A HAZARD MODEL OF U.S. AIRLINE PASSENGERS' REFUND AND EXCHANGE BEHAVIOR

Dan C. Iliescu, Ph.D. student
Georgia Institute of Technology
School of Civil and Environmental Engineering
790 Atlantic Drive
Atlanta, GA 30332-0355
Email: gth657x@mail.gatech.edu

Laurie A. Garrow (*corresponding author*)
Assistant Professor
Georgia Institute of Technology
School of Civil and Environmental Engineering
790 Atlantic Drive
Atlanta, GA 30332-0355
Ph: (404) 385-6634
Fax: (404) 894-2278
Email: Laurie.Garrow@ce.gatech.edu

Roger A. Parker
Senior Marketing Scientist and Chief of Technology
Marketing and Business Strategy
Boeing Commercial Airplanes
P.O. Box 3707 MC 21-33
Seattle, WA 98124-2207
Ph : (206) 766-2922
Fax : (206) 766-1030
Email : Roger.A.Parker@boeing.com

Last update: December 9, 2006

ABSTRACT

This study explores the use of discrete choice methods for airline passenger cancellation behavior. A Discrete Time Proportional Odds model with a retrospective time scale is estimated based on the occurrence of refund and exchange events in a sample of tickets provided by the Airline Reporting Corporation. Empirical results based on 2004 data from eight domestic U.S. markets indicate cancellation rates are strongly influenced by both the time of ticket purchase and the time until departure. In particular, while business travelers purchasing tickets close to flight departure are more likely to refund or exchange their tickets, both leisure and business travelers increase their refund and exchange activities as their flights approach departure. Cancellation rates are also influenced by several other covariates, including departure day of week, market, and group size.

Keywords: hazard model, discrete choice, revenue management, cancellation, air traveler behavior

A HAZARD MODEL OF U.S. AIRLINE PASSENGERS' REFUND AND EXCHANGE BEHAVIOR

1. INTRODUCTION

Currently, there is renewed interest in the airline industry in integrating discrete choice models of passenger behavior with traditional revenue management, scheduling, and other applications. This interest is renewed, not new, in the sense that as early as the 1980's several attempts were made to use discrete choice models in revenue management. However, with a few exceptions, these initial discrete choice modeling efforts were abandoned in favor of more simplistic probability models (*e.g.*, demand for booking classes on a flight arrives according to a Poisson process, cancellations are binomially distributed, etc.) and/or time-series methodologies based on historical averages (*e.g.*, the no show rate for a flight is a weighted average of no show rates for the previous two months). While these probability and time-series models were easier to implement, they did not capture or explain how individual airline passengers made decisions. Moreover, many of the models currently used in practice make strong independence assumptions; *e.g.*, it is common to assume the demand associated with a booking class on a flight is independent of the demand for all other booking classes on that (and surrounding) flights. However, over the last several years, these and other assumptions embedded in traditional revenue management algorithms have begun to be more openly challenged (Oliveira 2003; Boyd 2004; Boyd and Kallesen 2004; Hornick 2004; Lieberman 2004; Dunleavy and Westermann 2005; Ratliff and Vinod 2005; Van Ryzin 2005), forcing a re-examination of how one can model

individual airline passengers' behavior using discrete choice or other models ground in behavioral theory. Recent work using discrete choice methods for revenue management include that by Garrow and Koppelman (2004a; 2004b) for no show applications, Ratliff (2006) for demand unconstraining and recapture applications, and Talluri and van Ryzin (2004a) who explore the use of a simple multinomial logit (MNL) embedded in a optimization model to determine seat allocation levels.

This study explores the use of discrete choice methods for another component of revenue management, namely cancellation models. Specifically, a model of airline passengers' "cancellation" behavior is estimated based on the occurrence of refund and exchange events in a sample of ticketing data from the Airline Reporting Corporation (ARC). Survival analysis methods are used to explore the pattern of cancellation probabilities over time and to determine the extent in which the observed heterogeneity of tickets (*i.e.*, predictors) changes that pattern. With respect to the pattern of cancellation probabilities over time, survival models are used to predict the conditional probability that a purchased ticket will be cancelled in a time period given it survived up to that point (*hazard probability*). With respect to the observed heterogeneity, survival models are used to explore the amount of variation induced by different predictors in the hazard probability. Empirical results based on a Discrete Time Proportional Odds (DTPO) model indicate that cancellation rates are indeed influenced by both the time of purchase and time until departure, in addition to a many other covariates (including departure day of week, market, group size, etc.).

This study contributes to the existing literature in two ways. First, new behavioral insights into airline passengers' refund and exchange behavior emerge. Second, to the best of the authors' knowledge, this is the first published study to be based on ticketing data from ARC. The ticketing database reveals several new characteristics of airline "cancellation" behavior that provide additional impetus for challenging the traditional revenue management framework and exploring new research directions.

The remainder of this paper contains several sections. First, a review of airline cancellation models is provided followed by a description of the data used for the analysis. Next, the methodology and results are presented. Finally, directions for future research are discussed. Due to the uniqueness of ARC ticketing data, an Appendix is included that describes how the ticketing data used for this study differs from the publicly-available ticketing data more commonly reported in the literature.

2. REVIEW OF AIRLINE CANCELLATION MODELS

Airlines use revenue management to decide how many seats (associated with a set of prices) to make available for sale to customers. However, since all customers who request seats do not actually travel, airlines overbook to reduce the expected number of empty seats on flights when there is demand for those seats. Cancellation and no show rates are used to determine the overbooking level, *i.e.*, the number of seats authorized for sale that exceed the capacity of the flight. The difference between cancellation and no show models relates to when the airline knows passengers do not intend to travel. Cancellation models predict how many passengers

inform the airline they do not intend to travel prior to the departure of their flights while no show models estimate the number of remaining booked passengers, *i.e.*, passengers who have not cancelled, but fail to show for their flights.

Airline no show and cancellation models are based on booking information, which is distinct from ticketing information. The difference between these two data relates to whether the passenger has paid for a reservation. A reservation, or booking request, that has been paid for appears in both booking and ticketing databases, while a booking request that has not yet been paid for appears only in a booking database. Historically, many high-yield fares, such as unrestricted coach fares and business class fares, did not need to be purchased at the time of booking, but rather could be held in the reservation system until flight departure. However, while revenue management systems need to set seat allocation levels based on booking information (as some of these unpaid reservations, and particularly high-yield passengers, will show for flights), including information on whether a booking has been paid for has been shown in prior studies to be one of the most important factors for predicting no show rates (Garrow and Koppelman 2004a; 2004b). This is due to the ability to better identify speculative bookings (*i.e.*, low-yield bookings that have not been paid for within 24 to 48 hours after the original booking request was made).

Different types of cancellation models are discussed in the literature, but generally fall into two categories. The first category of models predicts the probability of survival from one period to the next while the second category directly predicts the probability a booking survives until flight departure. For example, the cancellation model described by Westerhof (1997) in

Figure 1 divides the booking horizon[1] into *T* separate booking periods. By definition, a booking that exists at booking period *t* survives until booking period *t*+1 with probability p_t and is cancelled with probability $(1-p_t)$ for *t* > the departure day of the flight. The probability a booking that exists at time *t* survives until departure is given as $p_t \times p_{t+1} \times \ldots \times p_{T-1} \times (1 - \text{no show rate})$. Values for p_t are empirically derived from historical data. However, it is important to note that this model assumes cancellation probabilities are independent of when passengers book. Thus, conceptually, p_T does not depend on whether the bookings that exist at time *T* are from business passengers who booked in period *T*-1 or from leisure passengers who booked far in advance of the flight departure at time period *t*.

[Insert Figure 1 about here]

While Westerhof's model predicts the probability a booking that exists at time *t* will survive until the next period, Talluri and van Ryzin (2004b) note that it is also common to directly model the number of bookings that survive until departure using a binomial distribution defined by q_t, the probability a booking at time *t* survives until departure[2]. The authors reference a Tasman Empire Airways study (Thompson 1961) as empirical evidence on the validity of the binomial distribution assumptions (*i.e.*, that (1) customers cancel independently of each other; (2) each customer has the same probability of cancelling; and, (3) cancellation

[1] The booking horizon is defined as the period during which an individual can make reservations for a flight. The typical booking horizon for an airline is 330 days.

probabilities are "memoryless" in the sense that they depend only on the time to flight departure and not on when the booking was first created). In contrast, Westerhof notes in his 1997 study of data from KLM that cancellation rates are not memoryless. However, while recognizing this phenomenon, Westerhof does not propose a methodology that can simultaneously incorporate these two dimensions of time, a question that is explicitly addressed in this research.

The cancellation models commonly reported in the literature are not representative of cancellation models used in practice. In general, cancellation models used in practice by airlines are based on two distinct components. The first component estimates the number of bookings currently active (or alive) at time *t* in the revenue management system that will survive until flight departure. The total number of active bookings at time *t* is defined as "current gross bookings" and those bookings expected to survive until flight departure are referred to as "current net bookings." The second component estimates the number of bookings that will arrive in between time period *t* and flight departure ("future gross bookings") and predicts how many of these will survive until departure ("future net bookings"). Different methods are used to forecast future demand and apply cancellation rates to current gross and future gross bookings. However, it is important to note that these four pieces of "cancellation information" are at the core of many revenue management systems used in practice and are distinct from models typically discussed in the literature.

---

[2] Note that for time period *T*, q_T is the no show probability. Talluri and van Ryzin (2004b) also discuss possible refinements to the binomial model (inflation of variance of the show demand; moment generating functions) for datasets in which the percentage of groups is large.

An example of the evolution of cancellation information during the booking[3] horizon is shown in Figure 2. For example, at time period five, there are five bookings in the reservation system (current gross bookings). Booking two is not included as a current gross booking because it was refunded or exchanged prior to period five. Of the five current gross bookings, one (booking six) cancels prior to departure, resulting in four current net bookings. Also, between time period five and departure, four bookings are "expected" to occur (future gross bookings), and one of these (booking eight) is expected to be refunded or exchanged, resulting in three bookings that survive (future net bookings).

[Insert Figure 2 about here]

3. DATA

It is important to emphasize from the outset that this study is based on ticketing data. This is in contrast to the booking data sources that are used by airlines to develop cancellation models. One of the key motivations for using ticketing data for this study is because 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

[3] To be consistent with the discussion of other cancellation models, the term "booking" is used; however, ticketing information is used in the analysis.

components of the passenger behavior models is a cancellation model. However, BCA does not have access to the same data that airlines do, including their detailed bookings and check-in information. Thus, one of the key motivations for this study was to determine the viability of using alternative data source containing disaggregate passenger records that could support development of a cancellation model across multiple carriers and markets[4].

In order to assess the viability of using ticketing data from the Airline Reporting Corporation (ARC) for cancellation modeling, a sample of ticketing data was obtained. However, 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 (including flight numbers, departure and arrival times, number of stops, etc.) has been suppressed. The data used for this study contains simple one-way and round-trip tickets[5] for which the outbound departure date occurred in 2004. 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[6]. 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

[4] For a detailed discussion on how the ticketing data used in this study relates to the management system used by an airline that is based on booking information, see Iliescu, Garrow, and Parker (2006).

[5] A "simple" ORD-HNL one-way itinerary is one in which the trip starts in ORD and ends in HNL. The passenger embarks at ORD (*i.e.*, there are no flight segments before ORD) and disembarks at HNL (*i.e.*, there are no flight segments after HNL). Similar logic applies to round-trip itineraries.

(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. While carrier confidentially considerations restrict the amount of flight-level information available for analysis, the sample data is unique in its ability to capture information about the time until refund and exchange events across multiple markets and multiple carriers.

From a modeling perspective, it is generally believed that cancellation rates differ for business and leisure passengers. For example, business passengers who are more time-sensitive and require more travel flexibility may be more likely to modify their itineraries than leisure passengers, leading to higher cancellation rates. While airlines do not explicitly collect information about trip purpose, trip purpose can be inferred from several other booking (or ticketing) information related to the passenger's itinerary. An itinerary is defined as a flight or sequence of flights that connects an origin and destination. Non-directional itineraries do not contain information on whether passengers on a flight from MIA-SEA are traveling outbound (as would be the case for a passenger living in MIA) or inbound (representing a passenger returning home to SEA after visiting MIA). While non-directional information is predominately used in airline's revenue management systems (and is becoming a larger percentage of all bookings due to the predominance of one-way pricing strategies favored by low-cost carriers that are being partially adopted by legacy carriers), directional itinerary information provides a much richer set

---

[6] Based on ticketing information available from the U.S. Department of Transportation (described in the Appendix), the markets selected for this analysis represent a market share of approximately 60% for American Airlines, 10% for Alaska Airlines, 10% for United Airlines and 5% for Delta Air Lines. Several other carriers are also represented.

of variables from which trip purpose can be inferred. For example, business passengers are more likely to depart early in the week, stay a few nights, and return home later in the week (and thus not stay over a Saturday night). In contrast, leisure passengers are more likely to depart later in the week, stay more nights than a business passenger, and stay over a Saturday night.

With respect to ARC data, 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[7] (*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 stay. Descriptive statistics for these and other variables are provided in Iliescu, Garrow, and Parker (2006).

4. METHODOLOGY

A Discrete Time Proportional Odds (DTPO) model is used to predict the conditional probability of a ticket being "cancelled" (*i.e.*, refunded or exchanged before outbound departure date) at each day from departure given the time the ticket was purchased. In addition, the effects of observed heterogeneity (*i.e.*, Saturday night stay, group size, departure day of the week, market and carrier) on the conditional probability are explored. The DTPO formulation adds to the

cancellation forecasting methods used in the airline industry (Polt 1998; Chatterjee 2001) by better quantifying the uncertainty of ticket cancellation rates and providing empirical evidence that the memoryless property of the cancellation rates is violated.

This section contains several sections. First, an overview of key concepts of time-to-event modeling that need to be considered in the context of airline cancellation models is provided, followed by a discussion of how the DTPO model can be applied to ARC ticketing data. Finally, the formulation and estimation of the DTPO model is described.

4.1 Fundamental Concepts for Time-to-Event Modeling

Time-to-event models are designed to analyze data for which the response variable is defined as the time to an event(s). In contrast to classical linear regression methods, time-to-event models typically exhibit two notable features. First, the outcome variable, "time" may be only partially observed for some individuals, which creates the presence of censored data. Second, distinct from cross-sectional studies that measure the impact of covariates at a single point in time, time-varying covariates are possible (McCullagh and Nedler 1989). Both aspects are governed by a "time at risk" mechanism in which the dynamics of conditional probabilities of an event happening (*i.e.*, the transition intensity) are assessed as a function of the elapsed time since the

[7] For round trip itineraries, outbound and inbound fares are obtained by prorating the total fare to each directional itinerary according to trip distance.

entry time.[8] The presence of multiple events complicates this mechanism and may create identification problems when the transition pattern from one event to another varies across different observations, *e.g.,* for the case of competing risks.

The medical field was the first to benefit from the time-to-event models capabilities. Statistical methods for "survival" data were introduced as epidemiological applications (that capture the time-to-occurrence of an event given exposure to an infection) or as clinical applications (that capture the time-to-occurrence of an event given exposure to treatment). Besides their scope difference, the two applications were distinguished by the way in which survival time was considered - either retrospective or prospective (Kim and 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. In contrast, in prospective studies investigators use a "forward looking" approach to analyze the evolution of disease for individuals exposed to various treatments. As discussed in subsequent sections, the use of a retrospective time scale proves particularly advantageous in airline cancellation modeling. For a detailed review of survival analysis applications and how they can be incorporated into the class of generalized additive models (GAM), see Shiboski (1998).

Although survival analysis concepts were first tested and validated by the medical field, their potential application to demography, econometrics, travel demand, and other areas was immediate. Nevertheless, if at their core, the new applications made use of "survival" concepts

[8] In the case when entry time is the same as the time when the subject becomes at risk. This might not always be the case (*e.g.* delayed entry).

(*i.e.*, they modeled the dynamics of the conditional probabilities of event happening given the length of "exposure"), the events of interests and spells[9] characteristics differed. The research community was prolific in exploiting these conceptual differences. Indeed, the multitude of methodological "add-ons" is testimonial to the degree of generalization that survival analysis concepts have reached. Comprehensive reviews of these applications are provided by several authors (Kiefer 1988; Jain and Vilcassim 1991; Hensher and Mannering 1994; Bhat 2000; Wu 2003). While an extensive overview of the methodological challenges associated with adapting survival models to different applications is beyond the scope of this paper, two on-going research topics are particularly relevant in the context of cancellation modeling.

The first topic relates to how to "appropriately" specify models when multiple time dimensions are present. This problem is frequently encountered in life course demographic studies based on cohort datasets. Specifically, an underlying identification problem exists because given knowledge of the respondents' age and duration in the study, their cohort (or entry in the study) is uniquely determined. Thus, the researcher needs to determine which two time dimensions are most appropriate to include in the analysis. Specifically, as noted by Wu (2003) "due to the complex nature of individual biographies, it has become clear that the notion of life course cannot be understood simply as a process of unilinear aging." If, by and large, life course researchers have agreed on the necessity to consider multiple dimensions of the dependent variable (*e.g.*, age, duration and cohort), the choice of which ones to include in the analysis is rather subjective and requires external knowledge that is often very difficult to acquire. This

[9] Spells, defined as the length of time spend in each state, are also called durations.

issue is relevant in the context of cancellation models when testing for the validity of the memoryless property (*i.e.*, how to simultaneous consider both the time of ticketing and days from departure).

The second topic refers to the difficulty of isolating the transition domain of a single event in an environment in which inherent interrelationships between different events exist. Nevertheless, solutions to the competing risks problem have been successfully designed and tested in the travel demand field. Under the more general taxonomy of "duration models," (Bhat 1996) addresses the problem of "multiple duration-ending outcomes" as a joint estimation of outcomes and outcomes-specific hazard durations using a proportional hazard model with a non-parametric specification of the base-line hazard. This issue becomes relevant in the context of cancellation models if one wants to consider ticketing refund and exchange as separate, competing events.

In addition to these specific methodological issues, it is important to note that when modeling time-to-event data, choosing the most appropriate model can be cumbersome since "methods are so similar in their underlying philosophy that they *usually* give similar results" (Allison 1995). In general, time-to-event models are characterized by two categories of assumptions: (1) distributional assumptions about the dependent variable, and (2) observed heterogeneity assumptions about the influence of the vector of covariates on the time-to-event process.

With respect to the first category of assumptions, it is important to note that the results of a time-to-event model can be interpreted using two closely related functions: survival and

hazard[10]. As such, assumptions about the time-to-event density function are closely related to assumptions about the hazard function (*e.g.*, the exponential density function will generate a constant hazard rate while the Gompertz density function will generate a linear increase in hazard rate). Furthermore, in the case when there is no clear evidence to favor one parametric specification over the other, or in cases where the time-to-event process evolves along a discrete-time scale, parametric specifications can be replaced by semi-parametric or non-parametric counterparts.

With respect to the second category of assumptions, the influence of the vector of covariates on the time-to-event process can be addressed using two classes of models – proportional hazard models (PH) and accelerated failure time models (AFT). The fundamental difference between them is that while AFT models coefficients represent changes in survival time due to a unit change in a given covariate, PH models coefficients represent changes in the hazard rates due to a unit change in a given covariate.

Among the time-to-event models, the Discrete Time Proportional Odds (DTPO) specification represents a viable alternative to its well-known parametric counterparts. First introduced by Cox (1972) as an extension to the semi-parametric COX model, the DTPO model has three main advantages: (1) is can be used with datasets in which a large number of ties[11] are

---

[10] The hazard function (the instantaneous risk that an event will occur at time $t = h(t)$) is defined by the survivor function (probability of survival beyond $t = S(t)$) and time-to-event density function ($f(t)$) .

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

[11] Events occurring at the same point in time.

present; (2) it has the ability to incorporate different assumptions related to the shape of the hazard function; and, (3) it has a built-in flexibility for incorporating time-varying covariates or time-varying effects of covariates. The main disadvantage of the DTPO model is that few statistical estimation packages exist that can be used to estimate a DTPO model using partial likelihood (PL) methods. However, this limitation can be addressed by reformatting a dataset prepared in time-to-event format, and estimating the model using maximum likelihood (ML) methods.

In the context of this research, two arguments favor the use of the DTPO model as an appropriate formulation to estimate the pattern of tickets' cancellation probabilities. The first argument refers to the computational efficiency of the ML estimators when compared to the PL estimators. Indeed, since the ARC sample dataset is a "consolidated" dataset, with tickets aggregated from eight different markets, the presence of a large number of ties is inevitable, a fact that eliminates the alternative of an exact COX model estimation. The second argument refers to an on-going debate in the revenue management field as to which is the most appropriate model to describe how cancellation probabilities evolve over time. Although several authors indicate that the value of cancellation probability over time is constant (Littlewood 2005) and independent of the time of booking (Talluri and Van Ryzin 2004b), empirical evidence suggests otherwise (Westerhof 1997; Chatterjee 2001). In this context, the DTPO model offers the flexibility of testing different scenarios with minor adjustments. In view of these advantages, the next section describes the DTPO model as an alternative way to estimate cancellation probabilities for the sample of ARC airline tickets.

4.2 Discrete Time Proportional Odds (DTPO) Model and ARC Data

The DTPO model extends previous research on the distribution of cancellation rates/proportions[12] in four aspects. First, it relaxes the general assumption of population homogeneity and tests the influence of observed heterogeneity on cancellation rates/proportions by considering different segmentations/covariates (Saturday night stay, outbound departure day of week, market, carrier, group size, pro-rated fare). Second, it assumes that heterogeneity across tickets is fully captured by these covariates and its effect is distinct from that of time (changes in covariates values produce only vertical shifts and no distortions in a "baseline" cancellation rate line, *i.e.*, the proportional hazard assumption). Third, by construction, the DPTO model accommodates time-varying effects of covariates, thus allowing for interactions between time of ticketing and days from departure to be explored. Finally, since the time-scale is discrete, the DTPO model has sufficient flexibility to test different distributional shapes for the baseline cancellation rate.

It is important to note that compared to the typical time-to-event datasets, the ARC sample ticketing data has several unique characteristics. The first characteristic is that the tickets "lifetimes" are completely determined, and end either in a cancellation (exchange/refund date) or in certain non-cancellation (outbound departure date). Both events are outside the control of the investigator. To compensate for this particularity of the data which can generate problems, the entry time (*i.e.*, *Time of Ticketing*) was included in the model as a means to control potential

[12] Chatterjee (2001) defines a **cancellation rate** at time *t* as the proportion of those booked at *t* which cancel by *t*+1 and a **cancellation proportion** at time *t* as the proportion of those booked at t which cancel by departure day.

estimation bias (Allison, 1995). In addition, a retrospective time scale was used. The sample hazard using a retrospective time scale for the ten tickets shown in Figure 2 is presented in Figure 3. Note that, in contrast to the prior example, at time $t=5$ the "population" includes all tickets that purchased in time periods five to nine, including ticket two that was purchased in period eight and refunded or exchanged in period six. The sample hazard at this point in time is 0/5 = 0% because all of the tickets in the population survive until the "next" retrospective time period, or time period six. While outside the scope of this paper, the use of a retrospective time scale in conjunction with booking information associated with "current" and "future" bookings can be translated into the cancellation information shown in Figure 2.

[Insert Figure 3 about here]

The second characteristic is that out of the total population of refund and exchange events, a significant part (*i.e.*, 25%) occurs after the outbound departure date. As discussed in Iliescu, Garrow, and Parker (2006), these events will generally appear as no shows. As such, the current analysis makes use of only those exchange or refund events that occur prior to the outbound departure. Furthermore, it assimilates these events into a single "cancellation event" and thus develops a cancellation model for the outbound legs of an airline itinerary.

The third characteristic is that the assumption of independence between observations is undoubtedly violated by the presence of groups. Therefore, the ARC dataset was transformed from an individual ticket level database to a group level database. More specific, observations determined to have the same values on the entire set of covariates with the same scrambled

passenger name record (PNR)[13] were eliminated and a variable indicating the group size added to the set of covariates. Finally, taking into account that the majority of tickets are booked in the 0 to 90 days from departure (DFD) time interval (95% of total number of tickets) and cancellation events for the rest of tickets are relatively scarce (5% out of total number of events), the ARC dataset was reduced to tickets purchased within 90 days of departure. Given all of the above characteristics, the ARC time-to-event application can be viewed as a *ticketing cancellation model on the outbound legs of an airline itinerary for groups for a ticketing horizon of 90 days from departure.* After this data reduction process, the original ARC dataset of 234,370 tickets (1.3% -Refunds; 1.2%-Exchanges) was transformed to 151,118 distinctive groups (2.03% - Cancellations).

4.3 Model Formulation and Estimation

Using the transformed ARC data, the Discrete Time Proportional Odds (DTPO) model partitions the time-to-event of the ith ticket (T_i) into a number of k disjoint time intervals $(t_0, t_1]$, $(t_1, t_2]$,$(t_2, t_3]$, …,$(t_{k-1}, t_k]$. The bounds of the time intervals $(t_0,t_1,\ldots,t_k)$ identify the days from departure (DFD) with outbound departure date as t_0 as the outbound departure date and t_k as either the time of ticketing (non-cancelled tickets) or the time of ticket refund/exchange (cancelled tickets). Furthermore, the discrete hazard of a cancellation event for the i^{th} ticket in the k^{th} interval is defined as the conditional probability that ticket i will experience the cancellation event in

[13] To ensure carrier and passenger confidentiality, ARC provided "scambled" PNR information and ensured that these records were unique within a specific market. PNRs can be used to determine how many passengers are traveling together on the same reservations.

interval *k* given survival up to that point (Equation 4.1). Using conditional probability theory, it follows that the probability that a cancelled ticket will experience the event in the k^{th} interval is equal to the product between the non-event conditional probabilities of 1 to *k*-1 time intervals and the event conditional probability of *k* time interval (Equation 4.2). Similarly, the probability that a non-cancelled ticket will experience the cancellation after the k^{th} interval is equal with the product of non-event conditional probabilities of all *k* time intervals (Equation 4.3).

$$h_{ik} = P(T_i = k \mid T_i \geq k) \tag{4.1}$$

$$\begin{aligned} &P(T_i = k) = \mathrm{P}(T_i = k \mid T_i \geq k) \cdot \mathrm{P}(T_i \neq k-1 \mid T_i \geq k-1) \ldots \mathrm{P}(T_i \neq 1 \mid T_i \geq 1) \\ &P(T_i = k) = h_{ik} \cdot (1 - h_{i(k-1)}) \cdot (1 - h_{i(k-2)}) \ldots (1 - h_{i1}) \end{aligned} \tag{4.2}$$

$$\begin{aligned} &P(T_i > k) = \mathrm{P}(T_i \neq k \mid T_i \geq k) \cdot \mathrm{P}(T_i \neq k-1 \mid T_i \geq k-1) \ldots \mathrm{P}(T_i \neq 1 \mid T_i \geq 1) \\ &P(T_i > k) = (1 - h_{ik}) \cdot (1 - h_{i(k-1)}) \cdot (1 - h_{i(k-2)}) \ldots (1 - h_{i1}) \end{aligned} \tag{4.3}$$

As a result, the likelihood contribution for canceled and non-cancelled tickets can be expressed using Equations 4.4 and 4.5 and further detailed as the product of all the individual likelihoods[14] – Equation 4.6 (Cox 1972).

$$L_i = h_{ik} \cdot \prod_{j=1}^{k-1} (1 - h_{ij}) \tag{4.4}$$

$$L_i = \prod_{j=1}^{k} (1 - h_{ij}) \tag{4.5}$$

[14] c_i is a censoring indicator equal to 0 for cancelled tickets and 1 for non-cancelled tickets.

$$L = \prod_{i=1}^{n} \left[h_{ik} \cdot \prod_{j=1}^{k-1} (1 - h_{ij}) \right]^{1-c_i} \cdot \left[\prod_{j=1}^{k} (1 - h_{ij}) \right]^{c_i} \tag{4.6}$$

Furthermore, since the exact time of tickets transition from the state of not-cancelled to cancelled can be captured using a binary variable y_{ij} equal with 1 if ticket is cancelled in the j^{th} day from departure and 0 otherwise, it follows that Equation 4.7 is an alternative form to express the log-likelihood function. Moreover, the likelihood function for the entire sample (Equation 4.8) is equivalent with the likelihood function of a binary logistic regression model for which y_{ij} are assumed to be a collection of independent variables and whose data structure is expanded[15] to represent an unbalanced panel dataset (*i.e.*, each ticket observation is replicated multiple times, one time for each day from departure of the ticket lifetime).

$$l = \log L = \sum_{i=1}^{n} \sum_{j=1}^{k} y_{ij} \cdot \log \left(\frac{h_{ij}}{1 - h_{ij}} \right) + \sum_{i=1}^{n} \sum_{j=1}^{k} \log \left(1 - h_{ij} \right) \tag{4.7}$$

$$L = \prod_{i=1}^{n} \prod_{j=1}^{k} h_{ij}^{\,y_{ij}} (1 - h_{ij})^{(1-y_{ij})} \tag{4.8}$$

The equivalence between the two likelihood formulations (Equations 4.6 and 4.8) defines the rationale behind the DTPO model, a model introduced by Cox (1972) and further detailed by several authors (Brown 1975; Thompson 1977). For a general set of covariates X_l, Equation 4.9 presents the general formulation of the DPTO model, while Equations 4.10 and 4.11 presents the estimation of hazard and survival probabilities.

[15] The creation of the expanded dataset process has several steps: (1) duplicating the set of time-invariant covariates over the entire life-time of a ticket, (2) filling in the time-variant covariates (if present) and (3) creating the binary indicators of the cancellation status y_{ij}.

$$\log\left(\frac{h_{ij}}{1-h_{ij}}\right) = \Psi_{ij} + \beta_1 \cdot X_{ij1} + \beta_2 \cdot X_{ij2} + ... + \beta_l \cdot X_{ijl} \quad (4.9)$$

where j=1,2,…,k time intervals ; i=1,2,…,n observations ; Ψ_{ij}- baseline hazard function

$$h_{ij} = [1 + \exp(-(\Psi_{ij} + \beta_1 \cdot X_{ij1} + \beta_2 \cdot X_{ij2} + ... + \beta_l \cdot X_{ijl})]^{-1} \quad (4.10)$$

$$S_{ij} = (1-h_{i1})(1-h_{i2})...(1-h_{ik}) = \prod_{j=1}^{k}(1-h_{ij}) \quad (4.11)$$

5. EMPIRICAL RESULTS

This section presents empirical results of the DTPO model. The choice of functional form for the baseline hazard and the effect of the observed heterogeneity on this pattern are detailed.

5.1 Interpretation of the Baseline Hazard

With respect to the choice of functional form for the baseline hazard, it is important to note that the DTPO model estimation is constructed using two fundamental assumptions. First, a linear relation between the covariates and the logistic transformation of ticket cancellation hazard is assumed (linearity assumption). Second, the effect of covariates over the odds of cancellation is considered to be constant over time (proportionality assumption). In view of these assumptions, the DTPO model formulation can be conceptualized as the multiplicative effect of the covariates' log-linear function on a baseline odds function (Equation 5.1).

$$\frac{h_{ij}}{1-h_{ij}} = \frac{h^0{}_{ij}}{1-h^0{}_{ij}} e^{\beta_1 \cdot X_{ij1} + \beta_2 \cdot X_{ij2} + \ldots + \beta_l \cdot X_{ijl}} \qquad (5.1)$$

When the magnitude of conditional probabilities is small (as is the case with the ARC data), Equation 5.1 indicates that the DTPO model can be a close approximation of the proportional hazard (PH) model[16]. In this context, non-parametric estimators of survival probability (Kaplan-Meier), cumulative hazard (Nelson-Aalen) and hazard rate (Cox-Oaks) are important since they provide a starting point for the assumptions on the time-to-event distribution, *i.e.* the shape of hazard rate with respect to time. Figure 4 presents the survival Kaplan-Meier estimator and lowess smoother with neighborhood bandwidth equal to 0.1 for the sample hazard $h_j=s_j/n_j$ (s_j- number of cancelled tickets during j^{th} day from departure; n_j- number of total tickets during j^{th} day from departure). A visual inspection of the two graphs reveals the large number of non-cancelled tickets and a decreasing trend of the hazard probability in the 0-60 days from departure (DFD) time horizon. Still, no conclusion can be draw about the shape of hazard probability for the 61-90 DFD time horizon, for which signs of instability are present (*e.g.*, at 69 and 75 days from departure).

[Insert Figure 4 about here]

The shape of the non-parametric hazard smoother was used as a basis for defining three baseline hazards specifications for the DTPO model: discrete, quadratic, and logarithmic.

[16] The odds of a cancellation event will be approximately equal to the conditional probability of cancellation (*i.e.*, $h_{ij} \approx h_{ij} / (1- h_{ij})$.

Estimation results for the transformed ARC dataset using these specifications are presented in Table 1 while graphs of the baseline hazards are presented in Figure 5.

[Insert Table 1 about here]

[Insert Figure 5 about here]

Although the visual inspection of Figure 5 indicates that the second order polynomial is the only functional form that can be used to capture the typical "bath-tub" shape of lifetime processes, the instability of hazard cancellation values for the 61 to 90 DFD booking horizon makes this result uncertain. However, from an interpretation perspective, it is important to note that the potential existence of a "bath-tub" shape points to the presence two underlying behaviors. First, increased refund and exchange activities occurring close to flight departure reflect an increase in individuals' rescheduling activities. This may occur as individuals becoming more certain of their travel plans (and/or ability to travel) as their flights near departure. Second, among travelers purchasing tickets 21-90 DFD, those purchasing further in advance (*e.g.*, 61-90 DFD) may be more likely to refund or exchange their tickets due to the longer "time of expose" during which there are increased opportunities for traveling conflicts to arise. Further, given these individuals are more price sensitive, they may be more likely to reschedule shortly after their initial purchase in order to take advantage of lower fares. In contrast, the use of a logarithmic formulation would capture only the first effect.

Although the discrete formulation has the advantage of estimating average cancellation rates over user-defined time periods consistent with those used in revenue management systems, it lacks parsimony, and is therefore excluded from further estimation. Finally, since the evidence

to favor the polynomial over the logarithmic (*i.e.*, Weibull) formulation is inconclusive, both specifications are kept for further estimation[17].

Before introducing the effect of observed heterogeneity on the baseline hazard, it is important to note that although by means of linearity in log-odds the DTPO model is similar to the binary logit model, its focus is mainly towards predicting conditional probabilities of an event at different times (*i.e.*, there is a time-to-event process associated with the DTPO model). Furthermore, when time-to-event datasets exhibit a significant number of tied events and the ratio between events and non-events is low (as in the case of the ARC sample database), the results of the DTPO model will be similar with the results of an "exact" Cox proportional hazards model. Therefore, without loss of generality, the coefficients estimates of the DTPO model can be interpreted in the proportional hazard framework (*i.e.*, changes in the hazard rates due to a unit change in a given covariate). In this context, Table 2 presents the results in odds-ratio format of the DTPO estimation for the logarithmic and quadratic baseline hazards.

[Insert Table 2 about here]

The *Time of Ticketing* variables show that, relative to the reference category of 60-90 DFD, the conditional probabilities of cancelling tickets (*i.e.* cancellation rates) are twice as high for those tickets purchased 0-14 DFD and decrease linearly as the ticketing horizon increases. Conceptually, this reinforces the commonly held opinion that business passengers, who tend to book close to departure, are more likely to refund or exchange tickets than leisure passengers.

[17] With the addition of the void data a more complete set of specifications can be investigated.

Moreover, the significance of *Time of Ticketing* categories in Table 2 reinforces previous empirical evidence (Westerhof 1997) that the memoryless property of cancellation rates is violated. In view of current methods used to forecast cancellation rates, this finding is particularly important. Specifically, it suggests that determining cancellation rates only as extrapolations of previously realized values[18] may not be valid, as different cancellation rates will be observed depending on when a passenger tickets. However, prior to generalizing this result to airline cancellation models based on booking data, the impact of the distribution of cancellations not included in current analysis (*e.g.,* due to booking churn) will have to be explored. Discussions are currently underway with ARC to obtain the data needed to look at part of this booking churn process.

5.2 Interpretation of Covariates

Several covariates were also examined in the study, including the outbound departure day of week, presence of a Saturday night stay on the itinerary, group size, carrier, market, and pro-rated fare. Similar to the trends observed with ticketing periods, those variables typically associated with leisure passengers exhibit decreased cancellation rates. Those passengers with a Saturday night stay are 1.3 times more likely not to cancel than those passengers without a Saturday night stay. In addition, those passengers traveling in groups are 2.5 – 3.3 times less likely to cancel than passengers traveling alone. Moreover, those traveling with two or more

[18] The use of separate cancellation rates for each booking class only partially corrects for this problem, as some booking classes are available for purchase over the entire (or large portion) of the booking horizon.

individuals are less likely to cancel than those traveling with just one other person. In addition, those traveling outbound early in the work week (typically associated with business travelers) are more likely to cancel than those departing later in the week. Specifically, departures on Monday and Tuesday are 1.16 times more likely to cancel than those departures on Saturday and Sunday. Departures on Thursday and Friday departures are 1.28 times less likely to cancel than departures on Saturday and Sunday.

The effects of the last three categories of covariates: *Market*, *Carrier* and the *Pro-Rated Fare,* although significant, are more difficult to generalize because of endogenity concerns (the fare is highly correlated with market, and different carriers may impose different refund and exchange ticketing policies). Additional ticketing data is currently being obtained and will be used in future analysis to help decompose the effects of market, carrier, and fare.

Finally, it is important to note that seasonality, defined by the month of departure, was not statistically significant.

6. SUMMARY AND DIRECTIONS FOR FUTURE RESEARCH

This study demonstrated how discrete choice methods can be used to model airline passengers' cancellation behavior. In particular, a "cancellation" model for the outbound legs of an airline itinerary for groups ticketing within 90 days of flight departure was estimated based on occurrence of refund and exchange events in sample of ticketing data from the Airline Reporting Corporation (ARC). Compared to cancellation models reported in the literature or used in

practice, the proposed model is more "customer-focused" in the sense that it captures underlying behavior of passengers. In particular, a Discrete Time Proportional Odds (DTPO) model with a retrospective time scale shows that cancellation rates are not independent over time. Both the both time of ticketing and time until departure are important. In particular, leisure passengers, who are more likely to book further in advance of flight departure, are less likely to cancel than business passengers. However, as the flight nears departure, both leisure and business travelers are more likely to refund and exchange their tickets. Several other covariates typically associated with leisure passengers are also associated with lower cancellation rates. These include itineraries with a Saturday night stay, itineraries departing on Thursday and Friday, and itineraries involving groups.

In addition to providing new behavioral insights to airline passengers' refund and exchange behavior, this study is one of the first to be based on ticketing data from ARC. The occurrence of cancellations as measured by refund and exchange events was much smaller (less than 8% across all markets) than rates reported using booking information. For example, Smith, Leimkuhler, and Darrow (1992) report combined no-show and cancellation rates/proportions of 50% for American Airlines; while these proportions vary across carriers and markets and may have decreased over time, cancellation proportions of 30% or more are not uncommon today. Therefore, one of the questions that naturally arises from this study is: Why is there a large discrepancy in cancellation proportions between booking and ticketing data? One possible explanation is that booking data (and revenue management systems) are capturing the initial searching and pre-purchasing behavior of passengers. This would occur, for example, if a

business traveler called a travel agent to booked a reservation, but then waited a few days to either modify or pay for the reservation once travel plans became more firm. In general, the time period a reservation can be held is short – 24 to 48 hours. Thus, failure to pay for a reservation could lead to rebooking the same (or similar) itinerary multiple times. Part of this booking activity or booking "churn" as it is more commonly referred to may be represented in ARC void data. The void data represents tickets that were created, but not purchased and thus "voided" before a financial transaction was required. Future analysis will extend the cancellation analysis presented in this study to include cancellations from voids. This should permit a better linkage between the cancellation rates developed from ARC ticketing data and the booking data used in airline revenue management systems. Moreover, given the cancellation hazard of voids is expected to exhibit a very different pattern than the cancellation hazard of purchased tickets (with the former having a much smaller probability of surviving past two days), other modeling methodologies (including competing risks or a multi-stage estimation approach) will be explored. Finally, ticketing data from 2005 departures will be used in future analysis to validate the DTPO cancellation model and compare its performance to the models currently used by airlines and/or reported in the academic literature.

ACKNOWLEDGEMENTS

The authors are also grateful to Moshe Ben-Akiva, Chandra Bhat, Richard Carson, and Yijian (Eugene) Huang for their valuable insights.

REFERENCES

Allison, P. D. (1995). Survival Analysis Using SAS: A Practical Guide. Cary, NC, SAS Institute Inc.

Bhat, C. R. (1996). "A Generalized Multiple Durations Proportional Hazard Model with an Application to Activity Behaviour During the Evening Work-to-Home Commute." Transportation Research-B **30**(6): 465-480.

Bhat, C. R. (2000). Handbook of Transport Modelling, Edited by D.A. Hensher and K.J. Button, Elsevier Science Ltd.

Boyd, E. A. (2004). "Dramatic Changes in Distribution will require Renewed Focus on Pricing and Revenue Management Models." Journal of Revenue & Pricing Management **3**(1): 100-103.

Boyd, E. A. and R. Kallesen (2004). "The Science of Revenue Management when Passengers purchase the Lowest Available Fare." Journal of Revenue & Pricing Management **3**(2): 171-177.

Brown, C. C. (1975). "On the Use of Indicator Variables for Studying the Time-dependence of Parameters in a Response-time Model." Biometrics **31**: 863-872.

Chatterjee, H. (2001). Forecasting for Cancellations. AGIFORS Reservations and Yield Management Study Group, Bangkok, Thailand.

Cox, D. R. (1972). "Regression Models and Life-Tables." Journal of the Royal Statistical Society. Series B (Methodological) **34**(2): 187-220.

Dunleavy, H. and D. Westermann (2005). "Future of airline revenue management." Journal of Revenue & Pricing Management **3**(4): 380-383.

Garrow, L. and F. Koppelman (2004a). "Multinomial and Nested Logit Models of Airline Passengers' No-show and Standby Behavior." Journal of Revenue and Pricing Management **3**(3): 237-253.

Garrow, L. and F. Koppelman (2004b). "Predicting Air Travelers' No-show and Standby Behavior Using Passenger and Directional Itinerary Information." Journal of Air Transport Management **10**: 401-411.

Hensher, D. A. and F. L. Mannering (1994). "Hazard-based duration models and their application to transport analysis." Transportation reviews **14**: 63-82.

Hornick, S. (2004). "Fare Play." Airline Business **20**: 72-75.

Iliescu, D. C., L. A. Garrow, et al. (2006). Analysis of U.S. Airline Passengers' Refund and Exchange Behavior Across Multiple Airlines. European Transport Conference, Strasbourg.

Jain, D. C. and N. J. Vilcassim (1991). "Investigating household purchase timing decisions: A conditional hazard function approach." Marketing Science **10**(1): 11-23.

Kiefer, N. M. (1988). "Economic duration data and hazard functions." Journal of Economic Literature **26**(2): 646-679.

Kim, M. Y. and S. W. Lagakos (1990). "Estimating the Infectivity of HIV from Partner Studies." Annals of Epidemiology **1**: 117-128.

Lieberman, W. H. (2004). "Revenue management trends and opportunities." Journal of Revenue & Pricing Management **3**(1): 91-99.

Littlewood, K. (2005). "Forecasting and control of passenger bookings " Journal of Revenue and Pricing Management **4**(2): 111-123.

McCullagh, P. and J. A. Nedler (1989). Generalized Linear Models. New York, NY, Chapman and Hall.

Oliveira, A. V. M. (2003). "Simulating revenue management in an airline market with demand segmentation and strategic interaction." Journal of Revenue & Pricing Management **1**(4): 301.

Polt, S. (1998). Forecasting is Difficult – Especially if it Refers to the Future. AGIFORS Reservations and Yield Management Study Group, Melbourne, Australia.

Ratliff, R. (2006). Multi-flight Demand Untruncation with Recapture. AGIFORS RMD and Cargo Study Group, Cancun, Mexico.

Ratliff, R. and B. Vinod (2005). "Airline pricing and revenue management: A future outlook." Journal of Revenue & Pricing Management **4**(3): 302-307.

Shiboski, S. C. (1998). "Generalized Additive Models for Current Status Data." Lifetime Data Analysis **4**: 29-50.

Smith, B. C., J. F. Leimkuhler, et al. (1992). "Yield management at American Airlines." Interfaces **22**(1): 8-31.

Talluri, K. and G. van Ryzin (2004a). "Revenue Management under a General Discrete Choice Model of Consumer Behavior." Management Science **50**: 15-33.

Talluri, K. T. and G. J. Van Ryzin (2004b). The Theory and Practice of Revenue Management. New York, NY, Springer.

Thompson, H. R. (1961). "Statistical Problems in Airline Reservation Control." Operation Research Quarterly **12**: 167-185.

Thompson, W. (1977). "On the Treatment of Grouped Observations in Life Studies." Biometrics **33**: 463-470.

Van Ryzin, G. J. (2005). "Models of Demand." Journal of Revenue & Pricing Management **4**(2): 204-210.

Westerhof, A. (1997). CO2 in the Air. AGIFORS Reservations and Yield Management Study Group, Melbourne, Australia.

Wu, L. L. (2003). Handbook of the Life Course; Edited by Jeylan T. Mortimer and Michael J. Shanahan. New York, NY, Kluwer Academic/Plenum Publishers.

APPENDIX: OVERVIEW OF TICKETING DATA SOURCES

The ticketing data used for this study is distinct from the data collected as part of the United States Department of Transportation (US DOT) O*rigin 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[19]. 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 though 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 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 or the Data Base Products, Inc. website at www.airlinedata.com for additional information.[20]

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.

[19] "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).

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.

[20] The website describes the data and federally-mandated reporting requirement for U.S. airlines.

LIST OF TABLES AND FIGURES

TABLE 1: Comparison of Baseline Hazard Specifications

Baseline hazard functional form (Ψ_{ij})	*Parameter Estimates*	*LL*	*Pseudo-R*
Discrete : $\alpha_1 \cdot D_{ij1} + ... + \alpha_5 \cdot D_{ij5} + \alpha_6$	α_1= 1.438 α_2= 0.734 α_3= -0.422 α_4= 0.044 α_5= -0.055 α_6= -8.082	-24421.975	0.0180
Weibull : $\alpha + \beta \cdot \ln(j)$	α =-5.882 β=.524	-24308.124	0.0226
Polynomial : $\alpha + \beta_1 \cdot j + \beta_2 \cdot j^2$	α=-6.157 β_1 = -.081 β_2 =0.00077	-29294.407	0.0231

TABLE 2: Discrete Time Proportional Odds Estimation Results

Covariates	Baseline hazard - Logarithmic		Baseline hazard - Quadratic	
	Parameter	**z-stat**	**Parameter**	**z-stat**
Time (DFD=Days from Departure)				
DFD			0.95	-14.68
DFD^2			1.00	9.91
Log (DFD)	0.72	-17.27		
Time of Ticketing (reference category 60-90 days from departure)				
0 – 13	2.09	9.02	2.11	8.6
14 - 20	1.84	7.63	1.79	6.64
21 - 29	1.68	6.77	1.67	6.06
30 - 44	1.50	5.6	1.55	5.48
45 - 60	1.23	2.46	1.30	3.03
Group Size (reference= one person)				
2 people	0.43	-13.91	0.43	-13.92
3 or more people	0.30	-10.85	0.30	-10.85
Saturday Night Indicator				
Saturday night	0.77	-6.22	0.77	-6.27
Outbound Day of the Week (reference = Saturday or Sunday)				
Mon or Tues	1.16	3.37	1.16	3.35
Wed	1.0	-	1.0	-
Thurs or Fri	0.78	-5.43	0.78	-5.41
Market				
Bos-Sea	0.65	-7.35	0.65	-7.34
Hnl-Ord	0.45	-5.37	0.45	-5.37
Mia-Bos	0.58	-7.73	0.58	-7.73
Mia-Sea	1.33	3.46	1.33	3.5
Ord-Hnl	0.73	-3.84	0.72	-3.87
Sea-Bos	0.66	-6.62	0.66	-6.63
SeaMia	0.65	-5.19	0.65	-5.18
Carriers (masked information)				
Carrier 2	1.08	1.28	1.08	1.26
Carrier 3	0.39	-10.44	0.39	-10.45
Carrier 4	0.79	-2.33	0.79	-2.33
Carrier 5	1.05	0.71	1.05	0.72
Pro-Rated Fare				
Fare	1.001	18.84	1.001	18.88
Goodness of fit statistics				
Number of obs.	3,691,317		3,691,317	
LR chi2(df)	2687.34 (23)		2679.81 (24)	
Pseudo R^2	0.054		0.054	
Log likelihood	-23,526		-23,530	

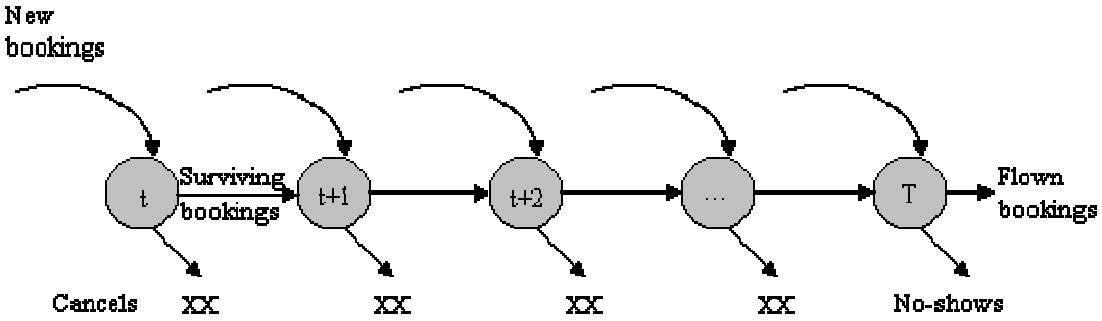

FIGURE 1: Cancellation Model Described by Westerhof (1997)

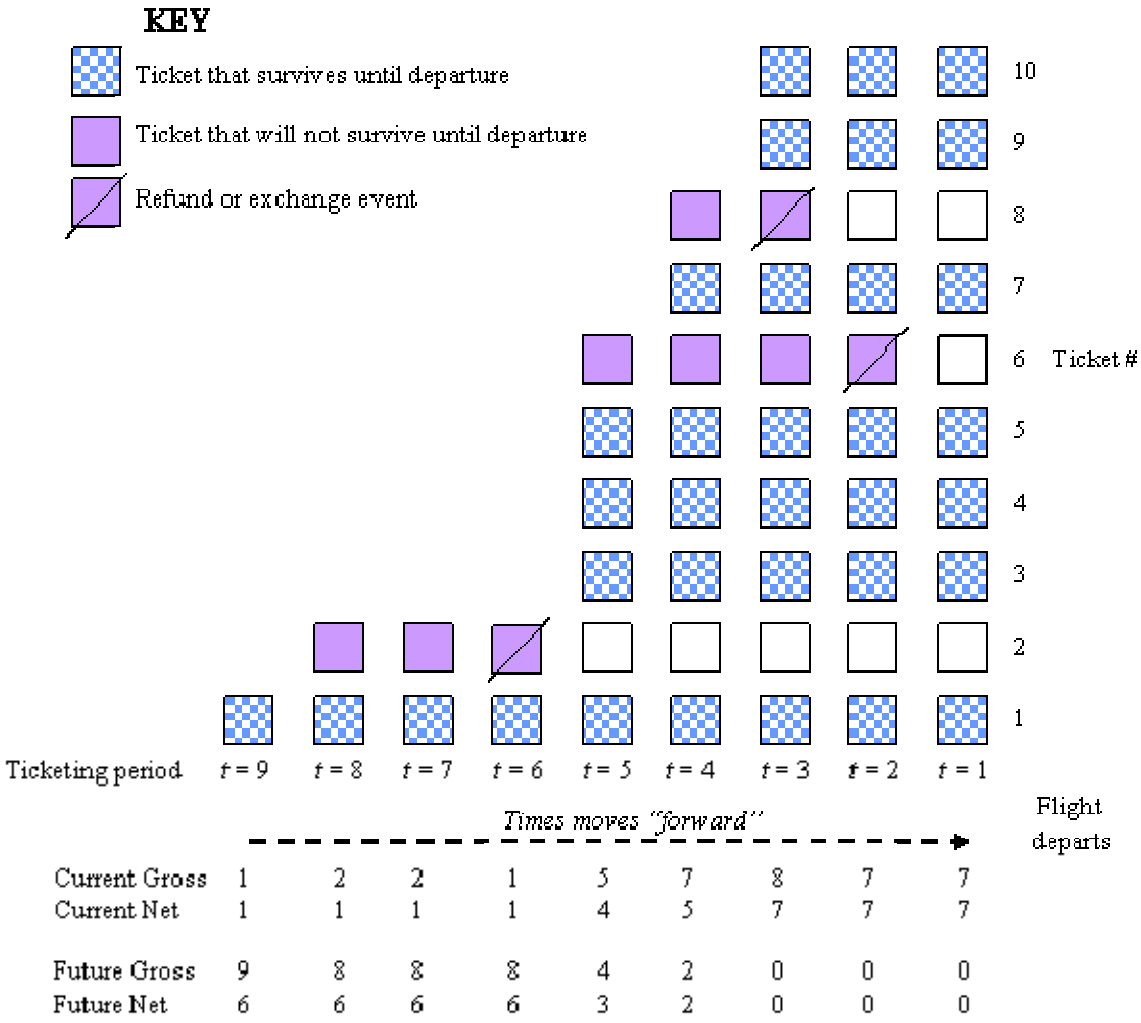

FIGURE 2: Cancellation Information Commonly Used in Practice

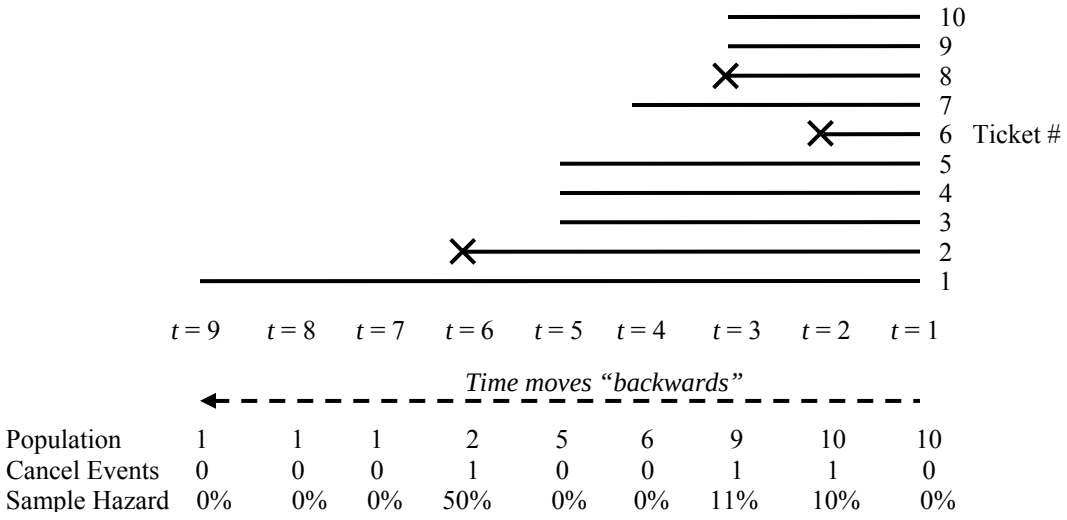

FIGURE 3: Sample Hazard Using a Retrospective Time Scale

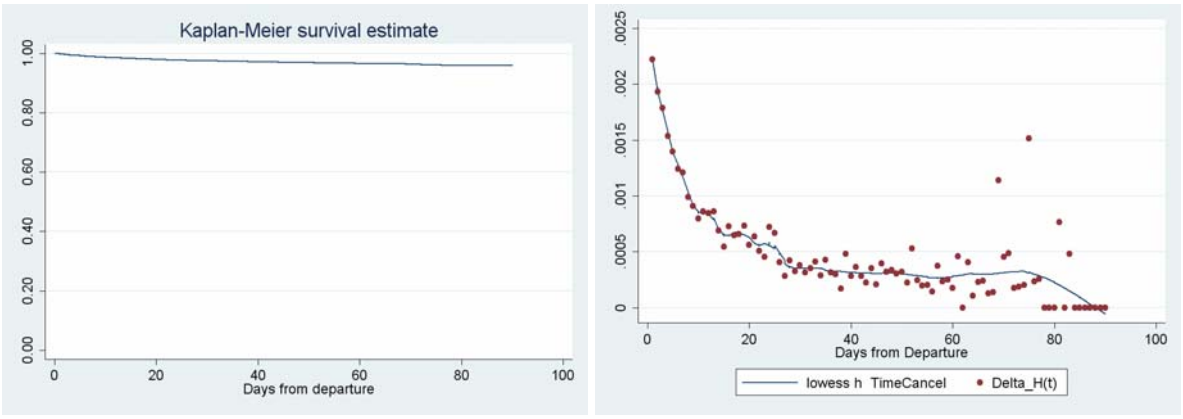

FIGURE 4: Non-parametric Estimators for S(t) and h(t)

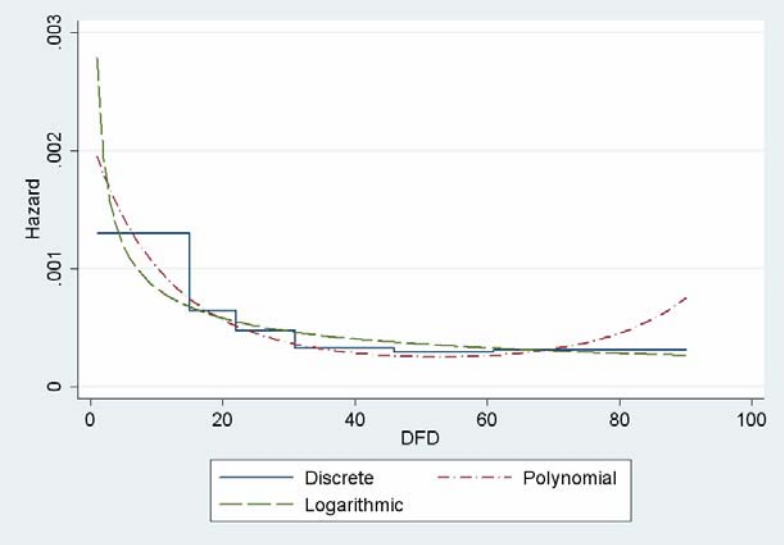

FIGURE 5: Non-parametric Estimators for S(t) and h(t)