

# **ANALYSIS OF U.S. AIRLINE PASSENGERS' REFUND AND EXCHANGE BEHAVIOR ACROSS MULTIPLE AIRLINES**

**Dan C. Iliescu**

Ph.D. Candidate, School of Civil and Environmental Engineering  
Georgia Institute of Technology

790 Atlantic Drive, Atlanta GA 30332-0355

Phone: (404) 894-2255; Fax: (404) 894-2278; Email: gth657x@mail.gatech.edu

Submitted to the  
2006 AGIFORS Anna Valicek Competition

May 15, 2006

***Acknowledgement:*** Support for this research was provided by The Boeing Company. The author would like to acknowledge the support of Boeing, and particularly Roger A. Parker, who inspired several of the ideas in this work. The author would also like to acknowledge the support of his doctoral advisor, Laurie Garrow. However, all opinions, errors, and omissions are the sole responsibility of the author.

***Abstract:***

This paper uses individual ticket data from the Airline Reporting Corporation to model customer refund and exchange behavior across multiple U.S. airlines. A conceptual model describing how ticketing data relates to and can be integrated into traditional no-show and cancellation models is outlined. Survival models are used to predict the number and timing of cancellation (defined as ticketing refunds and exchanges). Distinct from other cancellation models, time is modeled “backwards” relative to the outbound departure date to provide a clearer interpretation of cancellation effects occurring as a function of days from flight departure. Empirical results provide new insights into airline passenger behavior and constitute one of the first published studies to be based on ticket clearinghouse data. This is of interest to airlines because as the use of the Internet has increased, traditional data sources (such as booking reservations made through travel agencies) are becoming less reliable.

# **ANALYSIS OF U.S. AIRLINE PASSENGERS' REFUND AND EXCHANGE BEHAVIOR ACROSS MULTIPLE AIRLINES**

## ***1. Introduction***

Airlines throughout the world are experiencing tremendous pressure to control costs while competing in a low-fare market that is being overtaken by low cost carriers. Multiple factors have contributed to the fact that since 2001, more than 50% of the U.S. airline capacity entered into bankruptcy. While some of the factors leading to bankruptcy are well-recognized and include high fuel costs, high labor costs, and increased market penetration of low-cost carriers, other factors are less understood (such as airline customers' willingness to pay to travel by air and willingness to pay for service amenities such as a non-stop flight). Further, developing a better understanding of customer behavior and demand is seen as critical to the next generation of revenue management, pricing and scheduling models. Perhaps this urgency is best summarized by Suresh Achara, a director of Manugistics, who states that "there is a lot of focus on very, very sophisticated and fancy optimization models, and that's great, but frankly, if you don't have the right demand model, if you just assume that you have the right demand value, then you're making the wrong assumption" (Achara, 2005). Consequently, as airlines develop business plans for emerging from bankruptcy, it is clear that the ability to incorporate customer behavior into these models will be critical to succeeding in a more competitive environment.

This paper explores one facet of airline passenger behavior, namely their refund and exchange behavior, using individual ticket data from the Airline Reporting Corporation (ARC). In traditional models, refunds and exchanges are reflected in airlines' no-show and cancellation models. Many airlines currently forecast no-show and cancellation rates using time-series or averaging methods. These models consider differences due to booking-specific information (such as booking class and number of days prior to departure that the booking was made) and/or flight-specific information (such as departure time, day of week, month, origin, destination, etc.). However, as described by Garrow and Koppelman (2004a, 2004b), to the extent that different types of passengers or itineraries (such as round-trips and one-ways or outbound and inbound flights) exhibit distinct no-show and cancellation rates, current models based on historical flight or booking information cannot make accurate predictions when the underlying passenger and/or itinerary mix change.

The key objective of this paper is to develop a model of airline passenger cancellation behavior based on the occurrence of exchanges and refunds in the ARC ticket data. Multiple origin-destination pairs are selected to examine differences in leisure and business markets as well as differences in market structure (*e.g.*, percentage of low cost carriers, number of itineraries available, direction of travel, distance between origin-destination pairs, etc.). Survival models are used to predict the probability an individual will "survive" until the next time period or "die" due to cancellation or exchange. Unlike many applications of survival models, this application is unique in that the exact times of

“birth” and “death” are known, as are the exact causes of death (refund, exchange to fly on a different date, exchange to fly on the same date with a different itinerary, etc.).

This paper is one of first to examine refund and exchange behavior across multiple carriers. It is also one of the first published studies to be based on ticket clearinghouse data from ARC. This is of particular interest to airlines because as the use of the Internet has increased, traditional data sources (such as Computerized Reservation System, CRS, data on booking reservations made through travel agencies) is becoming less reliable since it represents a smaller percentage of the market. Moreover, given the backbone of many current airline systems are based on CRS booking data, there is increasing interest from airlines in assessing the viability of using alternative data sources, such as the ARC data. Consequently, this study provides one of the first assessments of the benefits of using ARC data to model passenger refund and exchange behavior.

The remainder of this paper contains five sections. The first section describes the ticketing data used for the analysis and describes how it differs from ticketing data used in previous studies. Next, the business objectives are described as these have a direct impact on the modeling methodology. The third section describes the methodology while the fourth section presents empirical results; the emphasis of these sections is on describing and interpreting hazard models, as their use within the airline community has been relatively limited. Finally, on-going work and extensions to the analysis presented in this paper are described.

## ***2. Data***

This study uses ticketing data from the Airline Reporting Corporation (ARC). ARC is the ticketing clearinghouse for many airlines in the U.S. and essentially keeps track of purchases, refunds, and exchanges for participating airlines and travel agencies. While not all U.S. carriers or distribution channels are represented, ARC data is unique in the sense that individual ticketing transactions across multiple airlines can be observed. In addition, the use of ticketing data is of particular interest to airlines because as the use of the Internet has increased, traditional data sources (such as Computerized Reservation System (CRS) that contains data on booking reservations made through travel agencies) are becoming less reliable as they represents a smaller percentage of the market. For example, according to NetRatings (2005), nearly 50% of U.S. airline ticket sales were conducted exclusively on-line in the first half of 2005 and approximately half of these on-line sales (or 25% of all airline ticket sales) occur on airline suppliers' websites. Further, the majority of these on-line bookings made on an airline suppliers' website are not captured in CRS data. Moreover, given the backbone of many current airline systems in the U.S. are based on CRS booking data, there is increasing interest from airlines in assessing the viability of using alternative data sources, such as the ARC data. Consequently, this study will provide the first assessment of the viability of using ARC data to model passenger ticketing, refund and exchange behavior.

Because ARC is owned by the airlines, extensive discussions were required to determine a data format that could support modeling objectives while protecting airline confidentiality. Specifically, individual tickets are used for the analysis, but each airline code has been replaced by a randomly assigned number and flight information has been suppressed. The data used for this study contains simple one-way and round-trip tickets for which the outbound departure date occurred in 2004. As shown in Table 1, a total of eight directional markets are included in the analysis and reflect a mix of business and leisure markets and a mix of round trip and one ways. Each market is served by at least three airlines and contains non-stop and connecting itineraries. The markets include travel in origin destination pairs involving Miami, Seattle, or Boston (specifically, MIA-SEA, SEA-MIA, MIA-BOS, BOS-MIA, BOS-SEA, SEA-BOS) in addition to travel between Chicago O’Hare airport and Honolulu (ORD-HNL, HNL-ORD). Overall, 1.3% of the tickets are refunded and 1.2% are exchanged, but there are large differences across markets. MIA-SEA is particularly distinct due to the large number of consolidator bookings appearing in the market.

	# tickets	# (%) Refunded	# (%) Exchanged	# (%) One Ways	# (%) Round Trips
MIA-SEA	8,599	623 (7.2%)	84 (1.0%)	4,095 (48%)	4,504 (52%)
SEA-MIA	18,059	210 (1.2%)	198 (1.1%)	3,433 (19%)	14,626 (81%)
BOS-MIA	84,752	858 (1.0%)	1,248 (1.5%)	9,013 (11%)	75,739 (89%)
MIA-BOS	23,800	106 (0.4%)	318 (1.3%)	9,778 (41%)	14,022 (59%)
BOS-SEA	35,204	374 (1.1%)	423 (1.2%)	6,337 (18%)	28,867 (82%)
SEA-BOS	34,564	288 (0.8%)	442 (1.3%)	6,178 (18%)	28,386 (82%)
HNL-ORD	5,261	62 (1.2%)	51 (1.0%)	1,715 (33%)	3,546 (67%)
ORD-HNL	24,131	416 (1.7%)	138 (0.6%)	1,664 (7%)	22,467 (93%)
TOTAL	234,370	2,937 (1.3%)	2,902 (1.2%)	42,213 (18%)	192,157 (82%)

Table 1: Refund and Exchanges by Market and Trip Type

Ticketing information includes the issue date (or date the ticket was purchased), the outbound and inbound departure dates, outbound and inbound ticketing class (*i.e.*, first letter of the fare basis code), ticketing cabin code (*i.e.*, first, business, coach, other/unknown), net fare (*i.e.*, fare that does not include taxes and fees), and total tax and fees. From the outbound and inbound departure dates, several variables commonly used to segment customers into business and leisure segments can be derived including departure and return days of week, length of stay, and trips that include a Saturday night. Tables 2 to 8 show the influence of these variables on exchange and refund rates. As shown in Table 2, the timing of advance purchase differs for refunds and exchanges. Exchanges exhibit more of a “tub” shape for tickets purchased 8 to 360 days from the outbound departure, that is, tickets purchased far from departure or 2-3 weeks from departure are more likely to be exchanged than tickets purchased between these periods. Exchanges drop dramatically one week from departure. In contrast, refunds tend to increase as the advance purchase decreases, that is, tickets purchased closer to the outbound departure date are more likely to be refunded. However, at 31-40 days from

departure, there is a slightly higher percentage of refunds which may be attributed to consolidator bookings (such as air travel associated with cruise lines that are present in the Miami and Seattle markets). Similar to exchanges, the percent of refunds drops very close to departure, or 0-3 days from the outbound departure date.

Advance Purchase	Exchanges		Refunds		Exchange & Refunds		Total Tickets
0-3	33	0.151%	267	1.223%	300	1.375%	21825
4-7	245	1.155%	483	2.278%	728	3.433%	21205
8-14	478	1.731%	430	1.558%	908	3.289%	27607
15-21	426	1.593%	312	1.167%	738	2.760%	26738
22-30	410	1.323%	370	1.194%	780	2.517%	30988
31-40	333	1.175%	445	1.570%	778	2.745%	28344
41-50	236	1.077%	259	1.182%	495	2.260%	21904
51-90	421	1.397%	197	0.654%	618	2.051%	30126
91-180	259	1.559%	118	0.710%	377	2.269%	16618
181+	61	1.921%	56	1.763%	117	3.684%	3176

Table 2: Advance Purchase

Table 3 shows the percents of exchanges and refunds by trip type. Due to the very small percentage of exchanges that occur on one-way trips, it is suspected that there may have been an error in pulling the data for these trips. Discussions are currently underway with ARC to determine if this is a coding error. As such, the model estimates presented in this document exclude one-way trips.

Trip Type	Exchanges		Refunds		Exchange & Refunds		Total Tickets
Round trip	2877	1.497%	2410	1.254%	5287	2.751%	192,157
One way	25	0.059%	527	1.248%	552	1.308%	42,213

Table 3: Round Trips and One Ways

Tables 4 and 5 show how a Saturday night stay and the number of nights away from home impact exchange and refund rates. Round trip tickets without a Saturday stay (that tend to be associated with leisure travel) are less likely to be exchanged or refunded. Similarly, as the number of nights spent away from home increases (which would indicate the trip is more likely to be leisure), the exchange and refund rates for round trip tickets decrease.

Saturday Stay	Exchanges		Refunds		Exchange & Refunds		Total RT Tickets
Saturday Stay	1,401	1.092%	1,263	0.984%	2,664	2.076%	128,333
No Saturday Stay	1,476	2.313%	1,147	1.797%	2,623	4.110%	63,824

Table 4: Saturday Night Stays (Round Trip Tickets)

Nights Away	Exchanges		Refunds		Exchange & Refunds		Total RT Tickets
0-1	304	3.280%	271	2.924%	575	6.204%	9,268
2	487	2.685%	344	1.897%	831	4.582%	18,138
3	538	1.929%	295	1.057%	833	2.986%	27,897
4	439	1.673%	240	0.915%	679	2.588%	26,241
5	213	1.205%	168	0.951%	381	2.156%	17,673
6	134	1.158%	61	0.527%	195	1.686%	11,569
7+	597	0.902%	854	1.291%	1,451	2.193%	66,152
<b>TOTAL</b>	2713		2233				176,938

Table 5: Number of nights away from home (Round Trip Tickets)

The differences in exchange and refund rates between business and leisure travelers is also seen in Tables 6 and 7 that show the effect of outbound and inbound departure dates on refund and exchange rates. Exchanges are more likely to occur on Sunday, Monday, and Tuesday outbound departures and Wednesday, Thursday, and Friday inbound returns. Refunds exhibit a similar pattern, but also show a relative high rate on Saturday outbound departures (this again may be attributed to the presence of consolidator bookings which requires further exploration with ARC).

Day of Week	Exchanges		Refunds		Exchange & Refunds		Total RT Tickets
Sunday	390	1.219%	564	1.763%	954	2.982%	31,989
Monday	515	1.684%	494	1.616%	1,009	3.300%	30,575
Tuesday	461	1.790%	337	1.308%	798	3.098%	25,759
Wednesday	477	1.548%	342	1.110%	819	2.659%	30,806
Thursday	416	1.066%	316	0.810%	732	1.876%	39,017
Friday	375	0.922%	390	0.959%	765	1.882%	40,653
Saturday	268	0.753%	494	1.389%	762	2.142%	35,571

Table 6: Day of Week for Outbound Departure

Day of Week	Exchanges		Refunds		Exchange & Refunds		Total RT Tickets
Sunday	475	1.236%	394	1.025%	869	2.261%	31,989
Monday	317	1.099%	227	0.787%	544	1.885%	30,575
Tuesday	241	1.153%	217	1.038%	458	2.191%	25,759
Wednesday	408	2.095%	305	1.566%	713	3.660%	30,806
Thursday	448	2.415%	361	1.946%	809	4.361%	39,017
Friday	522	2.158%	413	1.707%	935	3.865%	40,653
Saturday	301	1.135%	316	1.191%	617	2.326%	35,571

Table 7: Day of Week for Inbound Departure (Round Trip Tickets)

Finally, Table 8 shows the exchange and refund rates by month of the outbound departure date and refunds and exchanges. No clear pattern can be detected, suggesting the effects of seasonality may be limited.

Departure Month	Exchanges		Refunds		Exchange & Refunds		Total RT Tickets
January	245	1.412%	266	1.53%	511	2.94%	17,357
February	233	1.041%	306	1.37%	539	2.41%	22,384
March	237	0.983%	277	1.15%	514	2.13%	24,108
April	226	0.966%	264	1.13%	490	2.09%	23,402
May	244	1.262%	199	1.03%	443	2.29%	19,332
June	251	1.325%	239	1.26%	490	2.59%	18,946
July	204	1.136%	244	1.36%	448	2.49%	17,961
August	186	1.029%	202	1.12%	388	2.15%	18,071
September	273	1.693%	325	2.02%	598	3.71%	16,124
October	279	1.532%	224	1.23%	503	2.76%	18,209
November	293	1.604%	167	0.91%	460	2.52%	18,267
December	231	1.143%	224	1.11%	455	2.25%	20,209

Table 8: Month for Outbound Departure

In addition to the variables described above, tickets that are refunded or exchanged also contain the date the refund or exchange was processed. In addition, when one ticket is exchanged for another ticket, information on the exchange fee and fare difference from the original ticket is available. Indicator variables are also populated to show the reason for the exchanged ticket. Specifically, indicators are used to know whether the customer requested (1) a new outbound and/or inbound departure date, (2) a new outbound and/or inbound ticketing class and cabin code, and/or (3) a new outbound and/or inbound itinerary.

The ticketing data used for this study is distinct from the data collected as part of the United States Department of Transportation (US DOT) *Origin and Destination Data Bank 1A or Data Bank 1B* (commonly referred to as DB1A or DB1B). The data are based on a 10 percent sample of flown tickets collected from passengers as they board aircraft operated by U.S. airlines<sup>1</sup>. The data provide demand information on the number of passengers transported between origin-destination pairs, itinerary information (marketing carrier, operating carrier, class of service, etc.), and price information (quarterly fare charged by each airline for an origin-destination pair that is averaged across all classes of service). While the raw DB datasets are commonly used in academic publications (after going through some cleaning to remove frequent flyer fares, travel by airline employees and crew, etc.), airlines generally purchase Superset data from Data Base Products. Superset is a cleaned version of the DB data that is cross-validated

<sup>1</sup> “The raw materials for the Origin-Destination survey are provided by all U.S. certificated route air carriers, except for a) helicopter carriers, b) intra-Alaska carriers, and c) domestic carriers who have been granted waivers because they operate only small aircraft with 60 or fewer seats.” (Data Base Products, 2006).

against other data-sources to provide a more accurate estimate of the market size. See the Bureau of Transportation Statistics website at [www.bts.gov](http://www.bts.gov) or the Data Base Products, Inc. website at [www.airlinedata.com](http://www.airlinedata.com) for additional information.<sup>2</sup>

Data based on the DB tickets differs from the ticketing data obtained from ARC for this study in three important ways. First, DB data reports aggregate information using quarterly averages and passenger counts while ARC data contains information about individual tickets. Second, DB data contains a sample of tickets that were used to board aircraft, or for which airline passengers “show” for their flights. In contrast, ARC data provides information about the ticketing process from the *financial perspective*. Thus, historical information is available for events that trigger a cash transaction (purchase, exchange, refund), but no information is available for whether and how the individual passenger used the ticket to board an aircraft; this information can only be obtained via linking with the ARC data with airlines’ day of departure check-in systems. Finally, ARC ticketing information does not include changes that passengers make on the day of departure; thus, the refund and exchange rates will tend to be lower than other rates reported by airlines or in the literature.

Given an understanding of the ticketing data used for this study, the next section describes the key business objectives, as these have a direct impact on the research design.

### ***3. The Business Context and Conceptual Model***

From a business perspective, this research needs to integrate into the larger effort of Boeing Commercial Airplanes (BCA), the commercial products arm of The Boeing Company. Specifically, BCA has been engaged in a research effort to advance its models of passenger behavior. These models are a central part of the tools used by its marketing department to help potential airline customers estimate how much market share and revenue can be gained via the introduction of new service and equipment in a market. One of the core components of the passenger behavior models under development is the Universal Market Simulator (UMS) shown in Figure 1. The UMS is a Monte Carlo micro-simulation of airline revenue generation whose primary output is the revenue to an airline that results from the individual choices of thousands of passengers moving over a world-wide airline network. The UMS uses several models to represent different aspects of passenger behavior and airline competitive responses including models for synthetic population generation, induced demand, booking and ticketing curves, ticket cancellations, passenger itinerary choice, and airline revenue management models (Parker et al., 2004).

The UMS is designed to simulate passenger behavior and airline competitive responses on a global network. Unlike individual airlines, however, BCA does not have access to the same data (such as an airline’s bookings and check-in information). Further, the quality and ability to obtain data varies throughout the world. Thus, in practice, when BCA works directly with an airline to help assess the revenue and market share impacts

---

<sup>2</sup> The website describes the data and federally-mandated reporting requirement for U.S. airlines.

of introducing new equipment in a market, it is common to validate and adjust Boeing's forecasting input to more accurately reflect the individual airline's revenue management system, business process, and expectations of future market conditions. (For a flavor of the variability in airline policies related to ticketing, exchange and standbys, see the memo distributed by American Express, 2005). Consequently, given that booking (and not ticketing) data is the core of airlines' revenue management systems, it is essential that the refund and exchange models developed in this study can be integrated into the UMS (specifically linked to cancellation and day-of-departure no-show and standby models) and be adjusted based on an individual airline's business practice.

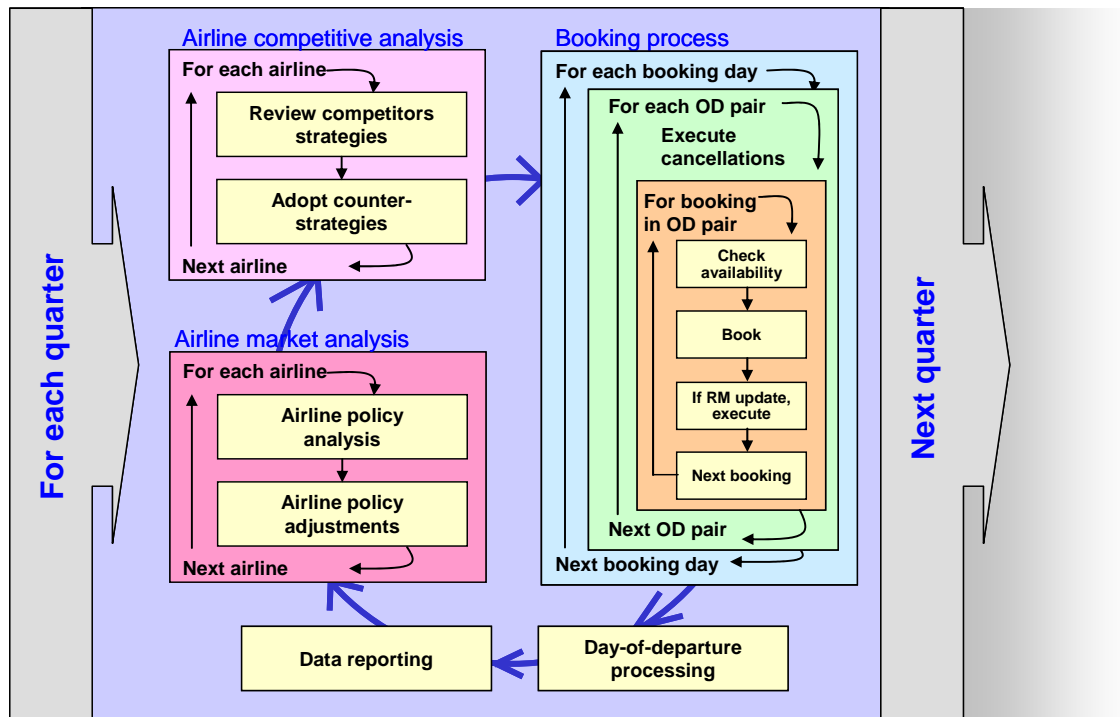


Figure 1: Boeing's Universal Market Simulator (Parker, et al. 2004)

Figure 2 shows the relationships among bookings, tickets, cancellations, no-shows, and standbys from the perspective of a typical airline's revenue management process. Revenue management is used to decide how many seats to allocate for sale to customers (prices for these seats are generally set outside the allocation decision and based on competitive market conditions). Booking data form the basis for these allocation decisions. Specifically, an airline decides the number of seats to sell based on forecasts of how many current bookings will show for flights and how many future bookings will show for flights. Cancellation, no-show, and standby models are used to predict the number of bookings that will show for flights. These forecasts differ depending on when the airline knows the passenger does not intend to travel on the ticketed flight. As described in Garrow and Koppelman (2004a), cancellation models predict how many passengers inform the airline they do not intend to travel prior to the departure of their flights. Some airlines further distinguish between bookings that are made and cancelled within a short time period (like 24 hours) and bookings that are cancelled outside this time period. If a booking is made and cancelled within a 24 hour period (which generally

corresponds to the amount of time after a booking is made that it must be ticketed), it is generally removed from the demand forecast as part of a “booking churn” model so it does not influence cancellation rates. There are several reasons for booking churn that include travel agencies making duplicate bookings or making bookings to hold inventory; for example, American Airlines defines churn as “any cancel/rebook activity intended to circumvent ticketing time limits or hoard inventory” (American Airlines, 2005).

In contrast to cancellation models, no-show models estimate the number of remaining booked passengers, *i.e.*, passengers who have not cancelled, but fail to show for their flights. Standby models are used to predict the number of passengers who arrive to the airport but take a flight on the ticketing carrier that is different than the one they purchased. The benefit of identifying standbys lies in the recognition that as flights become full, opportunities to standby on different flight decrease and passengers are not able to standby for alternate flights but rather “show” for the flight they purchased (Garrow and Koppelman 2004a, 2004b).

Given the business objectives supporting this analysis, it is important to understand how ticket information from ARC relates to the show, no-show, cancellation, and standby models used by airlines that are based on booking data. From an airline perspective, show, no-show, standby, and cancellation rates are influenced both from bookings that are ticketed (*i.e.*, paid for) as well as bookings that are not ticketed (*i.e.*, reservation made but never paid for). As shown in Figure 2, bookings that are not ticketed ultimately become a churn booking, a cancellation, or a no-show. In contrast, bookings that are ticketed can end up in one of four states, namely a cancellation, a no-show, a standby, or a show. These states can be related to the ARC and DB ticketing data by noting the different ways these states can occur. For example, a cancellation that triggers a financial transaction occurs when a passenger informs the airline prior to departure that she does not intend to take the ticketed flight. In this case, the original booking that was purchased is cancelled from the revenue management system and a new booking (and ticket transaction) is created for the new flight(s) she purchases. These transactions appear in ARC data. However, there are also ticketed bookings that are cancelled that will not appear in ARC data. For example, some airlines use automated data processes that cancel the inbound flights of an itinerary if the passenger no-shows on the outbound flights. In this case, the outbound flights that were never used or exchanged prior to departure become no-shows and the carrier automatically cancels in the inbound flights (without generating an automatic refund / exchange transaction). However, from the perspective of assessing the revenue generation to an airline, these cancellations are not relevant to the business question at hand as the airline still receives the revenue from the original ticket transaction.

Similar to cancellations, there are two ways a ticketed booking can become a no-show and only one of these cases will appear in the ARC data. No-shows that occur due to exchanges or refunds requested after the flight departure are captured in ARC ticketing data; however, no-shows that occur when an individual purchases a ticket yet never uses it or requests a refund are not captured in the ARC data. As before, these no-shows are not relevant as the airline receives the revenue from the original ticket transaction.

In contrast to no-shows and cancellations, a show occurs when a ticket was used exactly as purchased. This “snapshot” of tickets is what is captured in the lifted tickets collected in the DB data collected by the U.S. DOT. Finally, it is important to note that changes to tickets that occur on day of departure for the flight are not captured in the ARC data, but rather are part of an individual airline’s check-in processing.

To summarize, unlike DB ticketing data, the use of ARC ticketing data provides an opportunity to develop no-show and cancellation models. In addition, these models can be developed for multiple airlines and/or markets. Further, given the large percentage of bookings in revenue management systems are ticketed in U.S. markets, it is expected that the ARC data will be representative of overall no-show and cancellation rates. Most important, the no-show and cancellation rates developed in this study *directly tie to the revenue generation stream of an airline* which is one of the most important metrics to an airline considering aircraft purchases. Finally, through understanding the underlying causes driving these rates, they can be readily adjusted if needed when working directly with a carrier’s data.

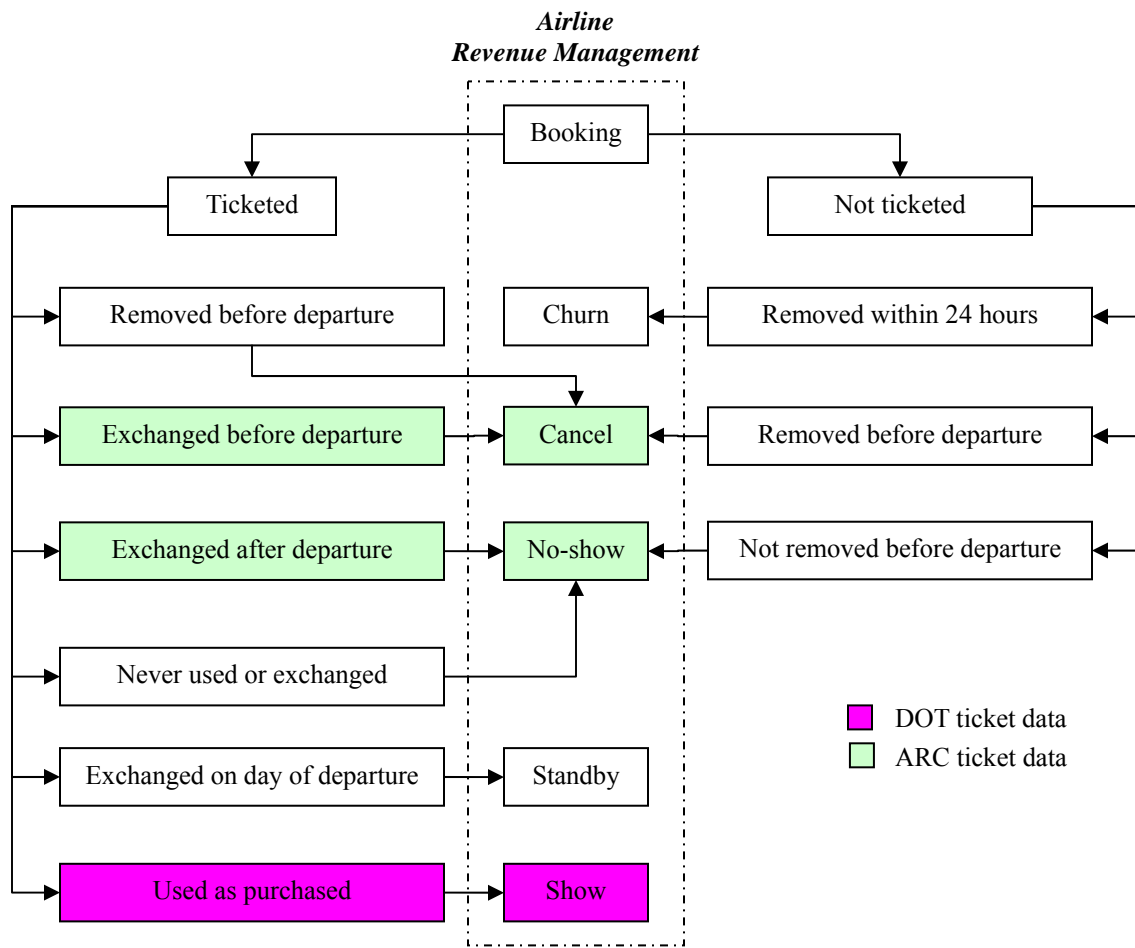


Figure 2: Relationships among Bookings, Tickets, Cancellations, and No-shows

#### 4. Methodology

Given an understanding of how ticketing refunds and exchanges relate to airline no-shows and cancellations, this section describes the study methodology. Survival models are used to predict the both the number of tickets that are refunded or exchanged and the timing of these refund and exchange events. Survival models differ from the time series smoothing or attribute-based cancellation forecasting methods commonly used in the airline industry; for examples of these models see Chatterjee (2001), Polt (1998), and Ratliff (1998). The survival models in this study are similar in spirit to the cancellation model proposed by Westerhoff (1998) shown in Figure 3. In Westerhoff's model, a booking that exists at time  $t$  survives until period  $t+1$  with probability  $p_t$  and is cancelled with probability  $(1-p_t)$  for  $t >$  departure day of the flight. This is also similar to a cancellation model proposed by Van Ryzin and Karaesmen (1999) that predicts the probability a booking will survive to the next time period as a random event according to a Binomial or Poisson process.

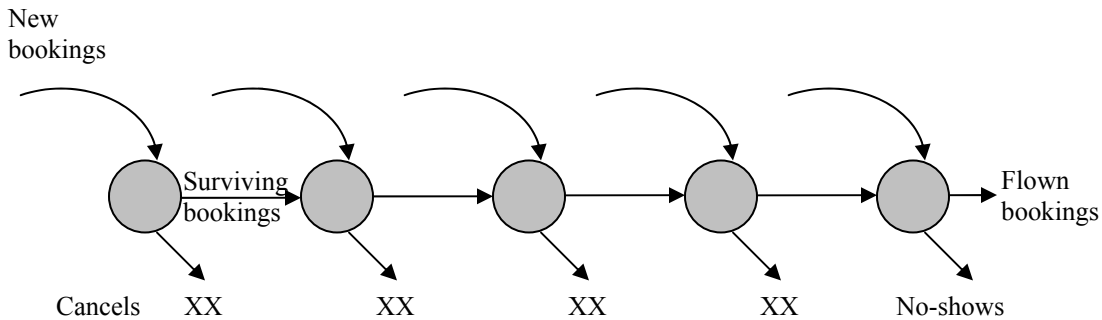


Figure 3: Cancellation Model Proposed by Westerhoff (1998)

However, this work differs from Westerhoff's model in two important ways. First, it extends the formulation above by examining the joint effects of multiple covariates affecting refund and exchange events within the ticketed population and by allowing these covariates to vary with time. Second, time is modeled "backwards" relative to the outbound departure date (versus "forward" relative to the booking date) in order to provide a clearer interpretation of cancellation effects as a function of the number of days before flight departure.

In order to better understand the flexibility of survival models and how they differ from the previous methods of forecasting cancellations reported in the literature, Section 4.1 provides an introduction to survival analysis (also known as "time to event" or "current status" analysis) while Section 4.2 describes how survival models are applied to the problem of interest in this study.

##### 4.1 General Concepts of Survival Analysis

Historically, statistical methods for current data developed in the medical field and were applied to epidemiological applications (that capture the time-to-occurrence of an event

given exposure to an infection) or clinical applications (that capture the time-to-occurrence of an event given exposure to treatment). The fundamental difference between the two categories of studies is defined by the way in which survival time is considered – either in retrospective or prospective (Kim & Lagakos, 1990). In retrospective studies, investigators analyze the disease incidence for exposed individuals “in hindsight” based only the prevalence of disease at the time the data is collected (Becker, 1989; Shiboski & Jewell, 1992). In contrast, in prospective studies investigators use a “forward looking” approach to analyze the evolution of disease for individuals exposed to various treatments (Hosmer & Lemeshow, 1999). Although the applicability of survival analysis arises naturally for the medical field, it has been used in a wide range of contexts spanning demography, econometrics, transportation, etc. For a comprehensive review of survival analysis applications and how they fit in the general context of generalized additive models, see Shiboski (1998). Since the taxonomy of survival analysis was developed outside of the airline industry, this section will introduce the main concepts related to these models.

Figure 4 portrays an example of a survival process for refund and exchange occurrences using a prospective or “forward looking” approach. All three tickets have the same departure date, but differ in when they were purchased. Using this “lifetime data” that records the beginning and ending times of a process, survival analysis seeks to answer two fundamental questions: (1) what is the population proportion that survives past a certain time (survival function), and (2) of those that survive, what is the rate at which they will they fail (hazard function).

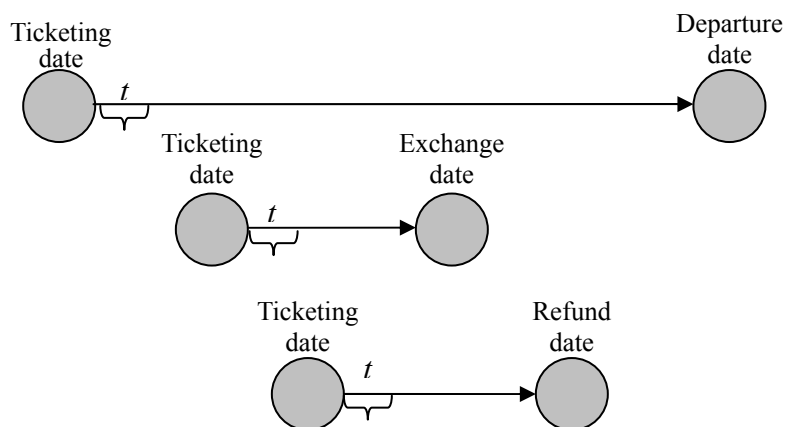


Figure 4: Example of Prospective Survival Analysis for Refunds and Exchanges

Formally, survival and hazard functions are used to describe this process. The general forms for survival  $S(t)$  and hazard  $h(t)$  functions are presented in Equation 4.1 (for continuous time) and Equation 4.2 (for discrete time).<sup>3</sup> In both equations,  $T$  is defined as a continuous (discrete) non-negative random variable representing the time until an event and is distributed according to pdf (pmf)  $f(t)$ .

<sup>3</sup> In the case of the continuous time the hazard rate represents an **instantaneous rate of occurrence**, in the case of discrete time is a **conditional probability** (defined later in this section).

$$S(t) = \Pr\{T > t\} = 1 - F(t) = \int_t^{\infty} f(x)dx \quad (4.1)$$

$$h(t) = \frac{f(t)}{S(t)}$$

$$S(t_j) = S_j = \Pr\{T \geq t_j\} = \sum_{k=j}^{\infty} f_k \quad (4.2)$$

$$h(t_j) = h_j = \Pr\{T = t_j \mid T \geq t_j\} = \frac{f_j}{S_j}$$

Given lifetime data defined by a survivor function  $S(t)$  or hazard function  $h(t)$ , the likelihood function for observation  $i$  can be expressed as the product of the two functions. Furthermore, under the assumption that censoring happens randomly in the population, the general likelihood function can be simplified using Equation 4.3 (Hosmer and Lemeshow, 1999). The function is derived by noting that under the random censoring assumption, the contribution to the likelihood function will be equal with the survival time cdf  $S(t)$ .

$$L = \prod_{i=1}^n L_i = \prod_i h(t_i)^{c_i} \cdot S(t_i) \quad (4.3)$$

where  $c_i$  represents a failure indicator taking the value of one for observations with materialized events during the observation time and the value of zero for observations with censoring or non-events.

Note that survival times can be modeled indirectly using a hazard rate. The hazard rate is defined as “the instantaneous death rate” or the probability of changing the current state from alive to dead at each time period given the survival experience up to that point. To illustrate the differences between survival and hazard functions when using a discrete versus a continuous time formulation, *Appendix A* presents partial estimation results for a sample of ten ARC tickets.

There are a variety of subjective assumptions related to the survival process that influence the appropriate choice of a survival model. These assumptions include definitions of the population at risk, the beginning and end of an observation, the censoring mechanism, whether the population is homogenous with respect to survival experience, the shape of the survival time distribution, and conditional versus unconditional analysis.

The *population at risk* is defined as independent “subjects” under observation during parts or the entire period of the survival study. The *beginning of an observation* is uniquely identified by the time at which the subject becomes at risk of failure during the period of study. In contrast, *the end of an observation* can be identified either by the time at which the failure of the subject is observed (non-censored) or by the time at which the subject observation ends due to maximum observation time or to the subject being lost in

the follow-up process (censored). Also, the typical taxonomy used to describe the time interval in which an observation is at risk is *spell length*. This study proposes a new definition of the beginning and end of observations, which is detailed in Section 4.2.

The main focus of the survival analysis changes depending on assumptions about *population homogeneity* with respect to survival experience. Conceptually, if the assumption of population homogeneity holds, then the lifetimes of all subjects are governed by the same survival function  $S(t)$ . In this case, the main focus of survival analysis is on determining the appropriate shape of the survival and hazard functions. However, if the assumption of population homogeneity does not hold, the focus of the analysis is expanded to include an exploration of the influence of a vector of covariates on survival time. Further, the way in which the vector of covariates impacts population survival process is used to define two main categories of models – proportional hazard models (PH) and accelerated failure time models (AFT). The fundamental difference between the two is that while AFT models coefficients represent changes in survival time due to a unit change in a given covariate, PH models coefficients represents changes in the hazard rates due to a unit change in a given covariate. As will be demonstrated in the preliminary results, the assumption of population homogeneity does not appear to hold in cancellation models (*e.g.* tickets with a Saturday night stay exhibit different survival functions than tickets without a Saturday night stay). Consequently, this forces us to explore the effects of multiple covariates using more complex PH and AFT models.

It is important to note that survival and hazard functions are alternative ways of representing the failure process. As such, assumptions about the *survival time distribution* are closely related to assumptions about the hazard function. Further, the function form of the hazard function is closely related to whether the survival process is assumed to evolve along a continuous or discrete time scale. The decision to model survival times as discrete or continuous depends on the general characteristics of the underlying survival process and the way in which survival data was collected. As a general rule, it is appropriate to use a continuous time scale when the ratio between the length of an observation (defined as the difference between the beginning and end of the observation) and the length of the time interval used for grouping the observations is high.

The assumption about whether observed survival times are continuous or discrete also influences the modeling approach. If time is discrete, then a discrete model with a *logit* link will be the most appropriate in terms of the ease of interpretation<sup>4</sup>. For example, if a ticket is purchased five days from departure and cancelled two days from departure (DFD), three logits would be “linked” together to express the probability of surviving when transitioning from 5 to 4 DFD, 4 to 3 DFD, and 3 to 2 DFD. On the other hand, if time is continuous but the observations are interval censored (as would be the case if a patient were examined every three months), then a discrete-time model based on a *c-log-log link* could provide more robust results. Finally, if the time is continuous, one could either explore parametric formulations of the hazard function or focus on estimating the

---

<sup>4</sup> For example, Cox’s (1972) extension to proportional hazards model is referred to as a *proportional odds model* where the odds refer to hazards.

effects of the covariates without making any assumption about the baseline hazard function<sup>5</sup> (Cox, 1972). Both discrete and continuous time specifications are explored in the preliminary results.

Finally, the choice of *a conditional or unconditional survival analysis* is determined by the characteristics of the events of interests with respect to the failure and censoring mechanism. If we assume that all subjects in the population will eventually experience the “failure” event, then hazard and survivor functions are defined as conditional. On the other hand, if there are observations for which the event of interest either ends on a certain “non-event” or censoring is present, then the hazard and survivor functions can still be computed, but will be defined as unconditional<sup>6</sup>. Stated another way, conditional is synonymous with “only those observations for which failure is present” (which in the study context would be only those tickets for which refunds and exchanges occur) while unconditional is synonymous with “considering all observations” (which in this study would be all tickets). This study is based on unconditional survival analysis (*i.e.*, uses all of the tickets).

#### ***4.2 Survival Analysis Applied to the ARC Ticketing Data***

In the context of this analysis, the study period was defined by the entire year of 2004, with each ticket considered as an independent observation (subject). The follow-up period for subjects was set up from a retrospective reference as shown in Figure 5, with the Outbound Departure Time (ODT) defined as the point in time when subjects enter the population at risk and the Refund Time/Exchange Time (RT/ET) or Issue Time (IT) defined as the points in time when subjects leave the population at risk (transition from one state to another). *Although one could argue that defining survival time in this manner is artificial, such a formulation represents cancellation behavior and provides a framework for interpreting results that is consistent with the “days from departure” concept used by the airline industry.* That is, in Figure 5, results are reported relative to the ODT. Intuitively, the cumulative survival function at time  $t$  represents the unconditional probability a ticket will not be refunded or exchanged 0 to  $t$  days from departure. This is in contrast to the representation in Figure 4 which defines the cumulative survival function relative to the issue date, *i.e.*, as the probability a ticket will not be exchanged or refunded 0 to  $t$  days after it was purchased.

Also, distinct from many other applications of survival analysis, this application is unique in the sense that in the ARC database, *censored observations are completely missing*. This is because all observations end in either failure (exchange/refund) or a certain non-

---

<sup>5</sup> In a semi-parametric framework the hazard function can be decomposed into the influence of the survival function and the influence of covariates. The general relationship is described in the following and identifies the Proportional Hazards (PH) Cox model:

$$h(t, x, \beta) = h_0(t) \cdot \exp(x' \beta)$$

$$r(x, \beta) = \exp(x' \beta)$$

Where the baseline hazard function is  $h_0(t)$  and is identified by the “null” vector of covariates  $r(x=0, \beta)=1$ .

<sup>6</sup> The definition of conditional observations should not be confused with the hazard rate, which by definition in the conditional probability of failure at time  $t$  given the individual has survived to time  $t$ .

event (issue date). In addition, only those exchange or refund events that occur prior to the outbound departure date are considered as part of this analysis; in typical airline models, this would correspond to a cancellation model for the legs on the outbound legs of the itinerary. The extension to include exchanges and refund events that occur between the outbound and inbound departure dates is straightforward, but is excluded from this analysis.

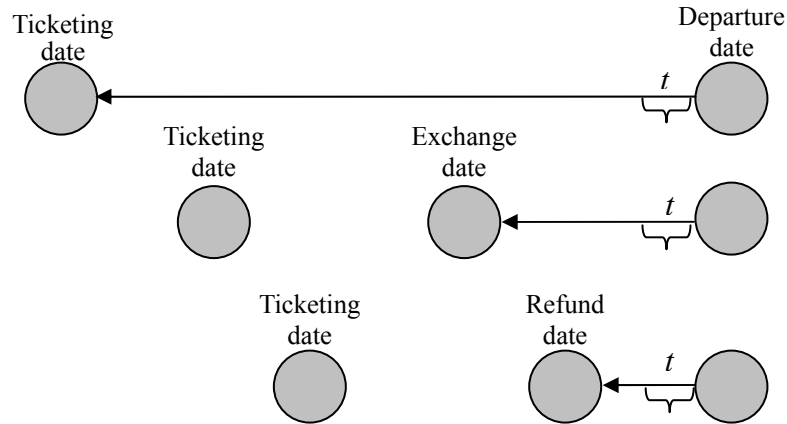


Figure 5: Retrospective Survival Analysis Structure Used in Study

This formulation associates the survival process to the entire ticketed population and “observes” the occurrence of “failures” (*i.e.*, tickets with exchanged or refunded events) from the ODT to the RT/ET or IT. The distribution of cancelled tickets over time can be described either by a survival function or a hazard function as given in Equation 4.1 or 4.2. Although similar in spirit, the two functions have different interpretations. While the hazard function represents the conditional probability of a ticket being cancelled during the time interval  $\Delta t$  given survival history up to that point, the survival function represent the unconditional probability of a ticket being cancelled during the time interval  $\Delta t$ .

Since refund and exchange events for tickets are observed on a daily basis and survival times are much larger than one day, analysis time was first modeled as a continuous random variable. In this framework, the proportional hazard assumption of Cox (PH) model was tested using a business-leisure segmentation of tickets. Since this assumption did not hold and one would still like to retain estimation results as tractable as possible, an alternative discrete specification of time was proposed with a Discrete-Time Proportional-Odds (DTPO) model. The DTPO model was estimated using a fully non-parametric base-line hazard function. The following paragraphs will describe the preprocessing of ARC tickets in terms of variable creation and methodological differences between the PH and DTPO models.

The first step of the ARC database preprocessing was to create a variable to characterize the classical segmentation business/leisure of passengers. Since passenger trip purpose cannot be defined based only on the higher categorization of cabin types (*i.e.*, economy/coach, business, first), a proxy representation of the underlying behavioral characteristics in terms of *Saturday Night Stay* and *Length of Stay* was used. The newly

created variable (*BvsL*) segmented the ARC tickets into three main categories: (1) business passengers tickets (associated with lengths of stay less than seven days and no Saturday Night stay); (2) leisure passengers tickets for short trips (associated with Saturday night stay and length of stay less than seven day); (3) leisure passenger tickets for long trips (associated with Saturday night stay and length of stay longer then seven days). The second step of the ARC database preprocessing was to create categorical variables capable to capture the seasonality of cancelled tickets in terms of the inbound departure day of the week (*InDOW*), outbound departure day of the week (*OutDOW*) and departure month (*DepMonth*). Finally, the third step of the ARC database preprocessing was to create a flag variable indicating whether a ticket was cancelled during its lifetime. Due to small percentages of refund and exchange events in the population (1.65% and 1.13% representing 1,248 and 858 tickets, respectively in the largest Boston-Miami market) both refund and exchange events were combined. Variables were also created to explore the effects of a dominant market carrier (*CarrierGrp*), pro rated ticket fare (*TktgFarePro*), and ticket advance purchase (*APurchase0\_3*, *APurchase3\_7*, *APurchase7\_21*, *Purchase21\_30*, *APurchase30\_45*, *APurchase45\_60*, *APurchase60\_90*, *APurchase90Plus*). In the context of the considered vector of covariates, the following paragraphs will describe the process used to perform exploratory analysis in addition to the two models estimated: a Cox Proportional Hazard (PH) model, and a Discrete Time Proportional Odds (DTPO) model.

As a preliminary step in survival analysis, the exploration of characteristics of survival and hazard functions across the considered vector of covariates is valuable in assessing the appropriate modeling approach. Two questions are of particular interest: (1) whether heterogeneity in survival experience is influenced by the vector of covariates, and (2) whether covariates contribution evolves proportionally with respect to hazard (PH models) or survival function (AFT models). Both questions can be addressed using non-parametric estimators of the survival (Kaplan-Meier) and hazard function (lowess smooth of hazard point-wise estimates).

The Kaplan-Meier estimator of survival functions is available in the majority of statistical software packages and can be computed using Equation 4.4 where  $d_j$  represents the number of failures and  $n_j$  the number of observations at risk at time  $t_j$ .

$$\widehat{S}(t) = \prod_{j:t_j < t} \frac{n_j - d_j}{n_j} \quad (4.4)$$

The formal test of survivor function equality across different strata is based on a contingency table which essentially follows the state of groups at each observed survival time (similar in spirit with Kaplan-Meier estimator) and creates an appropriate “contribution to the test statistic” assuming the survivorship function is the same for each group. As stated by Hosmer and Lemeshow (1999), the contribution to the test statistic depends on which of the various tests is used, but each may be expressed in the form of a ratio of weighted sums over the observed survival times.

In addition to Kaplan-Meier estimates, lowess smoothers of point-wise estimates are often employed in exploratory analysis. Lowess smoothers are based on the concept of “continuous histograms” of the conditional probability of failure. Histogram time intervals are reduced to point-wise estimates of the hazard ( $d_i/n_i$ ) and obtained values are then smoothed (typically using lowess) to obtain a graphical representation of the hazard function.

After the preliminary exploration, Cox Proportional Hazard models are typically estimated. The family of PH models assumes that relative to a baseline hazard function the influence of covariates develops proportionally and covariates are independent of survival time. The main advantage of such a formulation is that it permits semi-parameterization of the survival model, for which assumptions about the “error component of the model are unnecessarily stringent” and “desired inferences are based solely on the parameters in the systematic portion of the model” (Hosmer and Lemeshow, 1999). In that perspective in the class of PH models, the Cox model stands unique in providing only estimates of the “ $k$ ” parameters of covariates ( $\beta_1, \beta_2, \dots, \beta_k$ ) and no direct estimate for the baseline hazard function  $h_0(t)$ . However, the baseline hazard function can be recovered using point-wise estimators of conditional survival probabilities which have the disadvantage of being noisy or unstable with respect to derived confidence intervals (Hosmer and Lemeshow, 1999).

In cases where the proportional hazard assumption does not hold and time can be modeled discretely, an alternative formulation can be explored using the discrete time proportional odds (DTPO). DTPO models are similar in spirit with logistic regression in the sense that conditional odds are used to represent the odds of dying at time  $t_j$  given survival up to that point. The general formulation of the model is presented in Equation 4.5 where  $h(t_j | x_i)$  is the hazard at time  $t_j$ ,  $h_0(t_j | x_i)$  is the baseline hazard at time  $t_j$ , and  $\exp(x_i' \beta)$  is the relative risk associated with covariates values  $x_i$ . An empirical example of this model is provided in Appendix A.

$$\frac{h(t_j | x_i)}{1 - h(t_j | x_i)} = \frac{h_0(t_j | x_i)}{1 - h_0(t_j | x_i)} \exp(x_i' \beta) \quad (4.5)$$

In this study, the DTPO model was estimated using logistic regression on a set of generated pseudo-observations (expanded dataset). Each ticket observation was expanded to include an observation for each discrete time in present during the life of the ticket. For example, if the original ticket had a lifetime of 10 days that ended in failure (either a refund or exchange event), the expanded dataset would have 10 observations. These observations would be identical except for two variables, one describing the current day and one describing the current survival state. The immediate advantages of such a formulation are its flexibility in accommodating time-varying covariates and its flexibility in specifying the baseline hazard. Moreover, since the interpretation of estimated coefficients is done in terms of relative odds, this specification is particularly appealing.

## 5. Results

In the current research out of the eight available markets, Boston-Miami (BOS-MIA) was the only market for which the two model specifications (PH and DTPO) were explored. BOS-MIA was selected because it contains the largest number of tickets (83,504) and has a good representation of refunds and exchanges events (858 and 1,248 tickets). Since there were reasons to suspect that exchanges were not coded correctly on one-way trips<sup>7</sup>, as a preliminary filtering procedure the one-way tickets and round trip tickets with the inbound departure date missing were excluded from further analysis. Also for reasons mentioned in Section 4.2, ticket exchanges and refunds that happen after the outbound departure date were excluded from the analysis. After taking into account both considerations, the initial dataset was reduced to a total of 73,396 tickets out of which 1,215 had exchange events and 827 had refund events within the IT-ODT interval.

### 5.1 Exploratory Analysis

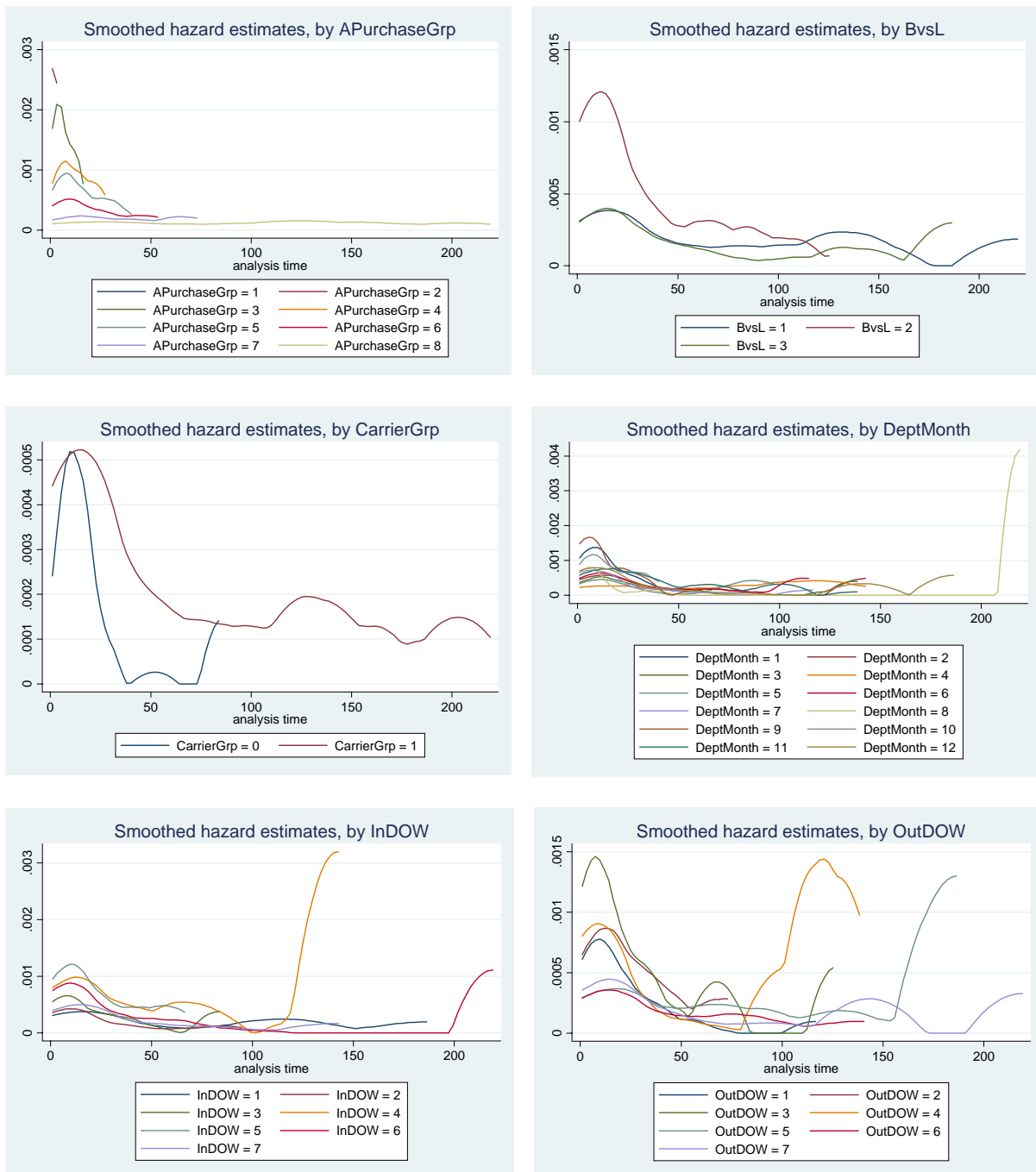
When conducting a survival analysis, exploratory analysis is generally used as a preliminary step to help assess the relative influence of different covariates on the survival and hazard functions. In this study, the Kaplan-Meier estimates are first used to determine if there are any difference in survival functions across different categorical covariates (*e.g.*, do leisure and business travelers have different survival functions?). Next, the validity of the proportional hazard assumption is explored using graphical techniques based on the lowess smoother of hazard point-wise estimates.

Kaplan-Meier estimates of the covariates described in Section 4.2 (business-leisure, advance purchase categories, outbound day of the week, inbound day of the week and dominant carrier) were used to test the equality of the survivor function. All tests (log-rank, Wilcoxon, Tarone-Ware, and Peto-Peto-Prentice) resulted in rejection of the null hypothesis that survivor functions are the same.

Figure 6 presents the graphical representation of the hazard function smoothers for these covariates. Visual inspection of the graphs suggests that the proportional hazard assumption will be violated for the majority of considered covariates. A possible exception is with the advance purchase (*APurchase*) variable. However, even in this case the shapes of the hazard function smoothers suggest that a segmentation based on tickets purchased 0-45 days from departure versus tickets purchased 46 or more days from departure would be more appropriate.

---

<sup>7</sup> One-way tickets had practically 100% chances of survival with only one exchange event.



**Legend**

*APurchaseGrp* = **1**- 0 to 3; **2**-3 to 7; **3**- 7 to 21; **4**- 21 to 30; **5**-30 to 45; **6** – 45-60; **7** -60 -90; **8**- more then 90  
*BvsL* = **1**-leisure short trips (< 7 days); **2**- business trips; **3**- leisure long trips (>7days)  
*CarrierGrp* = **0**-other carriers; **1**- leader of market carrier  
*DeptMonth* = **1**-January; **2** – February; ...; **12**- December  
*InDOW, OutDOW* = **1**- Sunday; **2** - Monday; ...; **7**- Saturday

Figure 6: Smoother Estimators of Hazard Function for Categorical Covariates

## ***5.2 Cox Proportional Hazard Model***

Two Cox Proportional Hazard model specifications were estimated. First, all covariates included in the exploratory analysis were included. Tests using the Schoenfeld scaled residuals of the covariates showed that the proportional assumption was violated. Next, a second Cox Proportional Hazard model specification was estimated using only those covariates for which the proportional hazard assumption that was “reasonably” satisfied.

The Cox Proportional Hazard model including all covariates is shown in Table 9. The bolded entries, which include inbound and outbound departure day of week, prorated fare, departure month, and trip purpose, are violating the proportional hazard assumption. Table 10 presents the Cox Proportional Hazard model that includes only advance purchase, dominant carrier code, and the implied trip purpose segmented by the presence of a Saturday night stay.

As shown in Table 10, the hazard function increases relative to the reference of 0-3 days from departure up until 21 days from departure and then begins to decrease. Further, the hazard function is not statistically different from zero at many of the advance purchase categories. However, trips with a Saturday night stay are twice as likely to have a refund or exchange event relative to trips without a Saturday night stay. In addition, for this market, passengers traveling on the dominant carrier are more than twice as likely to have a refund or exchange event relative to passengers traveling on non-dominant carriers. Figure 7 shows the shape of the survival and hazard functions using the mean values of the covariates. As clearly seen in the Figure, there is no obvious parameterization of the baseline hazard function with these covariates. Therefore, due to difficulties in satisfying the proportional hazard assumption, results of a discrete time proportional odds model is presented in the next section.

	Parameter estimate (z-stat)	Test of PH assumption
<b><i>Advance Purchase (Reference = 0-3 days)</i></b>		
3 to 7	2.81 (2.20)	0.7562
8 to 21	3.69 (2.87)	0.8328
22 to 30	3.15 (2.51)	0.8656
31 to 45	2.82 (2.27)	0.9144
46 to 60	1.81 (1.28)	0.8279
61 to 90	1.10 (0.21)	0.7914
90+	0.70 (-0.77)	0.9118
<b><i>Carrier Code (Reference = Not dominant carrier)</i></b>		
Dominant carrier	2.04 (6.51)	0.2602
<b><i>Inbound Departure Day of Week (Reference = Wednesday)</i></b>		
Sunday	0.734 (-3.06)	<b>0.0656</b>
Monday	0.559 (-5.23)	<b>0.0000</b>
Tuesday	0.674 (-3.51)	<b>0.0250</b>
Thursday	0.865 (-1.38)	0.6935
Friday	0.760 (-2.76)	<b>0.0033</b>
Saturday	0.666 (-4.10)	0.2202
<b><i>Outbound Departure Day of Week (Reference = Wednesday)</i></b>		
Sunday	0.579 (-5.56)	0.6282
Monday	0.626 (-4.67)	<b>0.0014</b>
Tuesday	0.872 (-1.36)	0.3529
Thursday	0.689 (-3.83)	<b>0.0002</b>
Friday	0.675 (-4.00)	0.3088
Saturday	0.830 (-1.82)	0.7581
<b><i>Prorated Fare</i></b>		
	1.00 (22.88)	<b>0.0300</b>
<b><i>Departure Month (Reference = September)</i></b>		
January	0.850 (-1.30)	<b>0.0138</b>
February	0.545 (-4.78)	<b>0.0016</b>
March	0.372 (-7.40)	<b>0.0001</b>
April	0.420 (-6.53)	<b>0.0001</b>
May	0.404 (-6.14)	<b>0.0010</b>
June	0.453 (-4.98)	<b>0.0001</b>
July	0.619 (-3.03)	0.8113
August	0.396 (-5.12)	0.6016
October	0.740 (-2.16)	<b>0.0094</b>
November	0.682 (2.88)	<b>0.0001</b>
December	0.720 (2.50)	0.3023
<b><i>Implied Trip Purpose (Reference = Short Leisure Trips)*</i></b>		
Business Trips	1.88 (6.54)	<b>0.1695</b>
Short Leisure	0.938 (-0.82)	0.9899
<b><i>Model Statistics</i></b>		
Log likelihood	-16241.16	
Number of obs. / failures	72,908 / 1,557	

\* Key: Short leisure trips are trips with length of stay < 7 and Saturday night.  
Long leisure trips are trips with length of stay ≥ 7 and Saturday night.  
Business trips do not include a Saturday night.

\* **Bolded** entries are violating the proportional hazard assumption.

Table 9: Cox Proportional Hazard Model with All Covariates

	Parameter estimate (z-stat)	Test of PH assumption
<b>Advance Purchase (Reference = 0-3 days)</b>		
3 to 7	2.450 (1.91)	0.8664
8 to 21	2.676 (2.17)	0.8830
22 to 30	1.983 (1.50)	0.8013
31 to 45	1.722 (1.19)	0.9938
46 to 60	1.065 (0.14)	0.9040
61 to 90	0.647 (-0.94)	0.7192
90+	0.433 (-1.80)	0.9044
<b>Carrier Code (Reference = Not dominant carrier)</b>		
Dominant carrier	2.313 (7.78)	<b>0.1664</b>
<b>Saturday night stay indicator (Reference = no Saturday night stay)</b>		
Leisure trips (Sat Stay)	0.491 (13.0)	<b>0.1354</b>
<b>Model Statistics</b>		
Log likelihood	-16465.263	
Number of observations	72,908	
Number of failures	1,557	

\* Key: Leisure trips include a Saturday night stay.

Business trips do not include a Saturday night.

\* **Bolded** entries are violating the proportional hazard assumption.

Table 10: Cox Proportional Hazard Model with Limited Covariates

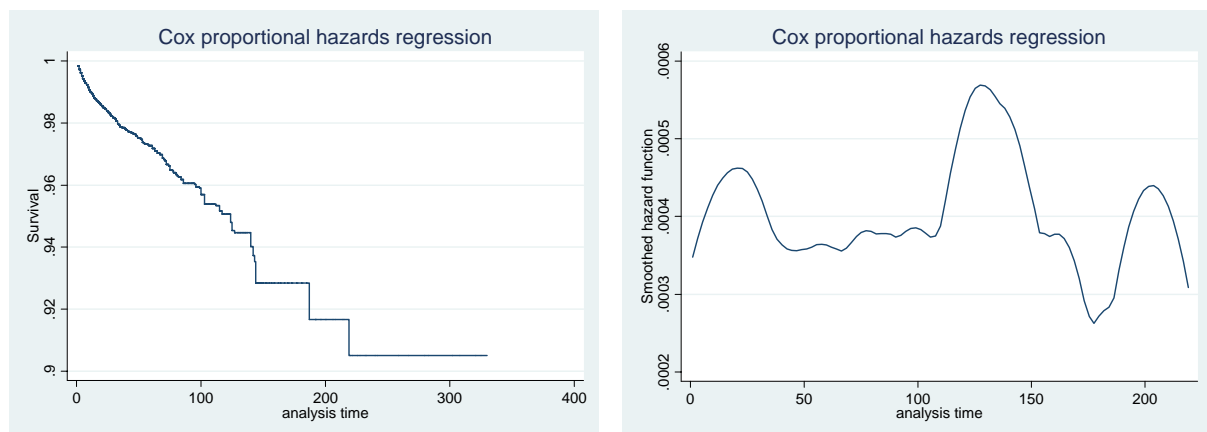


Figure 7: Cox PH Model Survival and Hazard Functions Using Mean Values of Predictors

### 5.3 Discrete Time Proportional Odds Model

A discrete time proportional odds model was estimated using the dominate carrier code and Saturday night stay indicators. The baseline hazard function was specified as a non-parametric, piece-wise function with intervals selected to ensure that at least 30 failures

were present in each interval<sup>8</sup>. As seen in Table 11, the effects of the dominant carrier and Saturday night stay indicators on the survival function are similar to the effects observed in the PH model. However, as seen in the parameter estimates of the hazard odds ratio coefficients, the DTPO model allows flexibility in specifying the baseline hazard function. In addition, the inclusion of advance purchase (represented as a continuous variable) is now significant. Figure 11 shows the impact of advance purchase on the survival functions. The top figure displays this relationship for all travelers, while the lower figure displays the relationship for business versus leisure travelers (defined by the presence of a Saturday night stay).

	Parameter estimate (z-stat)
<b><i>Hazard Odds Ratio</i></b>	
HO at 0	0.194456 (-10.93)
HO at 1	0.002752 (-45.22)
HO at 2	0.001940 (-44.09)
HO at 3	0.001963 (-43.76)
HO at 4	0.001783 (-42.98)
HO at 5	0.001567 (-41.96)
HO at 6	0.001229 (-40.13)
HO at 7	0.001205 (-39.56)
HO at 8	0.001225 (-38.95)
HO at 9	0.001359 (-39.16)
HO at 10	0.001662 (-39.92)
HO at 11	0.000891 (-35.54)
HO at 12	0.000992 (-35.96)
HO at 13	0.001209 (-36.89)
HO at 14	0.001112 (-35.90)
HO at 15 to 21	0.000764 (-53.92)
HO at 22 to 30	0.000621 (-51.87)
HO at 31 to 45	0.023302 (-49.95)
HO at 46 to 60	0.000360 (-38.27)
HO at 61 to 90	0.000617 (-38.74)
HO at 91+	0.001032 (-31.82)
<b><i>Advance Purchase (Continuous)</i></b>	
AP (# days)	0.98660 (14.24)
<b><i>Carrier Code (Reference = Not dominant carrier)</i></b>	
Dominant carrier	2.107 (7.26)
<b><i>Saturday Night Indicator (Ref = No Sat. night)</i></b>	
Leisure trips (Sat Stay)	0.4972 (13.3)
<b><i>Model Statistics</i></b>	
Log likelihood	-12,837.171
Number of obs / failure	3,274,781 / 1,557

Table 11: Discrete Time Proportional Odds Model

<sup>8</sup> In theory, one can define intervals that contain a minimum of 1 failure, but this was not done in this case to prevent over-parameterization.

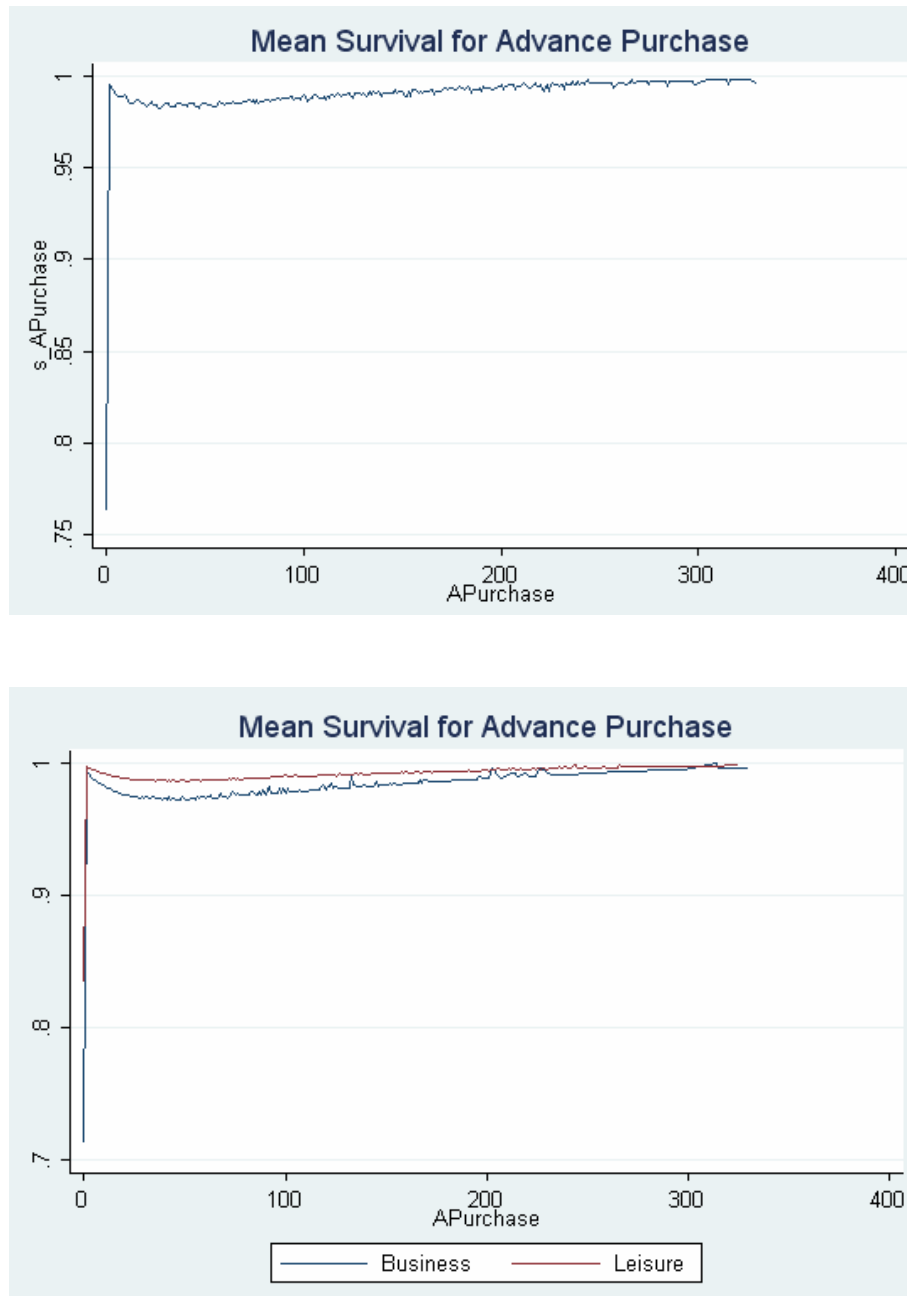


Figure 11: Survival Functions for Discrete Time Proportional Odds Model

## 6. Conclusions and Future Research

This analysis demonstrated how survival models can be used to forecast the refund and exchange behavior of airline passengers. The study differs from other studies in that it uses individual ticket data from the Airline Reporting Corporation. In addition, it builds upon the concepts of the cancellation models of Westerhoff (1998) by defining survival times relative to the outbound departure date. In addition, the influence of itinerary

covariates (Saturday night stay, length of stay) was explored using two different survival models. Preliminary results indicate that the survival function is influenced by many factors including advance purchase, carrier, and the presence of a Saturday night stay.

Several extensions to the models presented in this paper are in progress. These extensions can be thought of in terms of refinements to the model specifications and model validation. With respect to model refinements, future research efforts will focus on (1) exploring alternative functional forms for covariates (such as those that vary with time), (2) exploring heterogeneity of the population with respect to a baseline hazard function, (3) exploring alternative specifications of the hazard function, and (4) specifying separate models for refunds and exchange events. With respect to model validation, future research efforts will focus on verifying the robustness of current results on other markets and comparing the accuracy of model predictions with time-series cancellation methods currently used by the airline industry.

## References

- Achara, S. (2005). "Software Providers: Part II." Panel session of the Revenue Management and Price Optimization Conference, Georgia Institute of Technology, Atlanta, GA.
- American Airlines (2005). "Booking Policy Part 1: Updated November, 2005." <http://www.aa.com/content/uk/agency/bookingTicketing/bookPolOne.jhtml;jsessionid=X1K45TLTOWVRREAJN3U1C2QBFFXOVMD>. Accessed on April 30, 2006.
- American Express (2005). "Airline Policies as of April 1, 2005." <http://corp.americanexpress.com/gcs/travel/us/news/docs/AirlinePolicy-042005.pdf>. Accessed on April 30, 2006.
- Becker, N.G. (1989). *Analysis of Infectious Disease Data*. New York, NY: Chapman and Hall.
- Cox, D.R. (1972). "Regression models and life-tables (with discussion)." *Journal of the Royal Statistical Society*, B (34).
- Chatterjee, H. (2001). "Forecasting for Cancellations." Presentation to the AGIFORS Reservations and Yield Management and Study Group, Bangkok, Thailand, May 8-11.
- Data Base Products (2006). "The Origin-Destination Survey Of Airline Passenger Traffic." <http://www.airlinedata.com/Documents/O&DSURV.htm>. Accessed April 30, 2006.
- Garrow, L. & Koppelman, F. (2004a). "Predicting Air Travelers' No-show and Standby Behavior Using Passenger and Directional Itinerary Information" *Journal of Air Transport Management* 10: 401-411.
- Garrow, L. & Koppelman, F. (2004b). "Multinomial and Nested Logit Models of Airline Passengers' No-show and Standby Behavior," *Journal of Revenue and Pricing Management*, 3(3): 237-253.
- Gillen, D., Morrison, W. & Stewart, C. (2004). *Air Travel Demand Elasticities: Concepts, Issues and Measurement*. Final Report for the Department of Finance for Canada. [http://www.fin.gc.ca/consultresp/Airtravel/airtravStdy\\_1e.html](http://www.fin.gc.ca/consultresp/Airtravel/airtravStdy_1e.html) Accessed April 30, 2006.
- Hosmer, D.W. & Lemeshow S. (1999). *Applied Survival Analysis: Regression Modeling of Time to Event Data*. New York: John Wiley and Sons.

- Kim, M.Y. & Lagakos, S.W. (1990). "Estimating the Infectivity of HIV from Partner Studies." *Annals of Epidemiology* 1: 117-128.
- Nielsen//NetRatings (2005). "Online Travel Purchases Split Evenly Between Travel Agencies and Suppliers' Web Sites; Airline Supplier Sites Nearly Doubles Conversion Rates of Online Travel Agencies, According to Nielsen//NetRatings." [http://www.netratings.com/pr/pr\\_050621.pdf](http://www.netratings.com/pr/pr_050621.pdf). Accessed on May 1, 2006
- Parker, R., Lonsdale, R., Glans, T. & Zhang, Z. (2005). "The Naïve Universal Market Simulator," Presentation at the Advanced Market Techniques Forum of the American Marketing Association, Coeur d' Alene, Idaho. June 12-15.
- Polt, S. (1998). "Forecasting is Difficult – Especially if it Refers to the Future." Presentation to the AGIFORS Reservations and Yield Management Study Group, Melbourne, Australia, May 6-8.
- Ratliff, R. (1998). "Ideas on Overbooking." Presentation to the AGIFORS Reservations and Yield Management Study Group, Melbourne, Australia, May 6-8.
- Shiboski, S. C. & Jewell, N. P. (1992). "Statistical Analysis of the Time Dependence of HIV Infectivity Based on Partner Study Data." *Journal of the American Statistical Association* 87: 360-372.
- Shiboski, S. C. (1998). "Generalized Additive Models for Current Status Data." *Lifetime Data Analysis* 4: 29-50.
- Van Ryzin, G. & Karaesmen, I. (1999). "Overbooking with Substitutable Inventory Classes." Presentation to the AGIFORS Reservations and Yield Management Study Group, London, England, April 27-30.
- Westerhof, A. (1998). "CO<sub>2</sub> in the Air." Presentation to the AGIFORS Reservations and Yield Management Study Group, Melbourne, Australia, May 6-8.

### Appendix A- Survival and hazard function for a sample of the ARC Database

Table A-1 of the appendix presents the sample ARC data which was used to compare the differences between models that consider time as a continuous random variable versus models that consider time is a discrete random variable.

id	IssueDat	OutDepartDat	InDepartDat	RefundDat	ExchDat	EventFlag	TimeFail	APurchase
1	3-Dec-03	2-Jan-04	10-Jan-04	30-Dec-03		1	3	30
2	1-Jan-04	2-Jan-04		2-Jan-04		1	0	1
3	25-Nov-03	3-Jan-04	10-Jan-04	29-Dec-03		1	5	39
4	19-Dec-03	5-Jan-04	8-Jan-04		29-Dec-03	2	7	17
5	18-Nov-03	5-Jan-04	10-Jan-04			0	48	48
6	12-Dec-03	5-Jan-04	11-Jan-04			0	24	24
7	31-Dec-03	5-Jan-04	12-Jan-04			0	5	5
8	10-Dec-03	5-Jan-04	15-Jan-04			0	26	26
9	9-Oct-03	10-Jan-04	16-Jan-04			0	93	93
10	11-Dec-03	10-Jan-04	19-Jan-04			0	30	30

#### Legend

IssueDat – Issue Date for Ticket; OutDepartDat – Ticket Outbound Departure Date ; InDepartDat – Ticket Inbound Departure Date  
 RefundDat – Refund Date; ExchDat - Exchange Date ; EventFlag – Cancelled or Not (1- refunded , 2-exchanged, 0 – no event)  
 TimeFail – Survival Time ; APurchase – Difference between Outbound Departure Date and Issue Date

Table A-1 – Sample ARC data / BOS-MIA market

If in the **continuous time** specification we assume T to have a Weibull p.d.f. -  $f(t)$  then survival and hazard functions can be derived using the following formulas<sup>9</sup>.

$$f(t) = \beta t^{\beta-1} \exp(-t^\beta)$$

$$S(t) = \Pr(T \geq t) = \int_t^{\infty} f(x) dx = \exp(-t^\beta)$$

$$h(t) = \frac{f(t)}{S(t)} = \beta t^{\beta-1}$$

Figure A-1 presents the shape of the baseline survival and hazard function for the sample. It is important to note that since the largest values for the survival time (given by TimeFail) occur within the censored population, the baseline survival function does not reach the value zero. Also, the baseline hazard is monotonically decreasing with time which is represented by estimated shape parameter value ( $\beta \approx 0.55$ ).

The value of  $\beta$  can be also used to interpret the likelihood of ticket cancellation with respect to different point in time. In this case since  $\beta$  is less than one, we would expect that after 100 days the tickets will be 0.354 more times likely to be cancelled than after 10 days =  $(100/10)^{0.55-1}$ .

<sup>9</sup> Note that the possible influence of covariates was ignored. Typically the Weibull distribution can be extended to Weibull regression models considering either the scale parameter  $\alpha$  or the shape parameter  $\beta$  as functions of covariates. Parameter is estimated using MLE method.

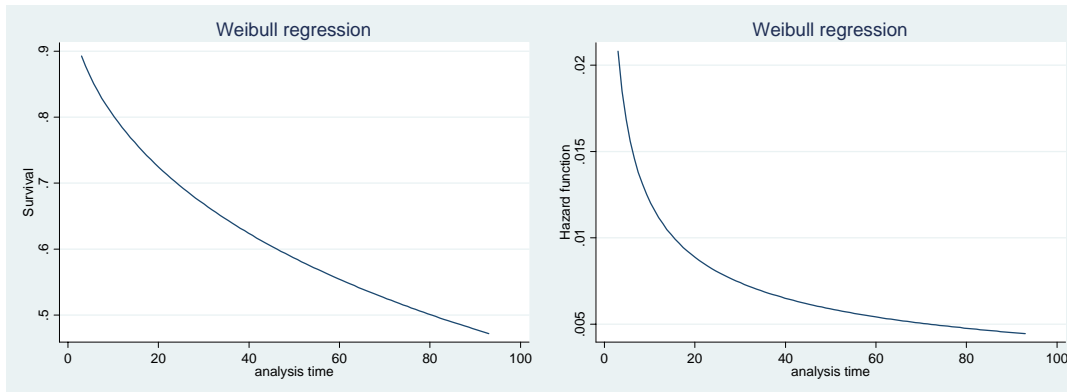


Figure A-1 Survival and hazard functions for a continuous time specification

A **discrete time** specification is appropriate when the survival process occurs in continuous time but survival times are observed only in intervals or when the time scale is intrinsically discrete. Depending on which of the two scenarios is more representative for the available data, one can either use a *complementary log-log* or *logit* transformation<sup>10</sup>. Considering the time to be intrinsically discrete (*i.e.*, refunds and exchanges are observed at the end of daily cycles) and a non-parametric specification of the baseline hazard, the survival and hazard functions can be derived using the following formulas<sup>11</sup>. Figure A-2 presents the shape of the baseline survival and hazard function.

$$\log \frac{h(j)}{1-h(j)} = \alpha_1 h_0^1(j) + \alpha_2 h_0^2(j) + \dots + \alpha_j h_0^j(j)$$

$$*h(j) = p(j)$$

$$p(j) = [1 + \exp(-(\alpha_1 h_0^1(j) + \alpha_2 h_0^2(j) + \dots + \alpha_n h_0^n(j)))]^{-1}$$

$$S(j) = (1 - p(1))(1 - p(2)) \dots (1 - p(j)) = \prod_{k=1}^j (1 - p(k))$$

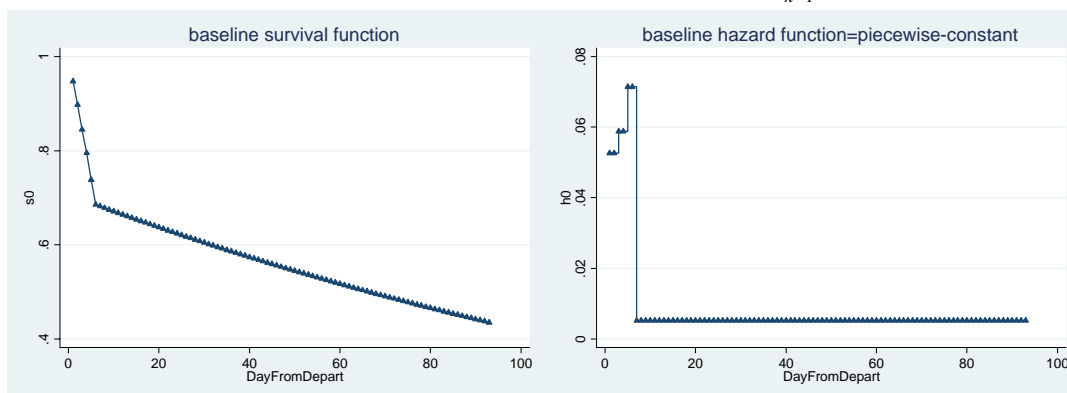


Figure A-2 Survival and hazard functions for a discrete time specification

<sup>10</sup> At time  $t_j$  and given covariates  $x_i$  the logit transformation is defined by  $\text{logit } h(t_j|x_i) = \text{logit } h_0(t_j|x_i) + X'\beta$  while the complementary log-log transformation is defined by  $\log(-\log(1 - h(t_j|x_i))) = \log(-\log(1 - h_0(t_j|x_i))) + X'\beta$

<sup>11</sup> Again the possible influence of covariates was ignored for this example.