory analysis that was undertaken to examine how highly-detailed, disaggregated financial transaction data could potentially be used to model air travel behavior and complement existing data sources, such as air traveler surveys. The analysis utilized a sample of de-identified financial transaction data that was made available to the research by a company that provides households with online financial management software. Users link this software to their bank and credit card accounts, which allows the software to create a database of financial transactions for each user. Since some of these transactions are associated with air travel (e.g., purchasing an airline ticket) the analysis attempted to determine how much information about each air trip could be inferred from the data and how many air trips different users appeared to have made.

Section E-3 provides a detailed discussion of four alternative ways that disaggregated socioeconomic data could be incorporated in air passenger demand studies, and in particular in analytical models of air travel demand. Obviously, one way in which disaggregated socioeconomic data can be used in air passenger demand studies is to incorporate variables that reflect such data in air travel demand models. However, it is less obvious how this can be done and there are a number of conceptual and practical issues that have to be addressed in order to do so. One approach to including disaggregated socioeconomic data in air travel demand models was implemented in the case study analyses described elsewhere in the Final Report (chapter 4 and Appendix D), but this is not the only possible approach and may not be the most appropriate one. The section describes four alternative approaches that were identified in the course of the research, including the approach that was used in the case study analysis, and discusses the advantages and disadvantages of each as well as some of the implementation issues that need to be taken into account.

Section E-4 contains references for the appendix. This appendix to the project final report presents a more detailed analysis and presentation of these topics than is present in Section 5 of the final report. In combination with the analyses presented in the Final Report, these additional analyses contributed to the research results and conclusions that are contained in the Final Report.

E.2 Exploratory Analysis of New Data Sources – Using Financial Transaction Data to Model Demand for Air Travel

a. Introduction and Description of Case Study

In recent years, there has been increasing interest in using non-traditional data sources for travel demand modeling applications. The interest is motivated in part by the explosion of large, third-party data sources. These big datasets, which range from mobile phone signal traces and global positioning system (GPS) data to transit smart card or credit card spending patterns, collectively provide detailed spatial and temporal data about individuals' behaviors and mobility patterns, often in real-time. The objective of this case study to is examine how highly-detailed, disaggregate financial transaction data could potentially be used to model air travel behavior and complement existing data sources, most notably air travel surveys.

Several companies have developed financial analysis programs that allow users to manage their finances by compiling information from different sources including credit card and bank accounts. These financial management companies maintain user panels and sell the data collected from this panel to other companies. To date, we are aware of no studies that have used these data to analyze travel behavior. The objective of this case study is to examine the potential of using these financial transaction data for modeling demand for air travel.

When modeling demand for air travel, there are many characteristics about the air travel trip that are typically used. These include the number of air trips taken by an individual and household and for each trip the:

1. Airline(s)
2. Booking date
3. Departure and arrival dates
4. Departure and arrival airports
5. Destination
6. Number in travel party
7. Amount paid (per person)

Our case study uses the financial database to identify the number of air trips taken by a card holder for those households that live in zip codes located within 100 miles of a major airport in the Los Angeles area. These airports include:

- John Wayne Airport (SNA)
- Los Angeles (LAX)
- Burbank (BUR)
- Long Beach (LGB)
- Ontario (ONT)

As part of the case study, we will provide a detailed assessment of the air travel information that is available for a random sample of 50 members. This assessment will shed light on how an algorithm could potentially be developed to identify all known characteristics of air trips (and identify when these characteristics are not in the database). In this sense, the case study

and detailed assessment provide a balance between (1) illustrating the potential of using the financial transaction data in the near future for some air travel demand studies; and (2) identifying more complicated analysis that could potentially be performed for future demand studies.

b. Data Description

To conduct the analysis, we pulled a random sample of 111,632 members from a user panel comprised of U.S. residents; the panel belongs to a non-disclosed company that compiles financial transaction data. The random sample contains almost 150 million banking and credit card transactions.

For the purposes of our analysis, we only included members who met the following criteria:

(1) First transaction date in January of 2012
(2) Last transaction date in 2015
(3) Frequency of transactions between five and 150 a month

The first two criteria together ensure that each member included in the analysis has at least three years of transactional history. The third criteria is used to screen out members with a low transactional volume (which would occur if an individual signed up to use the financial management software product, but infrequently used the product) or a high transactional volume (which would occur if the account was associated with a business versus a household). After applying these criteria, our analysis database contains 66,892,885 banking and credit card transactions from 35,458 members.

The database contains information about each member's banking and/or credit card transactions. Members decide which bank accounts and/or credit cards to associate with their financial management account, thus it is possible that only a subset of the transactions is included in the analysis database. An example of transactions for a member is shown in Table E-1. Due to space limitations, not all fields are shown (in particular the member id, account id, transaction id, street address, latitude, longitude, indicator for whether the account is for a bank or credit card, and currency have been suppressed).

Lines 1-3 represent transactions with clear geographical ties to the San Marcos and San Diego areas (i.e., we expect to see automotive and grocery store transactions close to the member's home zip code). Lines 4-5 show two transactions were made on 4/21/2013 for Delta Airlines (we assume this is for two separate tickets). The transactions continue to be predominately in the San Diego, San Marcos and surrounding areas until 5/18/2013 when a series of transactions appear in Salt Lake City. These transactions include a small charge at the Cotton Bottom Inn on 5/19/2013 and a small charge of $25 on Delta for 5/19/2013 (which is likely a checked baggage fee). From the sequence of transactions, we infer that the individual purchased two tickets on Delta Airlines to travel from an (unknown) airport near their home in the San Diego area to Salt Lake City on 4/21/2013 and departed for Salt Lake City sometime between 5/15/2013 (the last transaction in San Diego) and 5/18/2013 (the first transaction in Salt Lake City) and returned home on 5/19/2013.

Table E-1. Representative Financial Transactions for One Account Holder

Line	Amount	Transaction Date	Description	Merchant name	City	State	Zip	Transaction
1	110	4/18/13	JIFFY LUBE #1967 SAN MARCOS CA	Jiffy Lube	San Marcos	CA	92078	Physical
2	33	4/18/13	RALPHS #0167 SAN DIEGO CA	RALPHS	San Diego	CA	92130	Physical
3	39	4/18/13	RANCHO SANTA FE GAS SAN MARCOS CA		San Marcos	CA		Physical
4	300	4/21/13	DELTA XXXXXXXXXXXXX ATLANTA GA		Atlanta	GA		Physical
5	300	4/21/13	DELTA XXXXXXXXXXXXX ATLANTA GA		Atlanta	GA		Physical
6	25	4/21/13	TRADER JOE'S #221 QPS SAN DIEGO CA	TRADER JOE'S	San Diego	CA	92131	Physical
7	11	4/21/13	TRADER JOE'S #221 QPS SAN DIEGO CA	TRADER JOE'S	San Diego	CA	92131	Physical
8	356	4/21/13	USAA.COM PMT - THANK YOU SAN ANTONIO TX	Payment Thank You				Non-Physical
			. . .					
20	40	4/30/13	CHEVRON XXXXXXX SAN DIEGO CA	CHEVRON	San Diego	CA	92129	Physical
21	6	5/1/13	STARBUCKS #XXXXX SAN M SAN MARCOS CA	STARBUCKS		CA		Physical
22	29	5/1/13	URBAN OUTFITTERS #157 CARLSBAD CA	Urban Outfitters	Carlsbad	CA	92009	Physical
23	4	5/3/13	REDBOX *DVD RENTAL XXX-XXX-XXXX IL	REDBOX	Carlinville	IL	62626	Physical
24	35	5/5/13	24HOUR FITNESS USA,INC XXX-XXX-XXXX CA					Non-Physical
25	8	5/5/13	NETFLIX.COM NETFLIX.COM CA	NETFLIX, INC.				Non-Physical
26	26	5/6/13	ALBERTSONS #6772 SAN DIEGO CA	ALBERTSONS	San Diego	CA	92129	Physical
			. . .					
52	4	5/15/13	STARBUCKS #XXXXX SAN M SAN MARCOS CA	STARBUCKS		CA		Physical
53	33	5/18/13	H & M #257 SALT LAKE CITUT	H & M		UT		Physical
54	15	5/18/13	HATCH FAMILY CHOCOLATE SALT LAKE CITUT	Hatch Family Chocolates		UT		Physical
55	17	5/19/13	COTTON BOTTOM INN SALT LAKE CITUT	Cotton Bottom Inn	Salt Lake City	UT	84121	Physical
56	25	5/19/13	DELTA XXXXXXXXXXXXX SALT LAKE CTYUT			UT		Physical
57	5	5/19/13	ROCKY MOUNTAIN CHOCOLA SALT LAKE CITUT	Rocky Mountain Chocolate Factory		UT		Physical
58	25	5/21/13	ALBERTSONS #6772 SAN DIEGO CA	ALBERTSONS	San Diego	CA	92129	Physical
59	46	5/23/13	CASA SOL Y MAR SAN DIEGO CA	Casa Sol y Mar	San Diego	CA	92130	Physical
60	15	5/23/13	IN-N-OUT BURGER #68 SAN DIEGO CA	IN-N-OUT BURGER	San Diego	CA	92126	Physical

This example highlights some of the key challenges that we faced when using the database for analysis. First, we do not know where the card member lives and need to infer the home location from the transactional history. The analysis database is a partial view of financial transactions made by a member on one or more credit cards and/or bank accounts. Multiple individuals may be associated with an account. That is, the financial database represent account holders – which in many (but not all) cases will translate to household members that live in the same physical location. However, in some cases (most notably with college students), the members associated with an account can live in different physical locations (which gives results in transactions occurring simultaneously in different states). This makes it challenging to identify the "home zip code" for the card holder. In addition, although the merchant name and associated zip code where the transaction was made are present for many of the financial transactions, they are not present for all of the transactions. In the example shown in Table E-1, the merchant name for the transactions associated with Delta Airlines is missing; thus, the only way to reliably determine the merchant name for these transactions is by using the description field. In addition, the merchant name and associated zip code are incorrect for some of the transactions. The Redbox DVD rental shown on line 23 is noted as a "physical transaction" (meaning it was made at a brick-and-mortar store versus a "non-physical" internet store). However, the physical location is shown as being in Illinois (the site of Redbox's corporate headquarters). Thus, one cannot simply use the "first transaction" physically located more than a certain distance from the individual's (likely) home zip code as the likely location where the individual traveled by air. A more sophisticated algorithm that takes these and other subtleties into account was developed to identify detailed characteristics associated with the air travel. Even with a sophisticated algorithm, certain characteristics of the air trip (such as the departure date for the example shown in Table E-1) will be hard to infer from the financial transaction database.

c. Methodology

Figure E-1 summarizes the methodology we used to process the financial transaction data and conduct the case study and detailed assessment. First, we developed an algorithm to identify members who likely made one or more air trips. Next, we developed an algorithm to identify members who likely live within 100 miles of a major airport in the Los Angeles area. Applying these two algorithms on our analysis database revealed 1,028 members with 6,718 transactions associated with air travel (out of a total of more than 2.9 million transactions). Note the number of transactions does not directly translate into the number of air trips as multiple transactions can be associated with the same trip, e.g., in Table E-1 there are three "Delta" transactions associated with a single air trip involving a party of two. The third part of our methodology involves manually examining financial transactions for a random sample of 50 members to determine how often we can identify seven characteristics associated with an air trip: the booking date, outbound and inbound travel dates, airline (or airlines), destination, departure and arrival airports, fare, and number in the party. This manual examination provides insights into the complexity of an algorithm that will be needed to automatically identify these characteristics of air travel demand from the transactional database (and/or to know when it is possible to identify these characteristics).

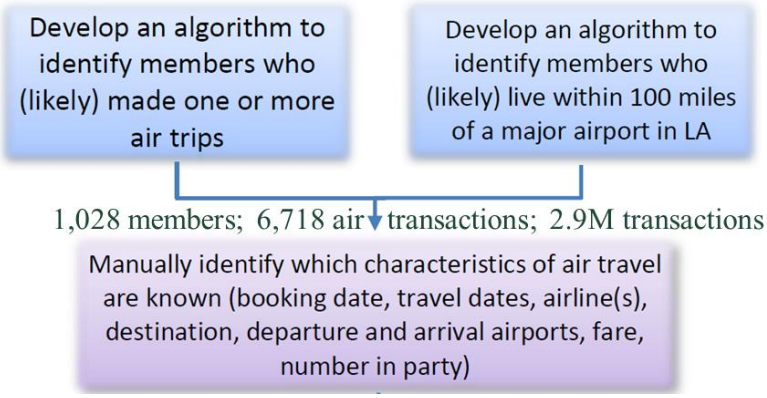

Figure E-1. Methodology for Processing Financial Data for Air Travel Studies

Step 1: Develop an Algorithm to Identify Members who Likely Made Air Trips

Based on the example shown in Table E-1, we determined that we could not use the "merchant name" field to identify transactions associated with air travel, as the merchant name was not populated for many air-related trips (such as the ones for Delta Airlines shown in Table E-1. We thus decided to use the "description" field. To understand how we developed an algorithm to identify members who likely made an air trip, we define "description codes" as the unique set of description fields in the analysis dataset. Using the description field posed a challenge, because there are often many description codes associated with a single company. Of the 66,892,885 transactions in the analysis database, there were 9,871,274 unique description codes. The frequency distribution for these description codes is shown in Table E-2 and reveals that there are a handful of description codes that appear very often in the database and many description codes that appear very infrequently. The frequency distribution in Table E-2 is sorted according to how often a description code appears in the analysis database with those codes appearing most frequently shown at the top. As shown in line one of the table, just three description codes (or 0.00003% of the more than 9.8 million unique codes) represent 2,317,699 (or 3.5%) of all transactions. Similarly, a mere 256 description codes (or 0.0026%) appear 7,500 or more times in the database and account for 20% of all transactions. However, there are also many description codes that appear infrequently – indeed, 60% of all description codes representing 8.9% of all transactions appear just once in the transaction database.

To identify members with likely air travel, we manually examined all description codes that appeared 125 times or more in the analysis database. These description codes represent 41.4% of all transactions in the analysis database. Based on an examination of these fields, we identified 29 text strings shown in Table E-3. These text fields were processed sequentially so that "SW AIR" or "AA AIR" (which are more likely to be associated with an airline) were identified and recognized as having a higher confidence of being associated with an air trip than the word "AIR." Some of the text codes, such as the one shown in line 7 "SW AIR" can be clearly associated with an airline (namely Southwest Airlines). In this case, all 20 description codes that contain the text "SW AIR" can be mapped to Southwest Airlines. However, most of the text strings are associated with multiple companies. The broader (or more popular) the text

string, the more likely it is to be associated with multiple companies, e.g., the text "American" appears in 45,655 description codes, "Spirit" in 9,750, and "Frontier" in 2,586.

Table E-2. Distribution of Description Codes and How They Relate to the Number of Transactions

# Times Description In Database	# Des. Codes	# Trans	% Des. Codes	% Trans	Cum % Des. Codes	Cum % Trans
750,000-784,604	3	2,317,699	0.00003%	3.5%	0.00003%	3.5%
500,000-749,999	4	2,758,168	0.00004%	4.1%	0.00007%	7.6%
250,000-499,999	4	1,522,890	0.00004%	2.3%	0.00011%	9.9%
100,000-224,999	15	2,345,174	0.0002%	3.5%	0.00026%	13.4%
50,000-99,999	13	895,657	0.0001%	1.3%	0.00040%	14.7%
40,000-49,999	6	265,155	0.0001%	0.4%	0.00046%	15.1%
30,000-39,999	11	395,392	0.0001%	0.6%	0.00057%	15.7%
20,000-29,999	38	917,505	0.0004%	1.4%	0.00095%	17.1%
10,000-19,999	99	1,400,962	0.001%	2.1%	0.0020%	19.2%
7,500-9,999	63	544,627	0.001%	0.8%	0.0026%	20.0%
5,000-7,499	98	586,644	0.001%	0.9%	0.0036%	20.9%
1,000-4,999	2,024	3,781,185	0.02%	5.7%	0.024%	26.5%
500-999	3,573	2,438,074	0.04%	3.6%	0.060%	30.2%
400-499	2,156	959,473	0.02%	1.4%	0.082%	31.6%
300-399	3,898	1,336,900	0.04%	2.0%	0.12%	33.6%
200-299	8,972	2,168,907	0.09%	3.2%	0.21%	36.8%
175-199	4,380	815,591	0.04%	1.2%	0.26%	38.0%
150-174	6,321	1,018,516	0.1%	1.5%	0.32%	39.6%
125-149	9,151	1,243,396	0.1%	1.9%	0.41%	41.4%
100-124	14,189	1,574,721	0.1%	2.4%	0.56%	43.8%
75-99	27,328	2,335,072	0.3%	3.5%	0.83%	47.3%
50-74	56,096	3,381,304	0.6%	5.1%	1.40%	52.3%
25-49	195,270	6,719,071	2.0%	10.0%	3.38%	62.4%
11-24	477,788	7,600,340	4.8%	11.4%	8.22%	73.7%
10	76,200	762,000	0.8%	1.1%	8.99%	74.9%
9	89,657	806,913	0.9%	1.2%	9.90%	76.1%
8	110,827	886,616	1.1%	1.3%	11.0%	77.4%
7	137,924	965,468	1.4%	1.4%	12.4%	78.8%
6	182,324	1,093,944	1.8%	1.6%	14.3%	80.5%
5	243,988	1,219,940	2.5%	1.8%	16.7%	82.3%
4	364,816	1,459,264	3.7%	2.2%	20.4%	84.5%
3	595,352	1,786,056	6.0%	2.7%	26.5%	87.2%
2	1,331,575	2,663,150	13.5%	4.0%	40.0%	91.1%
1	5,927,111	5,927,111	60.0%	8.9%	100.0%	100.0%
TOTAL	9,871,274	66,892,885				

However, there are many companies that have these text strings – American Red Cross, American Family Insurance, American Girl, Wine and Spirits, Frontier Communications, etc. Thus, the key challenge is that we need to define the text strings broadly enough to identify transactions associated with travel without making the search too broad.

Table E-3. Text Codes Used to Identify Transactions with Potential Air Trips

	Text String	# Desc. Codes
1	TRAVELOCITY	207
2	EXPEDIA	978
3	PRICELINE	364
4	HOTWIRE	159
5	GOGOAIR	199
6	TICKETFLY	99
7	SW AIR	20
8	JETBLUE	908
9	USAIR	1,677
10	US AIR	631
11	FRONTIER	2,586
12	AIRTRAN	387
13	ALLEGIANT	84
14	SPIRIT	9,750
15	VIR AMER	128
16	AA AIR	98
17	SOUTHW	5,746
18	UNITED	10,778
19	AMERICAN	45,655
20	DELTA	7,066
21	IN FLIGHT	116
22	INFLIGHT	346
23	ONBOARD	66
24	SMARTE CARTE	86
25	TRAVEL	14,735
26	CRUISE	1,458
27	HUDSON NEWS	1,552
28	AIR	133,581
29	FLIGHT	1,073

To achieve this balance, we used an iterative approach shown in Figure E-2. As the first step, we identified members that had a description field with at least one of the text strings shown in Table E-3. Next, we identified members that lived within 100 miles of an airport in Los Angeles (the methodology we used to do this is described below). This second step allowed us to reduce the size of the analysis database that was more manageable for manual processing. From

here, we sorted the description codes from most frequent to least frequent and identified text string that were clearly not associated with air travel, e.g., we could eliminate all transactions associated with the "American Girl Place" by eliminating description codes that contained the work "Girl," or the majority of beer and wine stores by eliminating description codes that contained the word "Spirits" (the plural of Spirit) or "Wine." We continued in this manner, eliminating description codes until all that we had remaining were (most likely) associated with air travel.

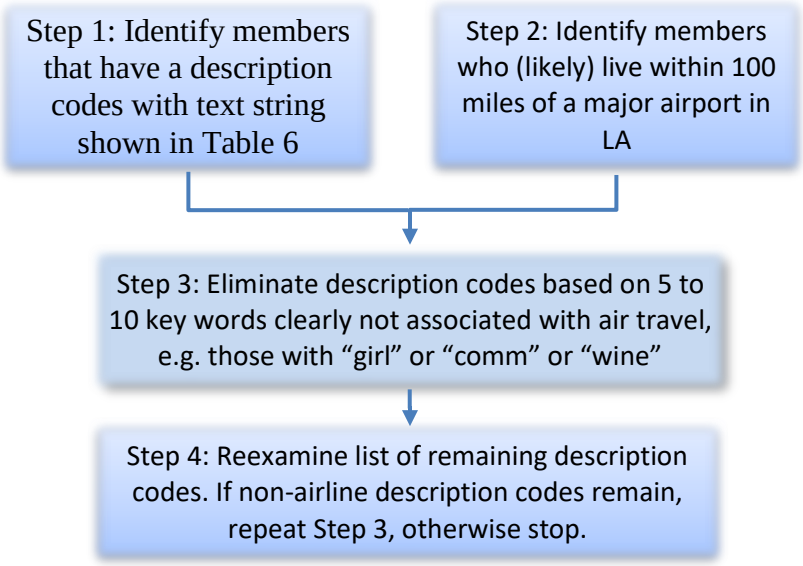

Figure E-2. Process to Identify Description Codes Likely Associated with Air Travel

Step 2: Develop an Algorithm to Identify Members who Likely Live within 100 Miles of a Los Angeles Area Airport

To identify members who likely live within 100 miles of an airport in the Los Angeles area, we first needed to identify the zip codes that were located within these 100-mile boundaries. We used the online site FreeMapTools[1] to identify zip codes located within 100 miles of each of the airports in the Los Angeles area. The site allows the user to enter a zip code and radius and then returns a list of zip codes that are located within the radius of the input zip code. The state, county, city, time zone, and distance (in km) associated with each zip code returned is also provided. In our application, we entered the zip code associated with an airport and a 100-mile radius. The number of zip codes returned for each airport is shown in Table E-4. All of the airports returned zip codes from the counties of Los Angeles, Kern, Orange, Riverside, San Bernardino, San Diego and Ventura. All airports except for John Wayne (SNA) and Ontario (ONT) also returned zip codes from Santa Barbara County. Given the close proximity of airports in this area, there was a high degree of overlap in zip codes associated with each airport, resulting in 1,474 unique zip codes.

[1] See https://www.freemaptools.com/find-zip-codes-inside-radius.htm.

Table E-4. Zip Codes and Cities Located Within 100 Miles of Airports in Los Angeles Area

Airport (IATA Code)	Zip Code Where Airport is Located	# Zip Codes Within 100 Miles of Airport
Burbank (BUR)	91505	967
John Wayne (SNA)	92707	1,075
Long Beach (LGB)	90808	1,059
Los Angeles (LAX)	90045	975
Ontario (ONT)	91761	1,066

Determining the set of unique zip codes located within 100 miles the five airports in the Los Angeles area was relatively straight-forward. Determining the home zip code associated with a member proved to be more challenging. Of the 66,892,855 transactions in the database 37,181,222 (or 56%) are associated with a "Physical" transaction origin (or a transaction that occurs in a brick-and-mortar store). In theory, each of these transactions should have an associated address. However, many of these physical transactions do not have associated address information, e.g., as shown in Table E-5 less than a third of these transactions have a zip code. More aggregate location-based information (namely city and state) is present for 68.2 – 78.5% of these transactions, respectively, but these more aggregate data are of limited value for the air travel demand case study.

Table E-5. Transactions with a Physical Origin that Have Location-Based Information

	# of Physical Transactions	% Physical Transactions
Zip code	11,173,632	30.1%
Street	11,081,580	29.8%
City	25,370,885	68.2%
State	29,179,414	78.5%
Latitude	10,966,225	29.5%
Longitude	10,966,225	29.5%
TOTAL	37,181,222	

To develop an algorithm for assigning a home zip code for each member, we compared three methods. For a given member, we assumed the home zip code was the one that:

- Method 1: Occurred most frequently across all days of the week
- Method 2: Occurred most frequently for transactions that occurred on Saturday and Sunday
- Method 3: Occurred most frequently among zip codes associated with purchases made at Walmart or Target stores.

We compiled a list of members who met at least one of the criteria identified above and included the member only if the zip code obtained using a given method was located within the 100 mile area of an airport in the Los Angeles area. A total of 2,559 members met these criteria. Among these members, 2,075 (or 81%) had a zip code populated based on method 1, 2,054 (or

80%) based on method 2, but only 1,005 (or 41%) based on method 3. Given that the majority of members did not have a zip code assigned based on method 3 (either because they did not shop at Target or Walmart or – more likely – because the zip code associated with these transactions was not in the analysis database) we subsequently focused our analysis on comparing results obtained between methods 1 and 2.

For cases in which the zip codes obtained using method 1 and method 2 differed, we used the distance (zip code 1, zip code 2) function in SAS to calculate the distance between these two zip codes. The distribution of zip codes is shown in Table E-6, e.g., 10% of these zip codes were located within 2.3 miles of each other, 50% within 6.4 miles, and 80% within 19.9 miles. Given zip codes that are located close to each other tend to have similar sociodemographic information, we decided to use the zip code calculated from method 1 (the zip code that occurred most frequently across the transactions) as a proxy for the member's home zip code.

Table E-6. Distribution of Distances Between Zip Codes Using Methods 1 and 2

	Distance (miles)
10%	2.3
20%	3.1
30%	3.9
40%	4.9
50%	6.4
60%	8.7
70%	12.3
80%	19.9
90%	40.1

Based on these assumptions, the resulting database (representing members who likely live within 100 miles of an airport in the Los Angeles region and who likely made at least one air trip) contains 1,028 members and 6,718 transactions potentially associated with air travel (out of a total of 2,904,920 transactions).

Step 3: Identify How Often Air Travel Characteristics Can Be Identified from a Sample of the Data

Using a sample of 50 random households, we determined how often the following air travel characteristics could be identified: departure arrival and connecting airports, booking date, price, airline, group size, departure date and return date. The logic that we used to identify air trip characteristics is described in this section. This logic could be built upon as part of future work to develop an algorithm to automatically identify these trip characteristics from the financial transaction data.

Step 3A: Identify a financial transaction that is likely associated with an air travel trip (see Step 1 described earlier for this methodology).

Step 3B: Compute the number of travelers associated with the air trip, defined as the number of transactions that share the same purchase date, airline, and price.

Step 3C: Determine if one or multiple airlines is associated with the air trip. Assume multiple airlines are involved if there are tickets purchased on different airlines on the same day.

Step 3D: Determine if there is any information about the destination that can be obtained from transactions that occur two days before or two days after the airline ticket purchase. In some cases, individuals pre-purchase their lodging accommodations and the location is included on the transaction details.

Step 3E: Determine when and/or where the air trip began. There are multiple ways this can be done and it is important to recognize that some transactions provide a higher degree of certainty that an air trip occurred. We call these "high confidence transactions"; examples include Gogo wireless purchases, onboard purchases, and checked baggage fees. The checked baggage fees are particularly helpful, as they often reveal the origin airport code as well as the date of travel.

a. Identify the first occurrence after an air ticket purchase that involves Gogo wireless, an onboard purchase, checked baggage fee or other transactions that are most likely to occur on the day of an air trip. Use this information to determine the travel date and if possible the departure airport.

b. Identify transactions that occur immediately before or immediately after these "high confidence" transactions and determine if the transactions are occurring in a location outside the home area. Assume these transactions are occurring at the destination associated with the air trip.

c. Scan transactions that occur at the end of the air trip (or a maximum of seven days after a high confidence transaction) for potential airport parking fees. This provides information on the airports used.

d. If there are no high confidence transactions, scan the financial transaction data to determine if there is a period of time (one – two days minimum) in which the location where transactions were occurring changed. Assume, unless this is a regular occurrence (which would occur if multiple cardholders associated with a household lived in different physical locations) that this is the location associated with the air trip. Identify the likely departure and/or return date ranges (for trips that do not contain high confidence transactions). For example, in the case of the outbound departure date, the earliest departure date is associated with the last transaction occurred in the home area and the latest departure date is associated with the first transaction associated with the "away" location.

Any algorithm developed would also need to recognize and consider "exceptions." Examples of these exceptions that we identified in our manual assessment include the following:

- An air trip that includes multiple destinations and/or multiple modes (such as a long driving trip)

- An air trip that has an associated on-board transaction but no corresponding air ticket
- Transactions that are likely associated with a cancellation or exchange
- Households that purchase air tickets for individuals that do not reside in the home (e.g., for others to attend a funeral or bachelor party)

As shown in Table E-7, based on a sample of 266 round-trip tickets purchased over a three-year period, the analysis showed that we could identify the airline, price, group size and booking date for almost all of the air trips; however, it was difficult to identify the departure and arrival airports because few transactions occurred on the day of travel near the airport location.

Table E-7. How Often Trip Characteristics Were Identified in Sample of Financial Data

Trip Characteristic	Percent Identified
Airline	100%
Price	97%
Group size	99%
Booking date	95%
Destination	38%
Outbound departure date	26%
Outbound departure date range	39%
Inbound return date	30%
Inbound departure date range	40%
Outbound departure airport (or inbound arrival airport)	6%
Outbound arrival airport (or inbound departure airport)	3%

We also discovered that the air trips in the financial transaction database are quite different than those represented in typical air passenger surveys. As shown in Table E-8, the financial transaction database contains a much larger percentage of air trips with two or more travelers than given by two representative air passenger surveys for personal trips. It was assumed that since many business trip expenses would not be charged to a personal credit card or bank account, most of the air trips appearing in the transaction data are likely to be personal trips.

Table E-8. Comparison of Group Size Distributions

Group Size	Financial Transaction Data	SFO 2014/15 Air Passenger Survey *(1)*	MWCOG 2013 Air Passenger Survey *(2)*
1	4.2%	53.7%	51.5%
2	74.1%	32.4%	35.6%
3+	21.7%	13.9%	12.9%
Average	2.46 pax	1.71 pax	1.71 pax

Notes: 1. Survey responses for U.S. residents making personal trips (domestic and international)
2 Survey responses for MWCOG regional residents making personal trips (domestic and international)

d. Summary

Based on the analysis of the financial transaction database, we conclude that financial transaction data show promise for being used in the future, but currently lack some critical information (most notably the ability to consistently identify the airports used. It is possible that as the market penetration and consumer acceptance of online financial tools increases, so too will the value of the financial transaction data that can be collected and analyzed for planning applications.

In addition, our analysis of the financial transaction data provides insights into the complexity of the algorithms that would need to be developed to identify air travel trips and their associated characteristics within financial transaction data.

E.3 Incorporation of Disaggregated Socioeconomic Data in Air Passenger Demand Studies

The research undertaken during the project has identified four different ways in which disaggregated socioeconomic data could be incorporated into air passenger demand models:

- Use of variables that reflect the shape of the distribution of the explanatory variable in a single relationship as well as variables that reflect the aggregate or average value of the explanatory variable;
- Use of separate variables for different ranges of the explanatory variable in a single relationship;
- Use of separate relationships for different ranges of each explanatory variable;
- Use of a simulation approach that generates a measure of trip propensity for individuals with specific values of the explanatory variables

This section of the appendix describes each of these approaches in more detail and discusses some of the implementation issues that need to be considered in applying them. Since three of the four approaches require fairly detained air passenger survey data and preferably data for several points in time, the selection of the approach to be used in any given study should take into account the availability and extent of air passenger survey data.

a. Discussion of Alternative Approaches

Use of Variables Reflecting the Shape of the Distribution of Explanatory Variables

This approach might include, for example, variables for the 20th and 80th percentiles of the household income distribution in addition to the average household income. Since air travel propensity varies with income, the resulting demand model could be expected to indicate how air travel demand is sensitive to changes in the distribution of household incomes as well as the average or aggregate income. For models where the explanatory variables are multiplied together (such as the common log-linear model), the variables for the distribution percentiles should be expressed as ratios of the average income rather than in monetary terms to avoid having two income values multiplied together.

Although in principle this approach could be applied to any socioeconomic variable, it makes most sense to apply it to the income variable, since it is known that income distributions have been changing over time, that air travel demand varies with household income, and data on household income distributions for each year are readily available from the U.S. Census Bureau. The U.S. Census Bureau report *Income and Poverty in the United States: 2015* (Proctor, *et al*. 2016) includes a table showing the household income distribution for each year from 1967 to 2015. The Census Bureau data show that households with an income of $200,000 or more in 2015 CPI-U-RS (consumer price index for all urban consumers using current methods – research series) adjusted dollars increased from 2.1% of all households in 1985 to 5.1% of all households in 2005 and 6.1% of all households in 2015. Conversely, households with an income of $25,000 per year or less declined from 24.6% of all households in 1985 to 22.0% of all households in

2005 and 22.1% of all households in 2015. Perhaps of greater relevance for air travel demand, the percent of households with incomes between $50,000 and $100,000 declined from 33.3% of all households in 1985 to 30.4% of all households in 2005 and 28.8% of all households in 2015.

Although the annual changes in these percentages are not large and the trend has been fairly stable over the last 30 years, some of the increase in air travel over that time period is clearly due to the changing income distribution (a higher percentage of households in the higher income brackets that have a higher air travel propensity) rather than the change in the average (and hence aggregate) household income. This showed greater year-to-year variability, increasing from $61,049 in 1985 to $76,878 in 2005 (a 25.8% increase) and to $79,263 in 2015 (3.2% increase from 2005), but declining in some years, particularly during recessions.

Relevant issues that were not resolved in the course of the project, but need to be addressed in future research, include the relative effectiveness of using different variables to reflect the shape of the distribution of explanatory variables, as well as the development of a theoretical basis for including such variables in air passenger demand models both to guide the choice of appropriate variables and to avoid adverse effects of including additional variables in a way that produces illogical or inherently unreasonable results.

This brings up the important issue of whether to include demographic or socioeconomic variables other than income in aggregate air travel demand models. Analysis of air passenger and household travel surveys have shown that air travel propensity varies by age, gender, race/ethnicity, and educational attainment, as well as income, although some of these effects may be correlated. For example, the analysis of demographic trends presented elsewhere in the project Final Report shows that the median income of households with a head of household age 65 or older has grown faster than all other households. The extent to which changes in the distributions of these characteristics may help explain changes in the demand for air travel can be addressed by examining the correlation between model residuals and variables measuring the changes in the distributions of these characteristics.

One potential challenge to the application of this approach that arose in the model development work undertaken as part of the project is the extent of correlation between the distributional and aggregate variables, which could make it difficult to obtain statistically significant coefficients for both types of variables in the same model. Both types of variables are slow-moving and generally changed in a consistent direction over the period from 1990 to 2010 that was used to estimate the air travel demand models developed in the course of the project, and were found to be fairly highly correlated. Although this correlation proved limiting with the simpler models examined, it was found that with careful definition of the relevant variables it was possible to obtain statistically significant estimated values for the coefficients of variables that measured both the average per capita household income in a region and the percentage of total regional household income received by the ten percent of households with the highest household incomes, which provided a measure of changes in income distribution over time.

Use of Separate Variables for Different Ranges of Explanatory Variables

This approach provides greater flexibility in reflecting the effect of the underlying distribution of particular socioeconomic variables, such as income or air traveler age, by defining

separate variables for different income or age ranges. However, this imposes two limitations on the analysis. The first is a more complex functional form, since the variables for each value range for a given factor cannot simply be multiplied together. Rather, terms for the demand generated by each subset of the total population formed by the various value ranges (each of which may well have a multiplicative form) must be added together to give the total demand.

For example, a simple model using total household income (*H*) and average airfare (*P*) might take the following form:

$$Pax = (a1.H1^{b1}.P^{c1}) + (a2.H2^{b2}.P^{c2}) + \ldots + (an.Hn^{bn}.P^{cn})$$

where *H1, H2, .. Hn* represent the total household income for different income ranges.

This formulation allows for a different income and airfare elasticity for each income range, although the model structure could be simplified by restricting each income range to have the same airfare elasticity (*i.e.* c1 = c2 = … = cn = c) or even the same income elasticity (*i.e.* b1 = b2 = … = bn = b) although it would be surprising if in fact air travel demand for households in different income ranges had the same sensitivity to changes in income. Nonetheless these are research issues that could be explored.

The second limitation is the need for a much larger number of data points in order to estimate the larger number of model coefficients implied by the approach. This suggests that this approach would be more appropriate to use with a large panel dataset, such as the demand in multiple origin and destination (O&D) markets or for a single model that is estimated across a number of airports or regions, such as a single model for all six case studies.

The resulting model segmentation is also likely to result in model functional forms (such as the one shown above) the coefficients of which cannot easily be estimated using standard linear regression estimation techniques. Therefore, further research is needed to identify appropriate statistical techniques to estimate such model functional forms, such as the use of maximum likelihood estimation or nonlinear regression.

Use of Separate Relationships for Different Ranges of Each Explanatory Variable

This approach avoids some of the technical challenges of the second approach by estimating separate models for different subsets of the population. Although this is likely to require a less complex functional form for each model, it requires an estimate of the demand generated by each subset of the population. Since there is no way to obtain this from the reported air passenger traffic, it requires air passenger survey data to segment the reported traffic into the air trips made by each subset of the population. However, segmenting the models in this way not only simplifies the model estimation process (since the model for each segment can have the same functional form as a model developed using aggregate data), but allows a much finer definition of the population subsets than would be practical with the second approach.

The challenge with applying this approach is that most airports have limited air passenger survey data (or none at all) that can be used to estimate the proportion of the total air passenger traffic in each subset of the population. Even those airports that have undertaken several air passenger surveys over time typically do not have survey data for every year. In these cases it would be necessary to interpolate the proportions between the years for which survey data is

available. In order to assess the likely validity of this approach, it would be helpful to undertake some analysis of how much these proportions vary from year to year (at least for those years for which survey data is available).

For airports where air passenger survey data is only available for one year, it may be possible to estimate the proportions for other years by applying the trends from airports for which multiple surveys are available. This is obviously less satisfactory than having survey data for multiple years, but may still be better than ignoring the differences between subsets of the population entirely.

One potential approach, although one that would require a significant amount of data analysis, would be to examine trends over time in the spending on air travel and number of air trips by households in the different subsets of the population, using data from the Consumer Expenditure Survey (CES), which is available on an annual basis. Although the sample size of the CES is not large enough to obtain reliable data for a given geographical area served by a specific airport or regional airport system, it is large enough to allow some segmentation between major metropolitan areas and smaller communities in general, as well as differences between different regions of the U.S. (defined broadly).

Although the work involved in applying these different approaches to a set of case study airports was beyond the resources of the current project, future research could explore both the practicality and benefits of the different approaches by applying them to a specific region for which several air passenger surveys are available in order to determine how consistent the different approaches appear to be. One potential candidate is the Baltimore-Washington metropolitan region, for which air passenger surveys have been undertaken in a consistent way at regular intervals for many years.

Use of Simulation to Generate Estimates of Air Travel from Trip Propensity Data

This approach takes a much more disaggregate approach to forecasting air travel demand based on air travel propensity relationships identified through analysis of air passenger survey data rather than a conventional econometric approach that attempts to fit a model to observed data using regression or other techniques. Although this approach requires a much greater amount of data than traditional econometric techniques that are based on reported air passenger traffic and fairly aggregate measures of potential causal variables, it avoids the inherent constraints imposed by the functional form of any given econometric model and provides much greater flexibility to vary the assumptions used for future values of the causal variables.

However, while past air travel propensity relationships can be determined from survey data, these relationships do not explicitly consider the effect of changes in pricing. Any practical application of this fourth approach must also include a way to incorporate changes in airfares and other travel costs as well as changes in air travel propensity that result from trends in the level and distribution of socioeconomic variables.

The effect on air travel propensity of changes in airfares and other air travel costs can be assessed by applying estimates of the price elasticity of air travel demand. This has been extensively studied for airfares and estimates exist for different types of air trip, as discussed in

the literature review documented in Appendix A of the project Final Report. Generally, it has been found that the airfare price elasticity for business travel is somewhat less than -1 in an absolute sense (*i.e.* air travel declines somewhat less than proportionately to an increase in cost) while airfare price elasticity for nonbusiness travel is somewhat greater than -1 in an absolute sense, reflecting that nonbusiness travel is generally more discretionary than business travel, so if airfares decline in real terms, households choose to make more air trips at the expense of other consumption, and *vice versa* if real airfares rise.

Although the exact airfare price elasticity for households with a given set of characteristics will not be known to any degree of precision, using an approximate elasticity value will correct for much of the effect of airfare changes and any errors this introduces can be corrected by calibrating the results to actual air passenger traffic levels.

Where air passenger survey data is available for surveys undertaken at the same airport over a period of time, as is the case for several of the airports or regions studied in the current project, trends over time in air travel propensities for survey respondents with a given set of socioeconomic characteristics can be determined and applied to years for which survey data is not available. If these changes in trip propensity can be shown to be reasonably consistent across different airports, even if the actual trip propensity values differ, then they could be applied to airports for which survey data is only available for one year, or even to airports for which no air passenger survey data is available, by adjusting the resulting estimates of total air passenger demand to conform to the actual air passenger traffic levels.

Given estimates of air travel propensity for households with any given set of household characteristics for a given year, taking onto account the effect of changes in airfares and travel costs, the total number of air trips at an airport or for a region in a given year can be estimated by generating a synthetic sample of households with appropriate characteristics from the regional distributions of household characteristics and then simulating the number of air trips that each of these households would be expected to take in the year. The resulting projected air trips would then be calibrated to the actual passenger traffic for past years and the resulting calibration factors used to forecast future air travel, based on scenarios for future trends in socioeconomic characteristics and future trends in airfares and other travel costs.

b. Implementation Considerations

Business and Personal Travel

The analysis of air passenger and household travel surveys undertaken in the current project has shown, not surprisingly, that the distributions of household characteristics of those making business and personal trips are significantly different. More important from the perspective of air passenger demand modeling, the factors influencing the demand for business and personal air travel are also likely to be different. In reality, it is not households that generate business trips but businesses (although of course those making business trips are members of households). Therefore the level of business trips at an airport is likely to depend on the composition and size of the local economy, as well as other factors unrelated or only indirectly related to the distributions of household characteristics in the region served by the airport.

This suggests that one dimension of disaggregation in air passenger demand modeling would be to distinguish between business and personal trips. This could be fairly easily accomplished by the last three approaches discussed above, but would be more difficult, to address using the first approach discussed. However, the more detailed modeling of air passenger traffic in the Baltimore-Washington region described elsewhere in the Final Report (chapter 4 and Appendix D) was able to obtain a statistically significant coefficient estimate for employment per capita, which was assumed to account for changes in the proportion of business air travel, as discussed in the description of that model development. The underlying concept in each of the last three approaches is to add terms to the demand function to cover business trips or (in the case of the third of these approaches) simulate personal and business trips separately. Of course, this requires knowing the split between business and personal travel, but this can be obtained from air passenger or household travel survey data in exactly the same way as estimating the proportion of trips by households with different characteristics.

Although the simulation of business travel could be based on the number of households with given characteristics in the region and the business trip propensity by household, it would be more logical to base the simulation of business trips on employment by sector. This would allow forecasts of future air passenger demand to reflect projected or assumed changes in the growth of employment by sector. Similarly, in the second and third approach, the demand function terms for business travel could use variables reflecting the composition of the local economy and employment levels rather than household characteristics.

Developing estimates of business air trip propensity relative to sectoral employment would require analysis of air passenger or household travel survey data, or other data sources, to identify the distribution of business trips by economic sector. Unfortunately, few air passenger surveys ask about the type of firm or organization that business travelers are employed in and given the large number of economic sectors and the likely variation in business travel propensity per employee, obtaining this information in sufficient detail from air passenger or travel surveys would be very unwieldy. Furthermore, including sufficient terms in a single air passenger demand model to address many different sectors would result in an excessively complex functional form for which it would be extremely difficult, if not impossible, to obtain statistically significant coefficient estimates.

A more practical approach would be to use total employment and an estimate of average business air trip propensity that is obtained from a separate analysis of business air trip propensity in each sector and the sectoral composition of the local economy. This is still sensitive to changes over time in the sectoral composition and can reflect anticipated future changes. Since estimating business air trip propensity by sector from air passenger or travel survey data is problematical, it may be more productive to analyze differences in business spending on air travel by sector. There is a considerable amount of data on business travel from a wide range of sources[2] that could provide estimates of trends in business travel expenditures and average costs per trip (which would allow expenditures to be converted to trips). The 2007 input-output model of the U.S. economy available on the website of the U.S. Bureau of Economic Analysis provides a detailed breakdown of spending on air transportation by economic sector. Although these data are only for one point in time, the relative business air trip propensities

[2] See https://www.creditdonkey.com/business-travel-statistics.html for examples of business travel data.

across different economic sectors are likely to be fairly stable, since they are largely determined by the types of activities undertaken by employees in firms or other organizations in each sector. Therefore these data can be combined with trend data from other sources to obtain estimates of business air travel propensity by sector for other years.

Air Trips by Residents and Visitors

Analysis of air passenger survey data has also shown that there are differences in the composition of air passenger trips at a given airport between those made by residents of the region served by the airport and visitors to the region. Although the approaches described above make sense for air trips generated by residents of a region, it is less clear that they apply to air trips by visitors. At a minimum, the distributions of household characteristics for visitors are likely to be different from those for residents, and in any case the air trips to a region made by visitors are not the only air trips that those people made. There may be a degree of symmetry for some types of trip between trips made by residents and those made by visitors. For example, trips by residents to visit family and friends elsewhere may be balanced by trips by visitors to visit family and friends in the region. However, for other types of trip, such as those for vacations, attending college, or medical treatment, there is no reason to expect that the levels of air trips by visitors are likely to be similar or proportional to those by residents.

Data from air passenger surveys performed at the same airport over time can provide some indication of whether the composition of the different types of trip appears to be fairly stable over time. The U.S. DOT 10% origin and destination (O&D) data can be analyzed to show the extent of the variability in the split of air trips at a given airport between residents and visitors over time as well as any overall trends. Since these data are available on an annual basis, they can be used to divide the total air passenger traffic at an airport into three components: outbound trips by residents, inbound trips by visitors, and connecting passengers. Since the latter are largely a consequence of airline network and hubbing strategies rather than the demographic and socioeconomic composition of the region where the hub airport is located, forecasting connecting traffic requires a different approach from that for O&D traffic that is beyond the scope of the current project.

Although all four approaches to incorporating disaggregated socioeconomic data in to air passenger demand models can be applied as well to modeling visitor trips as resident trips, the variables used may well be different, particularly for personal trips by visitors for purposes other than visiting friends and family. An important topic for future research is identifying appropriate variables for use in modeling visitor air trips. One obvious potential variable is the number of hotel rooms in a region. However, some care is needed in using such a variable, since it is not the number of hotel rooms that cause visitors to decide to make a trip to a particular region, but rather the number of rooms is a consequence of the number of visitor trips for purposes other than visiting family or friends. Thus while changes in the number of hotel rooms in the past may provide a good explanation of changes in the number of such visitor trips, this is of limited use for forecasting future air travel demand, since future growth in the number of hotel rooms is likely to depend on the growth in visitor trips, not the other way round.

In using the 10% O&D data to separate total traffic into the three directional components, care should be taken to make adjustments in the analysis for one-way trips. While these

undoubtedly do include some genuine one-way trips, most are an artifact of travelers purchasing two separate tickets, typically on different airlines to obtain a lower overall round-trip fare or more convenient flight schedules. Even for those that are genuine one-way air trips, it is not clear whether the traveler is a resident of or visitor to the region containing the first airport in the itinerary. It is therefore reasonable to assume that the proportion of one-way trips in each direction that are outbound trips by residents of the region containing the first airport in the itinerary reflect the directional split given by the round trip itineraries in the data.

E.4 References

Proctor, Bernadette D., Jessica L. Semega, and Melissa A. Kollar (2016). *Income and Poverty in the United States: 2015,* U.S. Census Bureau Report Number P60-256.