**Customer heterogeneity in revenue management for railway services**

Pratt Hetrakul and Cinzia Cirillo

Department of Civil and Environmental Engineering

1173 Glenn L. Martin Hall, Bldg #088

University of Maryland

College Park, MD 20742, USA

Tel. +1 301 405-7768

Fax +1 301 405-2585

e-mail: [hetrakul@umd.edu](mailto:hetrakul@umd.edu)

e-mail: [ccirillo@umd.edu](mailto:ccirillo@umd.edu)

**Abstract**

Choice models based on random utility theory are being used in revenue management problems because of their ability to deal with customer heterogeneity and preferences over a set of multiple products. In this paper, both discrete and random mixture of logit are proposed to model ticket purchase timing over a finite sale horizon; in particular, parametric and non-parametric mixed logit are formulated and estimated. The analysis relative to intercity railway trips is performed on real data extracted from internet booking records, which contain very limited information that can be used for customers’ segmentation. The parameter estimates are then integrated into a nonlinear optimization framework, and fare strategies and seat allocations are derived. Results show that up to 20% increase in revenue can be obtained when methodologies based on customer heterogeneity are considered. Eventually, this study demonstrates that advanced choice models can be efficiently estimated on real data with severe limitations, and operational optimization methods can be based on individual choice behavior.

**1. Introduction**

Revenue Management (RM) often relies on the premise that different customers are willing to pay different amounts for a product. In simple models, customers’ choices are based on individual reservation prices that can be derived from historical data or from customer characteristics by estimating the parameters of the distribution across the population of interest. The distribution of reservation prices is traditionally modeled by assuming an aggregate demand function (Talluri and van Ryzin, 2004a). More recently, models based on disaggregate consumer choices have been proposed in the revenue management field and were successfully applied to a number of both theoretical studies and empirical applications (Talluri and van Ryzin, 2004b; Chaneton and Vulcano, 2011; Vulcano *et al*., 2010; Newman *et al*., 2012). Probabilistic models based on random utility theory (Ben Akiva and Lerman, 1985) can be used to model preferences of heterogeneous populations. Under this framework, customers can be segmented based on their socio-demographic characteristics (Cherchi and Ortuzar, 2003), divided into groups with similar preferences or price responses (Carrier, 2003; Hetrakul and Cirillo, 2013), or assumed to have each a different set of coefficients drawn from random distributions.

In models based on deterministic heterogeneity (mainly multinomial or nested logit models), customers are segmented by income and/or travel purpose. When segments are not observable, more sophisticated modeling approaches are needed. In latent class choice models (or finite mixture logit models), customers in each segment have the same parameters and a certain probability to belong to that segment, which is estimated along with model coefficients (Talluri and van Ryzin, 2004a). In recent years, researchers from different disciplines are increasingly making use of mixed logit model to capture individual taste heterogeneity. Mixed logit models have been applied to many complex transportation phenomena, such as the analysis of the value of travel-time (Greene *et al.*, 2006), airport choice (Hess and Polak, 2005), airline choice (Carrier, 2003), vehicle choice (Brownstone *et al.*, 2000; Hess *et al.*, 2006), and congestion pricing (Bhat and Castelar, 2002).The model has an advantage of being able to retrieve random variations in sensitivities across customers, which often leads to significant improvements in model accuracy. However, important issues arise with the use of the random coefficient model in mixed logit. Modelers need to make a priori assumptions on the distribution for each random coefficient, with the majority of applications relying exclusively on the normal distribution. Fosgerau (2006) emphasized that an inappropriate distribution selection can lead to extreme bias. More specifically, Fosgerau and Bierlaire (2007) indicated that using only the goodness-of-fit to compare models does not provide valid conclusions of the appropriateness in the distribution of the random parameter model. They proposed a semi-nonparametric (SNP) specification to test if a random parameter of a discrete choice model follows a given distribution. Non-parametric approaches have been also suggested by Bastin *et al.*, 2010, who propose B-splines to parameterize the inverse cumulative distribution of model parameters. B-splines are known to provide concise formulation for curves that are composed of the polynomial pieces, thereby automatically controlling the overall curve smoothness (Farin, 1991).

As pointed out by Chaneton and Vulcano in 2011: “*Accounting for customer choice behavior involves two major challenges. The first one is to model the choice decision of a customer at a particular moment in time, and to estimate the parameters that describe that decision. The second one is to incorporate this sophisticated demand model in the optimization module of an RM system*”. There are a number of studies, mainly published on OR journals, related to revenue management models that incorporate customer choice. Talluri and van Ryzin (2004b) study a revenue management problem over a single flight leg; customers choose among the fare classes that are available for purchase and the optimization problem adjusts the assortment at each period in order to maximize the total expected revenue. This seminal paper has generated a florid literature on the subject. There are a number of papers that extend this work to a flight network; see Gallego *et al.* (2004), Liu and van Ryzin (2008), Kunnumkal and Topaloglu (2008), Gallego et al. (2011), Vossen and Zhang (2012) and Meissner et al. (2012). In addition, some recent work attempts to solve assortment problems under the nested logit model, which allows for correlations between the products available to a particular customer. Davis *et al.* (2011) give a linear programming formulation of the assortment problem under the nested logit model. Li and Rusmevichientong (2012) give a greedy algorithm for the same problem. Gallego and Topaloglu (2012) show how to impose a variety of constraints on the offered assortment when customers chose according to the nested logit model. However, as stated by Rusmevichientong *et al*., 2013, all of the current work on nested logit model is under the assumption that there is a single customer segment and more importantly with known choice model parameters. Rusmevichientong *et al*., 2013 also solve an assortment optimization problem under the mixture-of-logits model, where the parameters of the choice model are random. The authors motivate the randomness in the choice model parameters by the fact that in real problems there are multiple customer segments, each with different preferences for the products, and the segment of each customer is unknown to the firm when the customer makes a purchase. Again most of the research that involves discrete choice models for revenue management is based on simulated data or assumes that customers’ taste is known.

This paper models strategic customer behavior for revenue management using advanced forms of discrete choice models and demonstrates that it is possible to estimate individual heterogeneity using both discrete and continuous mixture of logit models. The models are estimated on real observations extracted from internet booking data, in the context of intercity rail trips for which segments cannot be determined on the base of socio-demographic characteristics or trip purposes. Furthermore, the parameters estimated are integrated into an optimization framework that maximizes fares and gives the railway operator indications about dynamic pricing strategies.

The remaining of this paper is organized as follows. Section 2 introduces the problem setting and model assumptions. The formulation of disaggregate choice models for revenue management is presented in Section 3, while the estimation of the aggregate demand function is given in Section 4. In Section 5, the revenue optimization problem is formulated; the problem jointly considers pricing and seat allocation while accounting for the network characteristics of the railway system. The optimization results are presented in Section 6 in terms of revenue improvement, seat allocation strategy, and passenger response to the new fare policies. Finally, conclusions and future research directions are given in Section 7.

**2. The Railway Revenue Management problem: data and modeling framework**

The railway industry is quite similar to the airline industry, as it sells seats for a transport from an origin to a destination. The price depends on journey, class, and date of booking. According to Thao Ly (2012) thecharacteristics common for the definition of RM strategies include:

- *Inflexible capacity*: the number of seats on a train is fixed;

- *Variable and uncertain demand*: the demand is different depending on the time of day, day of week, origin-destination pairs, and purpose of travel;

- *Perishable inventory*: the unoccupied seat cannot be inventoried, and the revenue of that seat is missed;

- *Low marginal cost and high production cost*;

- *Product can be sold in advance*;

- *Heterogeneous customers* with different segments, varying by age, purpose of travel, independent or group, time of day, day of week, etc.

However, differences exist betweenthe RM problem for railway industry and airline industry (Thao Ly, 2012).

- *A railway trip generally consists of more legs:* several station stops are usually in between the trip origin and the final destination. Each leg, defined by a pair of stops, must be determined in terms of opportunity cost (bid price) or capacity allocation;

- It is more difficult to forecast demand and to have the accurate number of travelers on a railway service: no check-in procedure, open tickets (rail pass) generally allow passengers to travel on any rail service, walk up tickets are very common (a large number of passengers purchase their tickets on the day from the station);

- It is less difficult to deal with train overbooking: train passengers are often allowed to stand during the journeys hence increasing capacity beyond the number of seat.

More specifically, in our modeling framework for railway services, we suppose that the number of seats is fixed, that the demand varies depending on the time of day, day of week, origin-destination pairs, unoccupied seats do not contribute to revenues, that tickets are sold in advance and that customers are heterogeneous. Moreover, the network is composed of several legs and we do not allow overbooking. Based on these assumptions, we propose an optimization framework that solves a joint pricing and seat allocation problem for revenue management. The RM optimization problem incorporates two modeling components: (1) disaggregate passenger choice models and (2) aggregate demand functions. The first component explicitly models customers’ heterogeneity and is estimated with discrete choice methods; it predicts the timing in which passengers purchase the ticket as a function of fare and other trip attributes (Section 3). The second component forecasts aggregate demand volumes between the origin-destination pairs considered. The aggregate demand functions account for passengers deciding not to travel with this service or for induced demand due to advantageous fare policy. Passenger volumes are estimated using log-linear regression, where independent attributes are fare and trip attributes (Section 4). The passenger choice models and the demand functions are then incorporated into a revenue optimization system that maximizes expected ticket revenue per each train trip in a network setting (Section 5).

The case study presented in this paper is based on monthly ticket reservation data of intercity passenger railway trips. The original dataset, whose source cannot be revealed for confidentiality issues, contains 110,828 records. The railway network is composed of sixteen stations (Figure 1); only trains travelling northbound are considered. Data refer to business class passengers, which constitute the predominant market of this rail service. The study focuses on trains that departed on Friday, March13, 2009 at four different departure times: 5:00 AM, 9:00 AM, 1:00 PM, and 4:00 PM; these four departure times (named Train#1 to Train#4, respectively) are selected to represent railway traffic conditions at different times of day.

It should be noted that for each reservation, we only observe fare on the day the ticket is purchased. To accommodate choice modeling, fares on other days in the sale horizon are approximated from the actual data by using average fares for each booking day considered from the monthly data available. Only confirmed bookings are used for model estimation, which implies that the mode choice decision has already been made by the passengers. Due to the limited number of observations available for some of the stations in the network, our optimization problem is based on the five stations in the second column of Figure 1. For a fair revenue comparison, seats currently occupied or empty at the excluded stations are not allowed to contribute to the revenue. In order to reduce the number of choice models to be estimated, the railway network has been aggregated into four station groups based on the geographical locations shown in the third column of Figure 1. Ten disaggregate choice models are estimated for trips between and within these station groups to account for different behavior on short, medium and long distances.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | |  |  | |  |  |
|  | Station1 | |  | **Station1** | |  |  |
|  |  |  |  |  |  |  |  |
|  | Station2\* | |  |  |  |  |  |
|  |  |  |  |  |  |  | Station group 1 |
|  | Station3\* | |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  | Station4 | |  | **Station2** | |  |  |
|  |  |  |  |  |  |  |  |
|  | Station5\* | |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  | Station6\* | |  |  |  |  | Station group 2 |
|  |  |  |  |  |  |  |  |
|  | Station7 | |  | **Station3** | |  |  |
|  |  |  |  |  |  |  |  |
|  | Station8 | |  | **Station4** | |  |  |
|  |  |  |  |  |  |  |  |
|  | Station9 | |  | **Station5** | |  |  |
|  |  |  |  |  |  |  |  |
|  | Station10\* | |  |  | |  |  |
|  |  |  |  |  |  |  | Station group 3 |
|  | Station11\* | |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  | Station12\* | |  |  | |  |  |
|  |  |  |  |  |  |  |  |
|  | Station13\* | |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  | Station14\* | |  |  | |  | Station group 4 |
|  |  |  |  |  |  |  |  |
|  | Station15\* | |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  | Station16\* | |  |  | |  |  |
|  |  | |  |  | |  |  |

\* Indicates stations excluded from the optimization.

Figure 1 Station numbering

**3. Discrete choice model formulation**

In this Section we formulate discrete choice models for revenue management; we present classical multinomial logit (MNL), and both discrete (LC) and continuous mixture of logit models (ML). Mixed logit are specified using parametric and non parametric random coefficients (in the form of B-spline) formulations. Discrete choice models are intended to capture heterogeneity over time and across customers. MNL models are able to accommodate only deterministic customer heterogeneity through socio-demographic variables, which are not available in our internet booking data. In order to have a better representation of users’ heterogeneity, we propose a more advanced form of discrete choice models that allow the estimation of random heterogeneity. Latent class models segment the population into a finite number of classes and estimate the probability for each customer to belong to each class. ML models estimate random taste heterogeneity and are used here to determine both parametric and non-parametric distribution of preferences over the price attribute. Results from model estimation are given in Table 1.

**3.1 Multinomial logit model (MNL)**

Customer *i* is assumed to choose the day *j* when to buy the ticket according to random utility maximization. In our RM framework, each customer associates to the ticket purchased the utility:

(1)

where the independent variables and their associated indexes are:

= Booking day,

=Booking period,

= Number of day before departure of booking day

= Fare of booking day ($)

= Weekend dummy (1 if departure is on weekend, 0 otherwise)

= error term, which is assumed to be identically and independently distributed Gumbel

Our data indicate that 98% of the tickets were purchased within 30 days before departure, thus it is reasonable to assume that the choice set is constituted of 31 days, from 30 days before departure (booking day 1) until departure date (booking day 31). The 31 booking days are grouped into six booking periods such that booking days within the same booking period have approximately the same number of reservations. These six booking periods are: (1) Booking day 1 to booking day 11, (2) Booking day 12 to booking day 20, (3) Booking day 21 to booking day 25, (4) Booking day 26 to booking day 29, (5) Booking day 30, and (6) Booking day 31.

Fare and advance booking variables are aimed to capture passenger tradeoff behavior between early booking with cheaper fare and late booking with higher fare. The actual fare of the railway service varies depending on departure time of day, day of week, how early the reservation is made in advance, and customer demand for each departure. The model specification allows passengers to have different price sensitivity across booking over the sale horizon.

The weekend dummy variables capture corresponding unobserved effects. Generally speaking, the trip purpose is believed to influence ticket booking time. For instance, leisure oriented passengers are expected to plan their trip in advance, while business passengers are likely to be less sensitive to price and tend to book later in time. The weekend dummy coefficients assume different values across booking periods to account for the departure day effect toward ticket booking time. The MNL models are estimated with AMLET (Another Mixed Logit Estimation Tool) (Bastin, 2011).

**3.2. Latent class model (LC)**

In the latent class model specification, passengers are segmented on the base of their trip characteristics because we believe that passengers traveling at particular time periods are relatively homogeneous in their behavior. It should be noted that internet booking data used for this analysis contain very limited socio-demographic variables and no information on trip purpose. For class specific choice model, the customers’ utility includes fare ($) and advance booking variables. The utility of choice for a specific class can be written as:

For the class membership model, other elements are extracted from the booking data to segment passengers’ demand and capture heterogeneity of behavior across different classes of customers:

*Departure time of day (TOD):*Dummy variables are used to indicate whether the trip is taken on a particular time of day. We use six departure times (Jin, 2007) for the intercity trip which are (1) early morning (0:00 am-6:29 am), (2) a.m. peak (6:30am-8:59 am), (3) a.m. off-peak (9:00 am-11:59 am), (4) p.m. off-peak (12:00pm-15:59 pm), (5) p.m. peak (16:00 pm-18:29 pm), and (6) evening (18:30 pm-23:59 am). Five departure times of day (except evening) are used for the class membership model.

*Departure day of week (DOW):*Dummy variables are used to indicate whether the trip is taken on a particular day of week; this results into six dummy variables for the class membership model, one for each day of the week (except Sunday).

Thus, the utility that customer belongs to class has the form:

Where is class specific constant.

The latent class model is estimated with Latent Gold Choice 4.5, a software package by Statistical Innovations specifically designed for latent class choice modeling (Vermunt and Magdison, 2005).

**3.3 Mixed logit model (ML)**

In mixed logit models taste heterogeneity is recovered by assuming that coefficients in the utility function are randomly distributed; the underlying distributions are specified by the analyst. Random distributions proposed for ML estimation include: uniform, normal, lognormal, truncated normal, triangular and Johnson SB. (Hensher and Greene, 2003; Train, 2009). Our parametric form of ML assumes that the fare coefficient is log-normally distributed. This distribution provided a better fit when compared to the normal distribution and ensured a negative value for the entire population (fare represents a disutility and is expected to have a negative coefficient). The resulting utility for passenger booking the ticket on day which falls within the booking period can be expressed as:

Where is distributed as normal with mean and standard deviation . The price sensitivity () has mean and variance equal to and . respectively. The price coefficient of mixed logit in Table 1 represents the mean ( and standard deviation ( of the log price sensitivity.

As stated in the Introduction, the parametric approach is prone to bias and the process of selecting the “right” distribution is time consuming. The selection process relies on trial and error, on model fit quality measures (often assessed using final log-likelihood values) and on the sign and value of the coefficient(s) of interest. Therefore, we have attempted to extend the use of non-parametric random coefficient ML to choice model for revenue management. If we suppose that random variables are independent and have a bounded support, an elegant way to construct the non-parametric distribution is constituted by the B-spline functions. The bounded support is considered consistent with the underlying behavior assumption (Bastin et al., 2010). By assuming that price sensitivity follows a uniform B-splines curve of degree three, the utility function has the form:

Where the piecewise cubic polynomial functions form an easily computable basis for B-spline function in [0,1]. The coefficients are called control points, and the function depends on the choice of special points of reference, called knots in [0,1]. With these basis and knots choice, is monotonically increasing if the control points have the property: . The constrained optimization proposed in Bastin *et al.* (2010) is adopted to estimate the knot points ) of the non-parametric distribution and to ensure its monotonicity. For price coefficient, seven control points have been estimated for each B-spline, where and give the bounds of the distribution (shown in Table 1), and the knot vector is defined on the percentile 0, 0.25, 0.5, and 1. More details on the numerical procedure to solve this simulated log-likelihood problem can be found in Bastin *et al.* (2010).

**3.4 Models estimation results**

Due to space limitations, we only show the model estimation results for passenger traveling from station group 3 to group 2 (medium distance journey) in Table 1. For this market segment, the non-parametric ML model provides the best statistical fit with the adjusted rho-squared being the highest among all the estimated models. As expected, parametric ML and LC models outperform MNL models, although the improvement in the value of the adjusted rho-squared is only marginal but statistically significant.

Results obtained with the MNL indicate that disutility is associated with advance booking and that the lack of flexibility to change travel plan is negatively perceived. Passengers generally prefer to hold their purchase and pay for the product as late as possible. The price sensitivities in all the booking periods have negative sign and are statistically significant at the 5% significance level. The decreasing magnitude of price sensitivity as booking period approaches departure is in line with the expectation that passengers are more sensitive to fare at the beginning of the sale horizon. As time approaches departure, passengers become less sensitive to fare especially on the departure day and more concerned about seats availability. The weekend dummies show the expected pattern (except for the last period); the value decreases as the booking period approaches departure indicating that passenger traveling on weekends generally purchase ticket in advance compared to passenger traveling on weekdays. It is also expected that weekend travelers are primarily leisure travelers.

The mixed logit model accounting for heterogeneity in price sensitivity with log-normal distribution shows a better model fit compared to the MNL and the LC model based on the adjusted rho-squared value. The ML model results show the expected sign for advance booking, fare and weekend dummies as observed in the MNL models. The estimates are all statistically significant at the 5% significance level.

In this context, the main advantage of the latent class specification over multinomial logit model is the ability to identify distinct groups of passengers. Passengers are assumed to have different preferences about trip schedule which results in different willingness to pay (WTPs) for delaying ticket purchase across the two classes considered. The results obtained with the latent class model are coherent with those obtained with the multinomial logit and mixed logit models except that passengers in class 1 have positive price coefficient meaning that they are insensitive to price. Given that the magnitude of the coefficients within the same model cannot be compared between different classes due to the scale parameter (Carrier, 2008), the ratio of advance booking coefficient to the price coefficient is calculated to represent the willingness to pay (WTP) to delay the ticket purchase for one day. We calculate the average WTPs obtained from the six price coefficients of each passenger class across three markets. The lower WTP value ($2.5) can be associated to leisure passengers, who generally know their travel plans in advance; while the higher value ($35.5) can be associated to business travelers who generally book their tickets closer to the departure date.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | |  | | | |  | | | |
|  | |  | |  | |  | |  | |  | |
|  |  | |  | |  | |  | |  | |  |
|  |  | |  | |  | |  | |  | |  |
|  |  | |  | |  | |  | |  | |  |

Table 1 Passenger choice model estimation

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **MNL** | |  | **ML** | |  |  | **Spline** | |  | **LC** | | | | | | |
|  |  |  |  |  |  |  |  |  |  |  | **Choice Model** | **Class1** | |  | **Class2** | |  |
| Variable | Est | T-Stat |  | Est | T-Stat |  |  | Est | T-Stat |  | Variable | Est | T-Stat |  | Est | T-Stat |  |
| advbk | -0.220 | 17.952 | \* | -0.450 | 89.143 | \* | advbk | -0.568 | 72.854 | \* | advbk | -0.199 | -11.688 | \* | -0.539 | -12.988 | \* |
| price.period1 | -0.008 | 1.942 |  | -6.225 | 54.812 | \* | price1 | -0.406 | 47.219 | \* | price.period1 | -0.083 | -10.519 | \* | 0.023 | 1.497 |  |
| price.period2 | -0.017 | 6.735 | \* | (3.630) | 56.304 | \* | price2 | -0.406 | 37.111 | \* | price.period2 | -0.084 | -11.562 | \* | -0.010 | -0.746 |  |
| price.period3 | -0.015 | 8.598 | \* |  |  |  | price3 | -0.406 | 27.721 | \* | price.period3 | -0.084 | -12.478 | \* | -0.012 | -0.950 |  |
| price.period4 | -0.013 | 9.154 | \* |  |  |  | price4 | -0.406 | 11.498 | \* | price.period4 | -0.076 | -11.299 | \* | -0.024 | -2.042 | \* |
| price.period5 | -0.004 | 3.109 | \* |  |  |  | price5 | 4.569 | 7.813 | \* | price.period5 | -0.106 | -6.418 | \* | -0.015 | -1.235 |  |
| price.period6 | -0.002 | 1.533 |  |  |  |  | price6 | 6.698 | 15.579 | \* | price.period6 | -0.059 | -8.864 | \* | -0.035 | -2.974 | \* |
|  |  |  |  |  |  |  | price7 | 6.698 | 15.578 | \* |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  | **Class Model** | **Class1** | |  | **Class2** | |  |
|  |  |  |  |  |  |  |  |  |  |  | Class Size | 0.575 | |  | 0.426 | |  |
|  |  |  |  |  |  |  |  |  |  |  | Variable | Est | T-Stat |  | Est | T-Stat |  |
|  |  |  |  |  |  |  |  |  |  |  | Intercept | 0.785 | 8.454 | \* | -0.785 | -8.454 | \* |
|  |  |  |  |  |  |  |  |  |  |  | Monday | 0.186 | 2.215 | \* | -0.186 | -2.215 | \* |
| wknd.period1 | 1.473 | 8.601 | \* | 0.468 | 12.337 | \* |  | 0.176 | 1.361 |  | Tuesday | -0.408 | -5.255 | \* | 0.408 | 5.255 | \* |
| wknd.period2 | 0.833 | 5.862 | \* | 0.491 | 42.742 | \* |  | 0.153 | 1.206 |  | Wednesday | -0.413 | -5.272 | \* | 0.413 | 5.272 | \* |
| wknd.period3 | 0.396 | 3.485 | \* | 0.518 | 36.889 | \* |  | 0.245 | 1.396 |  | Thursday | -0.456 | -5.764 | \* | 0.456 | 5.764 | \* |
| wknd.period4 | 0.352 | 4.280 | \* | 0.355 | 6.400 | \* |  | -0.495 | 4.276 | \* | Friday | -0.317 | -3.977 | \* | 0.317 | 3.977 | \* |
| wknd.period5 | -0.551 | 5.574 | \* | -0.163 | 3.125 | \* |  | -0.603 | 5.292 | \* | Saturday | 0.180 | 1.163 |  | -0.180 | -1.163 |  |
| wknd.period6 | 0.498 | 7.342 | \* | 1.332 | 23.902 | \* |  | 0.022 | 0.177 |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  | Early morning |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  | AM peak | -0.917 | -11.348 | \* | 0.917 | 11.348 | \* |
|  |  |  |  |  |  |  |  |  |  |  | AM off peak | -0.305 | -5.495 | \* | 0.305 | 5.495 | \* |
|  |  |  |  |  |  |  |  |  |  |  | PM off peak | -0.009 | -0.182 |  | 0.009 | 0.182 |  |
|  |  |  |  |  |  |  |  |  |  |  | PM peak | -0.190 | -3.484 | \* | 0.190 | 3.484 | \* |
| No. of observation |  | 11,536 |  |  | 11,536 |  |  |  | 11,536 |  | No. of observation | |  |  |  |  |  |
| Rho-squared: |  | 0.4327 |  |  | 0.4525 |  |  |  | 0.6736 |  | Rho-squared: |  |  |  |  | 0.4444 |  |
| **Adjusted rho-squared:** | | **0.4323** |  |  | **0.4523** |  |  |  | **0.6733** |  | **Adjusted rho-squared:** | | |  |  | **0.4437** |  |
| Log-likelihood at optimal | | -22,474 |  |  | -21,688 |  |  |  | -12,929 |  | Log-likelihood at optimal | | |  |  | -22,011 |  |
| Log-likelihood at zero | | -39,614 |  |  | -39,614 |  |  |  | -39,614 |  |  |  |  |  |  |  |  |
| Log-likelihood at constant | | -22,530 |  |  | -22,530 |  |  |  | -22,530 |  |  |  |  |  |  |  |  |

\*Indicates statistically significance at 5% confidence level. Parenthesis indicates standard deviation in the mixed logit.

**4. Aggregate demand function estimation**

In this Section, the passenger daily demands for each origin destination pair are estimated. A log-linear regression has been adopted to calibrate the demand function. The log-linear form restricts the estimated passenger demand (dependent variable) to be strictly positive and bounded at zero. Demand function estimation is based on the same dataset used for choice model calibration and contains 110,828 reservation records.

The optimization procedure is applied to the five selected stations in Figure 1, column 2. For a fair revenue comparison, we set the allowable seat capacity in each segment equal to the number of seats currently occupied by the selected five stations. Thus, seats occupied by the excluded stations and seats empty are not allowed to contribute to revenue. These seats are extracted from the total seat capacity in the constraints. The decision variables are the fares for each origin destination pair on each booking day over the sale horizon () and the fraction of demand to be accepted for each origin destination pair ().

The independent variables which enter the final model include the intercept, the square of the advance booking (), fare (), weekend dummy () indicating whether the departure day is on weekends, and booking period specific intercepts indicating whether the departure day is in a particular booking period (with the period defined in Section 4.2.2). The square of the advance booking is motivated by the non-linear relationship between advance booking and number of passenger booking observed from the data (Sibdari *et al.*, 2008). The weekend dummy accounts for the seasonality of train departing on different days. The specification of the log-linear demand function can be expressed as follows:

where:

= Number of reservations on booking day *j* for origin *o* destination *d*

= Square of the advance booking for booking day *j*

= Average fare ($) of booking day *j* for the observed departure day

= Weekend dummy (1 if departure day is weekend, 0 otherwise)

= Booking period dummy (1 if booking day *j* is in period *k*, 0 otherwise)

= Intercept

The daily passenger demand for each origin destination pair is then the summation of passenger reservation on each booking day over the sale horizon:

In Table 2 results from the estimation of the demand function between origin 16 and destination 8 are presented.

Table 2 Demand function estimation of selected market (origin 16 destination 8)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Coeff** | **Std Err.** | **T-Stat** | **P>|t|** | **[ 95% confidence** | **Interval]** |
| Advbksq | -0.0018 | 0.0003 | -6.46 | 0 | -0.0024 | -0.0013 |
| Fare | -0.0050 | 0.0015 | -3.36 | 0.001 | -0.0079 | -0.0021 |
| wkndmy | -0.2318 | 0.0664 | -3.49 | 0.001 | -0.3622 | -0.1014 |
| period 1 | -3.3826 | 0.2444 | -13.84 | 0 | -3.8623 | -2.9029 |
| period 2 | -3.2735 | 0.1811 | -18.08 | 0 | -3.6289 | -2.9181 |
| period 3 | -2.8051 | 0.1696 | -16.54 | 0 | -3.138 | -2.4723 |
| period 4 | -1.8583 | 0.169 | -10.99 | 0 | -2.1901 | -1.5265 |
| period 5 | -0.7223 | 0.2119 | -3.41 | 0.001 | -1.1381 | -0.3065 |
| constant | 6.4649 | 0.3018 | 21.42 | 0 | 5.8726 | 7.0572 |
| No. obs | 886 |  |  |  | R-squared | 0.657 |
| F( 8, 877 ) | 209.65 |  |  |  | Adj R-squared | 0.654 |
| Prob>F | 0 |  |  |  | Root MSE | 0.834 |

**4.1 Demand conversion**

Demand derived in Section 4 is converted to demand for each departure time using a conversion factor. The conversion factor is obtained from historical data by observing distribution of daily passenger demand across daily departure times. Our analysis focuses on trains that depart from the south end station (station 16\*) at four different departure times: 5:00 AM., 9:00 AM., 1:00 PM., and 4:00 PM. Departure times () at each station within the selected network where passengers are loaded into the train considered are shown in Table 3.

Table 3 Departure time for each origin

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Departure time ()** | | | | |
| **Origin** | **Train#1** | **Train#2** | **Train#3** | **Train#4** |
| 5 | 7:30AM | 11:30AM | 3:30PM | 6:30PM |
| 4 | 8:00AM | 12:00PM | 4:00PM | 7:00PM |
| 3 | 8:44AM | 12:44PM | 4:44PM | 7:44PM |
| 2 | 10:55AM | 2:55PM | 6:55PM | 9:55PM |

The conversion factor is denoted as where *t* represents the departure time and *o* represent the origin station. It is assumed that does not vary by destination (; from an empirical data analysis we found that is not significantly different across destinations. The passenger demand by departure time can be computed from the estimated demand function as follows:

where

= Number of passenger demand from origin *o* to destination *d*, at departure time *t*.

= Conversion factor from daily demand to demand by departure time.

= Estimated passenger demand on the entire day from origin *o* to destination *d* obtained from the demand function.

**5. The revenue optimization problem**

The problem is formulated to optimize revenues from the sale of coach class tickets for each train trip from the south end to the north end stations. The railway network is composed of a total of sixteen stations. However, in our optimization problem we just account for five of them, because passenger demand in the excluded markets is low and insufficient for estimation. Seats currently occupied on the excluded stations and seats currently empty are not allowed to contribute to the revenue. These seats are extracted from the total seat capacity in the constraint. The decision variables for the revenue optimization problem are: fare for each origin destination pair on each booking day over the sale horizon () and the fraction of demand to be accepted for each origin destination pair (). The purchase time probability is equivalent to the passenger share that purchases ticket on the booking day considered. The choice models specified as MNL, LC and parametric and non-parametric ML are used in this optimization framework.

**5.1 Notation**

Before giving the details of the optimization framework we provide the notation that will be used in the subsequent Sections.

= Number of stations (5 stations in selected problem)

= Boarding station index

= Alighting station index

= Passenger demand from origin *o* to destination *d* at departure time *t*

= Acceptance ratio; a fraction of demand ( ) to be accepted

= Fare for origin destination on booking day

= Total coach class seat capacity; equal to 260 (Railway Technology, 2011)

= Number of seats currently occupied by the excluded stations and seats currently empty

= Revenue per train trip ($) from south end to north end station

= Probability that passenger purchases the ticket on booking day for origin destination

**5.2 Revenue optimization with MNL choice model**

In the MNL model formulation the choice probability is expressed as:

where is the deterministic utility of booking on day from origin to destination . The denominator represents the sum of the exp of the utility to book the ticket on the 31 day time horizon considered.

The revenue optimization problem that makes use of the MNL customer choice is as follows:

(1)

The first two summations in eq. 1 account for passenger demand departing from origin to destination within the same train trip. Note that this index has different values for each station. The value represents station 5 and increase up to for station 1. The index is used to represent the corresponding departure time for each station which loads passenger into this train. The third summation computes the expected ticket revenue for each origin destination pair, which is computed as a weighted average of fare () and probability of being purchased () over the entire sale horizon. The fraction of demand that can be accepted for each origin destination pair is denoted as .

The optimization problem is subject to the following constraints:

* Capacity constraint:

for all .

The capacity constraint restricts the number of accepted passengers in each segment to be within the allowable seat capacity which is equals the total seat capacity subtracted by the seats currently occupied by the excluded stations and seats currently empty . The decision variable is an acceptance rate which is used to control the number of accepted passengers to be within the allowable seat capacity. Figure 2 represents the capacity constraint where the line connecting each origin destination pair represents the passenger demand.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Station1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Station2 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Station3 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | *+1* |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Station4 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Station5 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

**Figure 2 Capacity constraint**

* Fare policy constraint:

Fare policy constraint restricts the fare to be within the bound. The fare bound is obtained from the dataset and is based on the maximum and minimum of the average fare for a particular time of day, and day of week. This fare bound is also adjusted to ensure that fare for short distance trips does not exceed the fare of longer distance trips.

**5.3 Revenue optimization with LC choice model**

In latent class model, let represents individual and represents alternative from in the choice set . The model form can be written as:

where:

is class index;

is class membership explanatory variable

is class specific choice model explanatory variable

The utility function of alternative given the customer is in the class can be written as:

The class specific choice probability of alternative can be expressed as:

,

Where are the unknown parameters of the class-specific choice model. The utility function of customer belonging to class can be written as:

The probability of belonging to the latent class *s* can be written as:

Where are the unknown parameters for class membership model.

The application of latent class (LC) choice model in the revenue optimization allows for passenger taste heterogeneity in a discrete approach. Given the choice probability of LC specified for the choice model as:

The corresponding optimization problem can be expressed as:

The optimization is subjected to the same set of constraints in the MNL optimization.

**5.4 Revenue optimization with ML choice model**

Mixed logit probabilities are the integral of standard logit probabilities over a density of parameters (Train, 2003). Choice probabilities of a mixed logit model can be expressed in the form:

where is the logit probability evaluated at parameter :

andis a density function. is deterministic term observed by the analyst, which depends on the parameters . Usually, the utility is linear in , thus . The mixed logit probability then takes its usual form:

The application of mixed logit (ML) choice model in the revenue management problem allows for the estimation of continuous taste heterogeneity. Given that the choice probability of mixed logit is specified as follows:

The corresponding optimization problem can be expressed as:

The optimization is subjected to the same set of constraints in the MNL optimization.

**5.4.1 Random coefficient with parametric distribution**

To incorporate the random coefficient parameters, obtained from ML model estimation, into revenue optimization, we discretize the price sensitivity which is assumed to be log-normally distributed into six segments. It should be noted that the estimation procedure transforms the lognormal into a normal by applying the transformation where . Given that is distributed as a standard normal, we can derive the probability that lies within each of the six regions in Figure 3. The averaged price sensitivity of each segment can then be written as where is the mean value of within region calculated from the probability density function of the standard normal distribution. The averaged price sensitivity obtained is further truncated at [-1.0, 0.5] to exclude extreme behaviors.

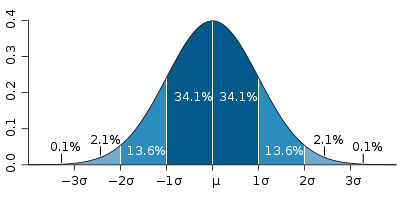


Figure 3 Standard normal distribution

**5.4.2 Random coefficient with non-parametric distribution**

The same procedure described in Section 5.5.1 is applied to parameters estimated using non-parametric B-spline distribution. The control points necessary to construct B-spline curves are derived using the De Boor’s algorithm (Lee, 1982) for a polynomial of degree three. Examples of the curves estimated are shown in Figure 4.

Distributions are truncated at [-1.0, 0.5] to exclude extreme behaviors which are difficult to interpret. To incorporate random coefficient into revenue optimization, we discretize the spline curve into seven regions equally spaced on the axis. For each region , the averaged price sensitivity is obtained by randomly drawing from the distribution based on the shape of the spline curve within the region considered. The probability mass for each averaged price sensitivity is equal to 1/7 since by constriction the spline region is equally partitioned on the axis. This approach allows us to approximate the random coefficient parameters estimated from mixed logit by segmenting price sensitivity into seven segments, and to assign each value to the population with the probability of 1/7 each.



Figure 4 Example of Spline curve estimated with De Boor’s algorithm

**6. Optimization results**

The optimization problems, as specified in Section 5, are solved with nonlinear programming techniques using LINGO 12.0, the optimization software by Lindo System Inc. (Lindo System Inc, 2010). The nonlinearity is caused by the exponential function of the logit choice probability. Results obtained are summarized in Table 4 to Table 6.

Table 4 provides acceptance ratio across the four departure times considered, which results into the accepted passenger rates shown in Table 5. The strategy deriving from the optimization procedure suggests increasing the total accepted passengers in all of the four trains. Based on our assumption that the optimal solution cannot utilize more seats than the one occupied in the current conditions we can derive that more short-haul passengers should be accepted. By allowing more short duration trips it is possible to serve a greater number of passengers and to produce more revenues for the operator.

Table 4 Acceptance ratio comparison

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **(O,D)** | **5AM (Train#1)** | | | |  | **9AM (Train#2)** | | | |  | **1PM (Train#3)** | | | |  | **4PM (Train#4)** | | | |
|  | **MNL** | **LC** | **ML** | **Spline** |  | **MNL** | **LC** | **ML** | **Spline** |  | **MNL** | **LC** | **ML** | **Spline** |  | **MNL** | **LC** | **ML** | **Spline** |
| (5,4) | 0.544 | 0.559 | 1 | 0.560 |  | 1 | 1 | 1 | 0.958 |  | 1 | 1 | 1 | 1 |  | 0.926 | 0.999 | 0.865 | 1 |
| (5,3) | 1 | 1 | 1 | 1 |  | 1 | 1 | 1 | 1 |  | 1 | 0 | 1 | 1 |  | 1 | 1 | 1 | 0.587 |
| (5,2) | 0 | 0 | 0 | 0 |  | 0 | 0 | 0 | 0 |  | 0 | 0 | 0 | 0 |  | 0 | 0 | 0 | 0 |
| (5,1) | 0 | 0 | 0 | 0 |  | 0 | 0 | 0 | 0 |  | 0 | 0 | 0 | 0 |  | 0 | 0 | 0 | 0 |
| (4,3) | 1 | 1 | 0.745 | 1 |  | 1 | 1 | 0.271 | 1 |  | 1 | 0 | 1 | 1 |  | 1 | 1 | 1 | 0 |
| (4,2) | 0.381 | 0.452 | 0.470 | 0.455 |  | 0.324 | 0.328 | 0.375 | 0.334 |  | 0.319 | 0.332 | 0.323 | 0.339 |  | 0.321 | 0.343 | 0.324 | 0.343 |
| (4,1) | 0.295 | 0.261 | 0.390 | 0.262 |  | 0.197 | 0.197 | 0.347 | 0.200 |  | 0.197 | 0.208 | 0.196 | 0.196 |  | 0.268 | 0.262 | 0.267 | 0.294 |
| (3,2) | 1 | 0 | 1 | 0 |  | 0 | 0 | 1 | 0 |  | 1 | 0 | 1 | 0 |  | 1 | 0 | 1 | 1 |
| (3,1) | 0.338 | 0.521 | 1 | 0.517 |  | 0.608 | 0.596 | 1 | 0.702 |  | 0.656 | 0 | 0.666 | 0.900 |  | 0.403 | 0.613 | 0.426 | 0.056 |
| (2,1) | 1 | 1 | 1 | 1 |  | 1 | 1 | 1 | 1 |  | 1 | 1 | 1 | 1 |  | 1 | 1 | 1 | 1 |

Table 5 Accepted demand comparison

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **(O,D)** | **5AM (Train#1)** | | | | |  | **9AM (Train#2)** | | | | |  | **1PM (Train#3)** | | | | |  | **4PM (Train#4)** | | | | |
|  | **Exist** | **MNL** | **LC** | **ML** | **Spline** |  | **Exist** | **MNL** | **LC** | **ML** | **Spline** |  | **Exist** | **MNL** | **LC** | **ML** | **Spline** |  | **Exist** | **MNL** | **LC** | **ML** | **Spline** |
| (5,4) | 1 | 5 | 5 | 3 | 5 |  | 0 | 6 | 6 | 3 | 5 |  | 0 | 1 | 1 | 1 | 1 |  | 0 | 3 | 3 | 3 | 3 |
| (5,3) | 0 | 6 | 6 | 3 | 6 |  | 0 | 4 | 4 | 2 | 5 |  | 0 | 1 | 0 | 1 | 1 |  | 0 | 2 | 2 | 2 | 2 |
| (5,2) | 1 | 0 | 0 | 0 | 0 |  | 1 | 0 | 0 | 0 | 0 |  | 0 | 0 | 0 | 0 | 0 |  | 1 | 0 | 0 | 0 | 0 |
| (5,1) | 9 | 0 | 0 | 0 | 0 |  | 9 | 0 | 0 | 0 | 0 |  | 4 | 0 | 0 | 0 | 0 |  | 4 | 0 | 0 | 0 | 0 |
| (4,3) | 1 | 4 | 4 | 2 | 4 |  | 0 | 4 | 4 | 1 | 4 |  | 0 | 5 | 0 | 5 | 5 |  | 1 | 3 | 3 | 3 | 0 |
| (4,2) | 24 | 27 | 32 | 28 | 32 |  | 21 | 30 | 31 | 25 | 31 |  | 44 | 46 | 48 | 46 | 48 |  | 27 | 27 | 28 | 27 | 27 |
| (4,1) | 45 | 43 | 38 | 47 | 38 |  | 49 | 41 | 41 | 52 | 41 |  | 81 | 77 | 81 | 77 | 75 |  | 59 | 60 | 59 | 60 | 63 |
| (3,2) | 4 | 5 | 0 | 1 | 0 |  | 3 | 0 | 0 | 1 | 0 |  | 0 | 1 | 0 | 1 | 0 |  | 0 | 2 | 0 | 2 | 1 |
| (3,1) | 1 | 9 | 14 | 7 | 14 |  | 3 | 14 | 14 | 8 | 15 |  | 0 | 5 | 0 | 5 | 6 |  | 1 | 4 | 5 | 4 | 0 |
| (2,1) | 2 | 5 | 5 | 2 | 5 |  | 1 | 6 | 7 | 2 | 7 |  | 2 | 5 | 6 | 6 | 6 |  | 2 | 2 | 2 | 2 | 2 |
| **Total** | **88** | **104** | **104** | **94** | **104** |  | **87** | **106** | **106** | **95** | **107** |  | **131** | **141** | **136** | **142** | **142** |  | **95** | **102** | **102** | **102** | **99** |

Table 6 Revenue comparison

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **(O,D)** | **5AM (Train#1)** | | | | |  | **9AM (Train#2)** | | | | |  | **1PM (Train#3)** | | | | |  | **4PM (Train#4)** | | | | |
|  | **Exist** | **MNL** | **LC** | **ML** | **Spline** |  | **Exist** | **MNL** | **LC** | **ML** | **Spline** |  | **Exist** | **MNL** | **LC** | **ML** | **Spline** |  | **Exist** | **MNL** | **LC** | **ML** | **Spline** |
| (5,4) | 32 | 191 | 197 | 138 | 202 |  | 0 | 255 | 254 | 153 | 204 |  | 0 | 60 | 59 | 62 | 47 |  | 0 | 156 | 170 | 145 | 103 |
| (5,3) | 0 | 488 | 480 | 193 | 474 |  | 0 | 343 | 325 | 211 | 282 |  | 0 | 79 | 0 | 85 | 45 |  | 0 | 228 | 205 | 249 | 87 |
| (5,2) | 93 | 0 | 0 | 0 | 0 |  | 124 | 0 | 0 | 0 | 0 |  | 0 | 0 | 0 | 0 | 0 |  | 140 | 0 | 0 | 0 | 0 |
| (5,1) | 795 | 0 | 0 | 0 | 0 |  | 978 | 0 | 0 | 0 | 0 |  | 520 | 0 | 0 | 0 | 0 |  | 523 | 0 | 0 | 0 | 0 |
| (4,3) | 71 | 220 | 221 | 136 | 221 |  | 0 | 289 | 288 | 56 | 201 |  | 0 | 428 | 0 | 432 | 308 |  | 71 | 246 | 244 | 248 | 0 |
| (4,2) | 2,366 | 2,703 | 3,206 | 2,771 | 3,214 |  | 2,263 | 3,373 | 3,416 | 2,758 | 3,411 |  | 5,293 | 6,182 | 6,426 | 6,228 | 6,409 |  | 3,061 | 3,575 | 3,816 | 3,590 | 3,645 |
| (4,1) | 4,781 | 4,413 | 3,905 | 4,848 | 3,910 |  | 5,430 | 4,757 | 4,743 | 5,934 | 4,668 |  | 10,629 | 10,739 | 11,304 | 10,614 | 10,479 |  | 7,117 | 8,388 | 8,175 | 8,296 | 8,805 |
| (3,2) | 396 | 481 | 0 | 132 | 0 |  | 370 | 0 | 0 | 154 | 0 |  | 0 | 161 | 0 | 161 | 0 |  | 0 | 203 | 0 | 202 | 96 |
| (3,1) | 104 | 882 | 1,360 | 719 | 1,347 |  | 343 | 1,573 | 1,543 | 859 | 1,416 |  | 0 | 620 | 0 | 627 | 776 |  | 119 | 480 | 641 | 506 | 28 |
| (2,1) | 70 | 181 | 185 | 83 | 190 |  | 25 | 248 | 268 | 93 | 266 |  | 90 | 245 | 278 | 274 | 268 |  | 75 | 100 | 51 | 112 | 47 |
| **Total** | **8,708** | **9,560** | **9,553** | **9,022** | **9,558** |  | **9,533** | **10,837** | **10,837** | **10,219** | **10,449** |  | **16,532** | **18,514** | **18,068** | **18,482** | **18,333** |  | **11,106** | **13,376** | **13,303** | **13,348** | **12,812** |
| **% Improve** | | **9.78** | **8.85** | **3.29** | **9.43** |  |  | **13.68** | **13.68** | **7.19** | **9.61** |  |  | **11.99** | **9.29** | **11.79** | **10.89** |  |  | **20.44** | **19.78** | **20.19** | **15.36** |

Table 6 compares revenue from each train trip and for the five stations accounted in the optimization problem. Result shows that the revenue improvement ranges from 3.29% to 20.44% depending on the departure time and the choice model type. Train#4 shows the highest revenue improvement when compared to other trains. The non parametric choice model results into revenue improvement ranging from 9.43% to 15.36%, these improvements are comparable to those obtained with other models for Train#1 and Train#3. B-spline (which is the model with the highest goodness of fit) revenue improvement for Train#2 and Tran#4 are relatively lower than those obtained with other model specifications. This could be explained by the variability to price sensitivity imposed by the spline model which better represents customers’ heterogeneity. We now turn to a more detailed analysis for the market (4,1), which is the one with the highest traffic in our selected network. For this market, we consider four different departure times which load passengers into Train#1 to Train#4, corresponding to the departure times from station 4 at 8AM, 12PM, 4PM, and 7 PM respectively.

Fare calculated for market (4,1) are shown in Figure 6, 8, 10, and 12 for the 4 train trips departing from Station4 at 8AM, 12PM, 4PM, and 7PM respectively. The results from all the models are compared to the existing fare, which is the representative fare pattern for a particular departure time of day and day of week, as obtained from the data. In general, while the MNL, LC, and ML with parametric distribution provide similar results across the four departure times, the optimization with B-spline choice model provides distinct fare pattern which is consistent with our expectation that price sensitivity should decrease as time gets closer to the departure date.

Figure 7, 9, 11, and 13 show the corresponding number of accepted passengers in market (4,1) on each day over the sale horizon across four departure times. The response of passengers demand to the choice model is consistent with our general expectations; a greater number of passengers are expected when the new fare is lower than the existing; and a lower number of passengers in the opposite case. This is behavioral pattern is observed across all the four train departure times.

Figure 5 Market (4,1) 8 AM departure: Fare

Figure 6 Market (4,1) 8 AM departure: Accepted number of passenger

Figure 7 Market (4,1) 12 PM departure: Fare

Figure 8 Market (4,1) 12 PM departure: Accepted number of passengers

Figure 9 Market (4,1) 4 PM departure: Fare

Figure 10 Market (4,1) 4 PM departure: Accepted number of passengers

Figure 11 Market (4,1) 7 PM departure: Fare

Figure 12 Market (4,1) 7 PM departure: Accepted number of passenger

**7. Conclusions and future research directions**

Accounting for customer heterogeneity in choice models for revenue management has emerged as an important topic in the operation research literature for its implications in terms of ticket pricing and seat allocation strategy. In this paper, we have introduced advanced random utility models in the form of discrete and continuous mixture of logit models and made them operational using internet booking data. Beside multinomial logit, latent classes and parametric mixed logit, we have also proposed semi-parametric methods based on B-splines curves, which are polynomial approximations of the unknown distributions. The approach allows the model to explicitly reveal taste heterogeneity in ticket purchase timing decisions without assuming a pre-defined shape of the underlying distribution. The estimates obtained from behavioral models have been incorporated into an optimization problem that maximizes railway revenue from ticket sales. The proposed approach jointly optimizes fare and seat allocation for each origin destination pair in the network considered. The method is able to derive fare policy in response to realistic passenger behavior and at the same time allows efficient utilization of the supply resources across the network.

The solution illustrates the impact of the strategies deriving from the optimization procedure on pricing, capacity distribution, and revenue. Results show that random coefficient parameters from mixed logit can be incorporated into the optimization problem and provide revenue improvement relatively similar to choice model based on discrete segmentation and homogeneous population. RM optimization with B-spline choice model provides pricing strategies which are consistent with the expected passenger purchase timing behavior, as price sensitivity is expected to decrease when departure date is approaching. The optimization results also indicate that seat allocation policy which accepts more short-haul trip contributes to greater revenue than long-haul trip under the same seat capacity. The solution from the proposed framework results in significant revenue improvements that range from 3.29% to 20.44% depending on the choice model and on the train departure time considered.

This research responds to the need to find parameter estimation techniques for discrete and continuous mixture of logit models in RM and opens up a number of directions for future work. First, it would be interesting to investigate other choice dimensions of the railway RM problem, for instance, choice of departure time or departure day when making ticket purchase decisions. However, for these choice dimensions, additional data collection to construct plausible choice sets for each customer might be necessary. It will also be interesting to optimize ticket revenue over multiple departures simultaneously by accounting for demand shifts across different departure time or departure day options, thereby efficiently balancing passenger demand and improve total revenue. From the RM revenue optimization perspective, it is desirable to consider the network with hub and spoke characteristic which involves station transfers and more complex capacity constraints. Finally, the aggregate demand functions can be substituted with sample path generators that produce streams of customer demand using random variables.

**ACKNOWLEDGMENTS**

We thank the editor and the anonymous reviewers for their constructive comments, which helped us to improve the manuscript.

**References**

Bastin F., Cirillo C., and Ph.L. Toint. 2010. Estimating non-parametric random utility models, with an application to the value of time in heterogeneous populations. *Transportation Science* 44(4), pp. 537-549.

Bastin, F. Another Mixed Logit Estimation Tool (AMLET) Reference Manual. <http://amlet.slashbin.net/> Accessed July 8, 2011.

Ben-Akiva, M., and S. Lerman. 1994. Discrete Choice Analysis: Theory and Applications to Travel Demand, 6th ed. MIT Press, Cambridge, MA.

Bhat, C.R. and S. Castelar. 2002. A unified mixed logit framework for modeling revealed and stated preferences: Formulation and application to congestion pricing analysis in the San Francisco Bay area. Transportation Research Part B 36 (7), 593–616.

Brownstone, D., Bunch, D., and K. Train, 2000. Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles. Transportation Research Part B 34, 315–338.

Carrier, E., 2003. Modeling airline passenger choice: Passenger preference for schedule in the passenger origin–destination simulator. (PODS). MS. thesis. Massachusetts Institute of Technology.

Carrier, E. 2008. *Modeling the Choice of an Airline Itinerary and Fare Product using Booking and Seat Availability Data*. Ph.D. Dissertation, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA.

Chaneton J., and G. Vulcano. 2011. Computing bid-prices for revenue management under customer choice behavior. *M&SOM* Vol. 13(4): 452-470.

Cherchi, E. and J.de.D. Ortúzar. 2003 [Alternative specific variables in non-linear Utility Functions: Influence of correlation, homoschedasticity and taste variations](http://unica2.unica.it/crimm/it/documenti/Papers/cherchi_papers/2003_Cherchi-Ortuzar_Alternative%20Specific%20Variables.pdf). *10th International Conference on Travel Behaviour Research*. Lucerne, Switzerland, 2003.

Davis, J. M., G. Gallego, and H. Topaloglu. 2011. Assortment optimization under variants of the nested logit model. Working paper, Cornell University.

Farin, G. E. 1991. NURBS for Curve and Surface Design. SIAM, Philadelphia.

Fosgerau, M., 2006. Investigating the distribution of the value of travel time savings. Transportation Research Part B 40 (8), 688–707.

Fosgerau, M., M. Bierlaire. 2007. A practical test for the choice of mixing distribution in discrete choice models. *Transportation* *Res. Part B* 41(7) 784–794.

Gallego, G., and H. Topaloglu. 2012. Constrained assortment optimization for the nested logit model. Working paper, Cornell University.

Gallego, G., G. Iyengar, R. Phillips, and A. Dubey. 2004. Managing exible products on a network. Working Paper, Columbia University.

Greene, W.H., Hensher, D.A., Rose, J., 2006. Accounting for heterogeneity in the variance of unobserved effects in mixed logit models. Transportation Research Part B 40 (1), 75–92.

Hensher, D. A. and W. H. Green. 2003 The Mixed Logit Model: The State of Practice, Transportation, 30, 133–176.

Hess, S., M. Bierlaire and J. W. Polak. 2005 Estimation of value of travel-time savings using mixed logit models, Transportation Research A, 39, 221–236.

Hess, S., and J. Polak. 2005. Mixed logit modelling of airport choice in multi-airport regions. Journal of Air Transport Management 11 (2), 59–68.

Hess, S., Train, K., and J. Polak. 2006. On the use of modified latin hypercube sampling (MLHS) method in the estimation of mixed logit model for vehicle choice. Transportation Research Part B 40 (2), 147–163.

Hetrakul, P. and C. Cirillo. 2013 “Accommodating taste heterogeneity in railway passenger choice models based on internet booking data.” *Journal of Choice Modeling*”. 6, pp.1-16.

Jin, X., 2007. Toward Time-of-Day Modeling for Long Distance Trips, Doctoral Dissertation, University of Wisconsin, Milwaukee.

Kunnumkal, S., and H. Topaloglu. 2008. A re\_ned deterministic linear program for the network revenue management problem with customer choice behavior. Naval Research Logistics 55 (6): 563{580.

Lee, E., 1982. A simplified B-spline computation routine, Computing [Volume 29, Number 4](http://www.springerlink.com/content/0010-485x/29/4/) (1982), 365-371.

Li, G., and P. Rusmevichientong. 2012. Technical Note: A Simple Greedy Algorithm for Assortment Optimization in the Two-Level Nested Logit Model. Working Paper, USC.

## Thao Ly (2012) [Railway Industry in France](http://blogs.cornell.edu/advancedrevenuemanagement12/2012/03/26/railway-industry-in-france/). Revenue Management in Action. http://blogs.cornell.edu/advancedrevenuemanagement12/2012/03/26/railway-industry-in-france/

Meissner, J., A. Strauss, and K. Talluri. 2012. An Enhanced Concave Program Relaxation for Choice Network Revenue Management. Production and Operations Management 22 (1): 71-87.

Newman, J., Ferguson, M., and Garrow, L. 2012 Estimating discrete choice models with incomplete data. Transportation Research Record 2302: 130-137.

Rusmevichientong, P., D. Shmoys, C. Tong, and H. Topaloglu. 2013. Assortment Optimization under the Multinomial Logit Model with Random Choice Parameters. Working paper, Cornell University.

Sibdari, S., Lin, K. Y., and S. Chellappan. 2008 Multiproduct revenue management: An empirical study of Auto train at Amtrak. Journal of Revenue and Pricing Management Vol 7(2), pp.172-184.

Talluri, K., G. J. van Ryzin. 2004a. Revenue management under a general discrete choice model of consumer behavior. Management Sci. 50(1) 15–33.

Talluri, K., G. J. van Ryzin. 2004b. The Theory and Practice of Revenue Management. Kluwer Academic, New York.

Train, K. 2009. Discrete choice methods with simulation, New York, NY, USA, Cambridge University Press.

Vermunt J.K., and J. Magdison. 2005 Technical Guide for Latent Gold Choice 4.0: Basic and Advanced. Statistical Innovations Inc., Belmont, MA.

Vossen, T. W. M., and D. Zhang. 2012. A Dynamic Disaggregation Approach to Approximate Linear Programs for Network Revenue Management. Technical report, University of Colorado at Boulder, Boulder, CO.

Vulcano, G.J. van Ryzin, and W. Chaar. Choice-based revenue management: An study of estimation and optimization. *M&SOM* Vol. 12(3): 371-392, 2010.