Mixed Logit and Latent Class Model for Railway Revenue Management
MIXED LOGIT AND LATENT CLASS MODEL FOR RAILWAY REVENUE MANAGEMENT

FINAL REPORT

UMD-2010-02
DTRT07-G-0003

Prepared for
U.S. Department of Transportation
Research and Innovation Technology Administration

By
Dr. Cinzia Cirillo, Principal Investigator
Pratt Hettrakul, Graduate Research Assistant

Department of Civil & Environmental Engineering
1173 Glenn L. Martin Hall, Bldg #088
University of Maryland, College Park, MD 20742

November 2011

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the University of Maryland.
In this paper advanced demand modeling approaches are proposed to study intercity passenger booking decision and to segment preferences; all the models are calibrated on internet booking data. Modeling formulations considered include multinomial logit, mixed logit, and latent class models; markets are segmented on trip distances: long, medium, and short. The results show that the following variables: fare price, advance booking (number of day before departure), and departure day of week, can be used as determinants affecting ticket booking. Mixed logit and latent class models are then applied to account for taste heterogeneity. The results indicate that mixed logit model provides the best statistical fit for the long distance and medium distance markets, while the latent class model provides the best statistical fit for the short distance market. Results also indicate that segmenting passengers by booking period provides better fit than segmenting passengers by socioeconomic information.

Results from this study prove that advanced demand models can be estimated on internet booking data and that market segmentation can be obtained even with limited knowledge of socio-demographic characteristics of the population. A number of directions could be considered for future research. First, the proposed choice model could be integrated into an optimization algorithm for maximizing revenues. Second, the proposed model is based on a static approach where decisions are made at a single point in time. Dynamic discrete choice models are expected to provide a significant improvement in prediction accuracy by offering the possibility to account for the evolving characteristics of the market over time and for sequential purchasing decisions. Third, the modeling approach applied here to railway revenue management could be applied to other problems for which just internet booking data are available (i.e. shippers, or couriers selling their services online).
# TABLE OF CONTENTS

1. INTRODUCTION ...................................................................................................................... 1

2. LITERATURE REVIEW ........................................................................................................... 2

3. METHODOLOGIES .................................................................................................................. 4
   3.1 Latent Class (LC) Model ....................................................................................................... 4
   3.2 Mixed Logit (ML) Model ...................................................................................................... 6

4. DATA ANALYSIS ..................................................................................................................... 7

5. PASSENGER CHOICE MODEL .............................................................................................. 8
   5.1 Model Setup .......................................................................................................................... 8
   5.2 Sample Selection ................................................................................................................... 9
   5.3 Choice Generation ............................................................................................................... 9

6. MODEL FORMULATION ...................................................................................................... 10
   6.1 Multinomial Logit (MNL) model specification .................................................................. 10
   6.2 Mixed Logit (ML) model specification ............................................................................... 10
   6.3 Latent Class (LC) model specification ................................................................................ 10
   6.4 Multinomial Logit (MNL) model with socioeconomic information specification ............. 11

7. RESULTS AND DISCUSSIONS ............................................................................................. 11
   7.1 Results of multinomial logit model ..................................................................................... 15
   7.2 Results of mixed logit model .............................................................................................. 15
   7.3 Results of latent class model ............................................................................................. 15
   7.4 Results of multinomial logit model with socioeconomic information ................................ 18
   7.5 Comparison of Model Fit .................................................................................................... 20

8. CONCLUSIONS ....................................................................................................................... 20
LIST OF FIGURES

Figure 1 Total passenger demand by trip schedule 7
Figure 2 Total number of booked seats prior to departure 8
Figure 3 Probability of passenger belonging to class 1 in long distance model 17
Figure 4 Probability of passenger belonging to class 1 in medium distance model 18
Figure 5 Probability of passenger belonging to class 1 in short distance model 18
LIST OF TABLES

Table 1 Long distance model 12
Table 2 Medium distance model 13
Table 3 Short distance model 14
1. INTRODUCTION
Discrete choice analysis has been introduced in revenue management (RM) in recent years due to its ability to account for customer preferences in RM problems. Despite the number of research efforts toward discrete choice analysis in airline, and hotel RM, there is limited empirical research on the application of discrete choice model to the railway industry. Recently, the increased use of internet as a major distribution channel for railway tickets has resulted in railway passengers becoming more strategic to fare policies. Consequently, railway has become more aware of the benefit that can be derived from the development and the deployment of RM; AmTrak, U.S., VIA Rail Canada, and Eurostar and SNCF in France are known to actively use RM techniques.

Specifically, most of the empirical studies in railway RM reported so far do not account for taste heterogeneity across passengers. Passenger taste heterogeneity is one of major characteristics of the railway market. Railway passenger preferences vary by distance (short haul or long haul) and by time of day. Given that RM relies on the premise that different customers are willing to pay different amounts for a product, accounting for passenger heterogeneity is expected to provide high yield toward RM strategy. Garrow (1) suggests that calibrating models by segments to distinguish between time-sensitive and price-sensitive customers is potentially a relevant research topic in RM.

Generally speaking, consumer taste heterogeneity is an important issue in web based marketing; existing studies indicate that customer segmentation is crucial for electronic commerce (2). Burke (3) raised the issue of consumer market heterogeneity in the context of internet marketing. Peterson et al. (4) dealt with the problem concerning retailers’ segmentation for online shopping. Miller (5) has proposed the use of demographics to depict the profile of internet users.

This study aims to investigate the possibility to model taste heterogeneity from internet tickets sale in order to support the development of efficient railway RM strategies. To this scope we review, in Section 2, the literature related to the class of discrete choice models able to recover taste heterogeneity. In Section 3 methodologies used for this analysis are reported. Section 4 describes the data set and provides statistical analysis on the sale data. Section 5, 6 and 7 are
dedicated respectively to the choice set generation, model specification and model results. Conclusions and suggestions for future research are given in Section 8.

2. LITERATURE REVIEW

The methods that deal with behavioral heterogeneity generally relies on market segmentation. Originally introduced by Smith (6), market segmentation is the process of splitting a market into distinctive groups or segments of customers who have similar characteristics or needs. It offers valuable implication for a company’s market strategy and resource allocation. A number of empirical studies accounting for heterogeneity in intercity travel behavior can been found in the literature. Advanced choice model such as mixed logit model have been broadly applied in air travel studies to account for preference heterogeneity and flexible substitution pattern across alternatives (7, 8). Greene and Hensher (9) compared latent class (LC) with the mixed logit (ML) model using a data set on long distance travel by three road types in New Zealand in 2000. After a detailed comparison of model goodness of fit and other parameters including value of travel time savings, direct share elasticities, choice sensitivities to travel time increase, and choice probabilities, the study does not report a definite conclusion on the superiority of one of the models proposed. Future research is needed to compare and contrast these models in terms of model stability in explanation and prediction.

Bhat (10) estimated an intercity travel mode choice model which accommodates variations in response to level of service measures due to observed and unobserved individual characteristics. The analysis applied multinomial logit (MNL), fixed-coefficients logit (FCL), and random coefficients logit (RCL) model to examine the travel mode choice behavior of weekday business travelers. Data used for the analysis have been extracted from the VIA Rail (the Canadian national rail carrier) data set collected in 1989 and are relative to the Toronto-Montreal corridor. The model was applied to forecast future inter-city travel demand on the corridor and estimate modal shift in response to potential rail service improvement. Based on a nested likelihood ratio test, the RCL rejects the FCL, and the FCL rejects the MNL model when applying a non-nested adjusted likelihood ratio test. Bhat (11) formulated a mixed logit model of multiday urban travel mode choice which accommodates variations in mode preference and response to level of service factors. The model was applied to examine the travel mode choice for workers in the San
Francisco Bay Area in California. These two studies (10, 11) emphasize the need to accommodate observed and unobserved heterogeneity across individuals in travel mode choice modeling.

Shen (12) compared the difference between latent class and mixed logit models using two stated choice survey data sets from Osaka, Japan relative to mode choice. The first survey aims to investigate environmental consciousness of transport mode choice consisting of monorail, car, and bus. The second survey again on mode choice focuses on both environmental impact and network accessibility. The latent class and mixed logit models are compared in term of values of time savings, direct choice elasticities, predicted choice probabilities, and prediction success indices. In addition, the prediction success indices proposed by McFadden (13) and a non-nested model test based on the Akaike Information Criterion (AIC) proposed by Ben-Akiva and Swait (14) were used for model comparison. Latent class model is found to be statistically superior to mixed logit model based on prediction success indices and nonnested model test in both data sets.

Teichert et al. (15) applied the latent class modeling approach to explore preferences within airlines segments and subsequently analyzed respondents’ profiles in terms of individual socioeconomic and trip characteristics. The segmentation approach is proposed where behavioral and socio-demographic variables are used to profile segments. Passengers are segmented into five latent classes based on preference toward the product attributes and additional attitudinal and socio-demographic variables which are: efficiency/punctuality, comfort, price/performance, and catch all/flexibility (broadly balanced across product features). The authors suggest a route specific optimization for each product features, given that the demand structure may be geographically different. They also suggest the extension of the analysis in the context of a more general problem.

Carrier (16) analyzed the choice of airline itinerary and fare product based on the latent class model framework. The choice set was constituted from booking data, fare rules, and seat availability data. The segmentation applied is based on passenger and trip characteristics including variables such as frequent flyer membership, ticket distribution channel, and travel day of week. The approach is shown to provide a distinct and intuitive segmentation across
passengers. Two passenger segments were defined in the study: business and leisure passenger. Based on this result, business travelers tend to travel on weekdays, book the ticket with traditional offline travel agents, and rely on the frequent flyer membership. The model extends the application of passenger choice model to airline pricing and revenue management.

Wen and Lai (17) used latent class model to identify airline passengers’ potential segments and preferences toward international air carriers. The analysis is based on stated preference data relative to airline choices for flights from Taipei to Tokyo and Taipei to Hong Kong. Passengers are segmented by individual socioeconomic and trip characteristics. The analysis compared three model types, multinomial logit, latent class model without individual characteristics, and latent class model with individual characteristics. The MNL model shows that airfare, schedule time difference, flight frequency, on-time performance, check-in service, in flight seat space, and cabin crew service influence airline carrier choices. The latent class model provides additional insights on choice probabilities sensitivity to changes in service attributes (i.e., flight frequency, schedule time difference, and on-time performance) and for distinct segments. This study found that willingness to pay for service attributes improvement is substantially different across air routes and vary by travelers’ segments. It is suggested that future studies should recognize the diverse characteristics of air routes and distinct segments of passengers especially for international air carrier choices.

3. METHODOLOGIES
In this section, we summarize methodologies used in passenger modeling which accounts for passenger heterogeneity across booking.

3.1 Latent Class (LC) Model
Latent class model is based on discrete segmentation which assumes that heterogeneity in passenger behavior is likely to be driven by specific elements. The approach generally groups observations into classes with similar needs, constraints, and preferences with class membership model. For instance, passengers can be segmented to business-oriented and leisure-oriented travelers with different preferences parameters defined. The class membership model is
combined with choice model enabling the model to account for differences in choice behavior between different segments of the market (16).

The structure of the latent class passenger choice model could be described as follows. Let \( i \) represents alternative from \( 1, \ldots, J_b \) in the choice set \( C \) of booking \( b \). The model form can be written as:

\[
P(i / X_M, X_C) = \sum_{s=1}^{S} P(s / X_M) P(i / X_C, s) \quad \forall i \in C
\]

Where \( s \) is class index; \( \{1, 2, \ldots, S\} \)

\( X_M \) is class membership explanatory variable

\( X_C \) is class specific choice model explanatory variable

The utility function of alternative \( i \) given the customer is in the class \( s \) can be written as:

\[
U_{ib} = X_{Cib} \beta_C + \epsilon_{ib}
\]

Where \( X_{Cib} \) is a \((I \times K)\) vector of choice model explanatory variables

\( \beta_C \) is a \((K \times 1)\) vector of parameters

\( \epsilon_{ib} \) is a random disturbance (i.i.d. extreme value)

The class specific choice probability of travel option \( i \) can be expressed as:

\[
P(i / X_{Cib}, s) = \frac{e^{X_{Cib, \beta_{C,s}}}}{\sum_{j=1}^{J} e^{X_{Cib, \beta_{C,s}}}} \quad \forall s \in S \quad \forall i \in C
\]

Where \( \beta_{C,s} \) are the unknown parameters of the class-specific choice model. The probability of belonging to the latent class \( s \) can be written as:
\[ P(s / X_{Mb}) = \frac{e^{X_{Mb} \beta_M}}{\sum_{t=1}^{S} e^{X_{Mb_t} \beta_M}} \]  \hfill (4)

Where \( \beta_M \) are the unknown parameters for class membership model.

### 3.2 Mixed Logit (ML) Model

Mixed logit is a highly flexible model capable of approximating any random utility model \( (I^8) \). It obviates three limitations of standard logit model by allowing for random taste variation, unrestricted substitution patterns, and correlation in unobserved factors over time. Mixed logit probabilities are the integral of standard logit probabilities over a density of parameters \( (\beta) \). Choice probabilities of a mixed logit model can be expressed in the form:

\[ P_m = \int L_n(\beta) f(\beta) d\beta \]  \hfill (5)

Where \( L_n(\beta) \) is the logit probability evaluated at parameter \( \beta \):

\[ L_n(\beta) = \frac{e^{V_{ni}(\beta)}}{\sum_{j=1}^{J} e^{V_{nj}(\beta)}} \]  \hfill (6)

And \( f(\beta) \) is a density function. \( V_{ni}(\beta) \) is deterministic term observed by the analyst, which depends on the parameters \( \beta \). Usually, the utility is linear in \( \beta \), thus \( V_{ni}(\beta) = \beta x_{ni} \). The mixed logit probability then takes its usual form:

\[ P_m = \left\{ \frac{e^{\beta x_{ni}}}{\sum_{j} e^{\beta x_{nj}}} \right\} f(\beta) d(\beta) \]  \hfill (7)

Mixed logit model can be viewed as a mixture of the logit function evaluated at different \( \beta' \)s with \( f(\beta) \) as the mixing distribution.
4. DATA ANALYSIS

Booking data for intercity passenger railway trips are used for this study. Data collected over a period of two months in 2009, contain information in terms of trip origin, trip destination, fare class, reservation date, departure date, departure time, arrival time, fare price, and accommodation charge. Booking data from the first month are used for the analysis; it contains a total of 406,422 reservation records.

This railway service under study consists of two ticket classes: first class, and business class. This study focuses on business class passenger which is the predominant market for this service and on reservations which are eventually confirmed and paid. To reduce the problem size, the analysis focuses only on the northbound trip. This results in a final data set of 110,828 reservation records. General statistics on trip schedule and advance booking are reported in Figure 1 and Figure 2.

![Figure 1 Total passenger demand by trip schedule](image)

**Figure 1 Total passenger demand by trip schedule**

Figure 1 shows a high passenger demand on weekday and a significant lower number of passengers on weekend, with the lowest values on Saturday. Passengers consistently prefer PM off peak (12:00-3:59 PM.) and PM peak (4:00-6:29 PM.) times of day.
Figure 2 represents the number of reservation by number of days before departure. Data indicate that about 98 percent of the passengers make the reservation no earlier than 30 days before the departure date. The majority of the passengers book the ticket about one week before departure and that an extremely high portion of passengers book the ticket within 2 days before departure.

5. PASSENGER CHOICE MODEL

5.1 Model Setup

In this Section, we propose disaggregate choice models for booking time decisions, based on the assumption that each individual purchases the ticket at the time that maximizes his/her utility. While some study suggests a nested structure model where the mode choice decision is made at the upper nest, and the ticket product decision is made at the lower nest (19), this approach generally requires dedicated stated preference survey which combines mode choice and ticket choice decision. Given that this study is based on confirmed booking data, we assume that the mode choice decision has already been made by the passengers. Thus, the proposed model aims to capture the passengers’ purchase timing decisions as a function of booking time specific and trip specific attributes.
5.2 Sample Selection

The data set of 110,828 observations mentioned in Section 4 is used for the analysis. It contains business class passengers who eventually confirmed and paid for their trip. The railway service has 16 stations, which have been aggregated into 4 groups based on a geographical order (from North end to South end stations). Due to space limitation, it has not been possible to consider all the station groups defined above. Instead, we select three representative markets, each characterized by trip distance as follows:

1. Long Distance: trip departing from station group 4 to station group 2
2. Medium Distance: trip departing from station group 2 to station group 1
3. Short Distance: trip traveling within station group 4

5.3 Choice Generation

The fare price of this railway service varies depending on departure date, departure day of week, departure time, the time the reservation is made, and customer demand for each departure. Different passenger groups are also subjected to different discount policy such as seniors, children, military, and group travel. Based on this fare price variation over the sale horizon, passengers are assumed to make the choice of when to purchase the ticket. Given that 98 percents of the tickets were purchased within 30 days before departure, we assume that the choice set is constituted of 31 days. Each passenger is therefore assumed to make the purchase timing decision amongst 31 booking day alternatives, from 30 days before departure (booking day 1) until departure date (booking day 31).

Based on the data set, we can only observe the fare price on the day in which passenger purchases the ticket but not on other days over the sale horizon. To accommodate choice modeling, fare prices on other days of the sale horizon have been simulated from the actual data by averaging over the observed fare prices within the same booking day over the observed month period. Different choice models have been estimated for the three market segments identified: (1) Multinomial Logit (MNL) model without socioeconomic information, (2) Mixed Logit (ML) model, (3) Latent Class (LC) model, and (4) Multinomial Logit (MNL) model with socioeconomic information.
6. MODEL FORMULATION

6.1 Multinomial Logit (MNL) model specification

The independent variables that enter the final models are advance booking (number of day before departure), fare price ($), and weekend departure dummies. Fare price and advance booking variables aim to capture passenger tradeoff behavior between early booking with cheaper fare and late booking with higher fare. The model specification allows for price sensitivity and assumes different value parameters across booking periods to accommodate the assumption that passengers have different price sensitivity over the sale horizon. The booking periods are grouped such that booking days within the same booking period have approximately the same number of reservations. These 6 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. The weekend dummy variable aims to capture corresponding unobserved effects. For instance, trip purpose is believed to influence ticket booking time; leisure oriented passengers are generally expected to plan their trip in advance. The weekend dummy coefficients are allowed to take different values across booking periods to account for the departure day effect toward ticket booking time.

6.2 Mixed Logit (ML) model specification

In the mixed logit specification, the fare price is specified as a random coefficient with log normal distribution. The log normal distribution is imposed to restrict the price estimate to have negative sign.

6.3 Latent Class (LC) model specification

In the latent class model, the explanatory variable of the choice model includes fare price ($) and advance booking (number of day before departure). The specification allows for passengers to have different price sensitivities for different booking periods similar to MNL specification. In addition to the explanatory variable of the choice model, other elements of the booking data are extracted to segment demand and capture heterogeneity of behavior across passenger. The explanatory variables for the class membership model include:
Departure day of week: A dummy variable is 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).

Departure time of day: A dummy variable is used to indicate whether the trip is taken on a particular time of day. We use six departure times as suggested by Jin (20) 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 time of day (except evening) is used for the class membership model.

6.4 Multinomial Logit (MNL) model with socioeconomic information specification

The first multinomial logit model estimated is based on the assumption that the booking data do not contain socioeconomic information. However, given that different passenger groups for this railway service are subjected to different discount policy such as seniors, children, military, and group travel, we could use ticket discount information to identify passenger types of different socioeconomic characteristics. We have been able to identify nine passenger types from the data set. Unidentified passengers are grouped into passenger type ten.

In the model specification, we specify price sensitivities differently for each passenger type to accommodate the assumption that price sensitivities vary by passenger type. However, on the departure day (booking period 6), given that passengers generally become insensitive to price and behave similarly across passenger types, price sensitivity of booking period 6 is specified separately and assumed to be the same across all passenger types.

7. RESULTS AND DISCUSSIONS

The estimation results derived from multinomial logit, mixed logit, and latent class model are reported in Tables 1 to 3. The multinomial and mixed logit models are estimated with AMLET (Another Mixed Logit Estimation Tool) (21). 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 (22).
Table 1 Long distance model

<table>
<thead>
<tr>
<th>Variable</th>
<th>Class1</th>
<th>Class2</th>
<th>MNL with Socioeconomics</th>
</tr>
</thead>
<tbody>
<tr>
<td>advbk</td>
<td>Est  -0.184</td>
<td>T-Stat  42.872*</td>
<td>Est  -0.370</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>price.period1</td>
<td>Est  -0.006</td>
<td>T-Stat  3.634*</td>
<td>Est  -5.318</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>price.period2</td>
<td>Est  -0.012</td>
<td>T-Stat  9.173*</td>
<td>Est  (2.830)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>price.period3</td>
<td>Est  -0.011</td>
<td>T-Stat  10.783*</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>price.period4</td>
<td>Est  -0.010</td>
<td>T-Stat  10.380*</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>price.period5</td>
<td>Est  -0.005</td>
<td>T-Stat  5.287*</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>price.period6</td>
<td>Est  -0.003</td>
<td>T-Stat  3.366*</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Class Size       | 0.619      | 0.381      | 0.3428                  |
| R²               | 0.2945     | 0.2904     | 0.2442                  |
| Log-likelihood at optimal | -90,226 | -91,070 | -90,487                  |
| No. of observation | 37,373    | 37,373     | 37,373                  |

*Statistically significant at 5% significance level. Parenthesis indicates standard deviation.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>advbk</td>
<td>-0.160</td>
<td>30.287</td>
<td>*</td>
<td>-0.392</td>
<td>100.986</td>
<td>*</td>
<td>advbk</td>
<td>-0.084</td>
<td>-10.870</td>
<td>*</td>
<td>-0.672</td>
<td>-16.043</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>price.period1</td>
<td>-0.024</td>
<td>5.481</td>
<td>*</td>
<td>-0.640</td>
<td>35.723</td>
<td>*</td>
<td>price.period1</td>
<td>0.011</td>
<td>3.826</td>
<td>*</td>
<td>-0.014</td>
<td>-0.917</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>price.period2</td>
<td>-0.021</td>
<td>9.799</td>
<td>*</td>
<td>-0.024</td>
<td>25.505</td>
<td>*</td>
<td>price.period2</td>
<td>0.012</td>
<td>3.851</td>
<td>*</td>
<td>-0.130</td>
<td>-2.310</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>price.period3</td>
<td>-0.021</td>
<td>9.799</td>
<td>*</td>
<td>-0.024</td>
<td>25.505</td>
<td>*</td>
<td>price.period3</td>
<td>0.012</td>
<td>3.826</td>
<td>-0.073</td>
<td>-4.953</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>price.period4</td>
<td>-0.012</td>
<td>5.593</td>
<td>*</td>
<td>-0.012</td>
<td>5.593</td>
<td>*</td>
<td>price.period4</td>
<td>0.022</td>
<td>4.696</td>
<td>*</td>
<td>-0.087</td>
<td>-5.805</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>price.period5</td>
<td>-0.008</td>
<td>4.074</td>
<td>*</td>
<td>-0.008</td>
<td>4.074</td>
<td>*</td>
<td>price.period5</td>
<td>0.029</td>
<td>6.196</td>
<td>*</td>
<td>-0.093</td>
<td>-6.019</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>price.period6</td>
<td>-0.008</td>
<td>4.074</td>
<td>*</td>
<td>-0.008</td>
<td>4.074</td>
<td>*</td>
<td>price.period6</td>
<td>0.010</td>
<td>1.792</td>
<td>-0.077</td>
<td>-5.371</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wknd.period1</td>
<td>1.163</td>
<td>21.812</td>
<td>*</td>
<td>4.299</td>
<td>57.340</td>
<td>*</td>
<td>Intercept</td>
<td>0.952</td>
<td>8.602</td>
<td>*</td>
<td>-0.952</td>
<td>-8.602</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wknd.period2</td>
<td>1.203</td>
<td>29.228</td>
<td>*</td>
<td>1.747</td>
<td>43.787</td>
<td>*</td>
<td>Monday</td>
<td>-1.212</td>
<td>-13.484</td>
<td>*</td>
<td>1.212</td>
<td>13.484</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wknd.period3</td>
<td>0.751</td>
<td>19.745</td>
<td>*</td>
<td>0.168</td>
<td>4.589</td>
<td>*</td>
<td>Tuesday</td>
<td>-1.061</td>
<td>-12.027</td>
<td>*</td>
<td>1.061</td>
<td>12.027</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wknd.period4</td>
<td>0.537</td>
<td>17.042</td>
<td>*</td>
<td>-0.664</td>
<td>20.754</td>
<td>*</td>
<td>Wednesday</td>
<td>-1.161</td>
<td>-12.777</td>
<td>*</td>
<td>1.161</td>
<td>12.777</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wknd.period5</td>
<td>-0.149</td>
<td>3.310</td>
<td>*</td>
<td>-1.187</td>
<td>27.385</td>
<td>*</td>
<td>Thursday</td>
<td>-1.196</td>
<td>-13.013</td>
<td>*</td>
<td>1.196</td>
<td>13.013</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wknd.period6</td>
<td>-0.504</td>
<td>13.923</td>
<td>*</td>
<td>-1.353</td>
<td>30.550</td>
<td>*</td>
<td>Friday</td>
<td>-0.859</td>
<td>-11.172</td>
<td>*</td>
<td>0.859</td>
<td>11.172</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Saturday</td>
<td>-0.837</td>
<td>-9.565</td>
<td>*</td>
<td>0.837</td>
<td>9.565</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Early morning</td>
<td>0.977</td>
<td>11.940</td>
<td>*</td>
<td>-0.977</td>
<td>-11.940</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AM peak</td>
<td>0.852</td>
<td>14.129</td>
<td>*</td>
<td>-0.852</td>
<td>-14.129</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AM off peak</td>
<td>0.428</td>
<td>8.998</td>
<td>*</td>
<td>-0.428</td>
<td>-8.998</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PM off peak</td>
<td>0.109</td>
<td>2.949</td>
<td>*</td>
<td>-0.109</td>
<td>-2.949</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PM peak</td>
<td>-0.076</td>
<td>-2.027</td>
<td>*</td>
<td>0.076</td>
<td>2.027</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Statistically significant at 5% significance level. Parenthesis indicates standard deviation.
Table 3 Short distance model

<table>
<thead>
<tr>
<th>Variable</th>
<th>Est</th>
<th>T-Stat</th>
<th>Variable</th>
<th>Est</th>
<th>T-Stat</th>
<th>Variable</th>
<th>Est</th>
<th>T-Stat</th>
<th>Variable</th>
<th>Est</th>
<th>T-Stat</th>
</tr>
</thead>
<tbody>
<tr>
<td>advbk</td>
<td>-0.534</td>
<td>20.090 *</td>
<td>advbk</td>
<td>-1.164</td>
<td>-6.080 *</td>
<td>advbk</td>
<td>-0.674</td>
<td>32.119 *</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>price.period1</td>
<td>0.056</td>
<td>7.863 *</td>
<td>price.period1</td>
<td>-0.378</td>
<td>-1.451</td>
<td>price.period1</td>
<td>-0.215</td>
<td>22.716 *</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>price.period2</td>
<td>0.011</td>
<td>2.760 *</td>
<td>price.period2</td>
<td>-0.518</td>
<td>-1.953</td>
<td>price.period2</td>
<td>-0.327</td>
<td>-10.734 *</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>price.period3</td>
<td>-0.007</td>
<td>2.626 *</td>
<td>price.period3</td>
<td>-0.448</td>
<td>-2.712 *</td>
<td>price.period3</td>
<td>-0.327</td>
<td>-10.734 *</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>price.period4</td>
<td>-0.019</td>
<td>11.389 *</td>
<td>price.period4</td>
<td>-0.606</td>
<td>-3.504 *</td>
<td>price.period4</td>
<td>-0.089</td>
<td>-3.597 *</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>price.period5</td>
<td>-0.016</td>
<td>12.133 *</td>
<td>price.period5</td>
<td>-0.595</td>
<td>-3.399 *</td>
<td>price.period5</td>
<td>-0.081</td>
<td>-3.202 *</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>price.period6</td>
<td>-0.008</td>
<td>8.293 *</td>
<td>price.period6</td>
<td>-0.426</td>
<td>-2.788 *</td>
<td>price.period6</td>
<td>-0.090</td>
<td>-4.031 *</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| wknd.period1  | 2.701 | 9.230 * | wknd.period1  | 0.318 | 4.444 * | wknd.period1  | 0.318 | 4.444 * |
| wknd.period2  | 1.560 | 7.202 * | wknd.period2  | 1.438 | 3.920 * | wknd.period2  | 1.438 | 3.920 * |
| wknd.period3  | -0.268 | 1.183 | wknd.period3  | -0.255 | 0.875 | wknd.period3  | -0.255 | 0.875 |
| wknd.period4  | -0.639 | 4.205 * | wknd.period4  | -1.129 | 4.460 * | wknd.period4  | -1.129 | 4.460 * |
| wknd.period5  | -0.758 | 4.708 * | wknd.period5  | -1.108 | 3.839 * | wknd.period5  | -1.108 | 3.839 * |
| wknd.period6  | 0.404 | 4.267 * | wknd.period6  | 0.618 | 2.516 * | wknd.period6  | 0.618 | 2.516 * |

| Early morning | -1.117 | -5.476 * | AM peak        | -0.922 | -9.202 * | AM peak        | -0.922 | -9.202 * |
| AM off peak   | -0.297 | -3.406 * | AM off peak    | -0.297 | 3.406 * | AM off peak    | -0.297 | 3.406 * |
| PM off peak   | -0.083 | -1.007 | PM off peak    | 0.083  | 1.007 | PM off peak    | 0.083  | 1.007 |
| PM peak       | -0.287 | -3.548 * | PM peak        | 0.287  | 3.548 * | PM peak        | 0.287  | 3.548 * |

| No. of observation | 4,454 | 4,454 | 4,454 | 4,454 |
| Rho-squared:       | 0.5637 | 0.5493 | R²(0) | 0.6766 |
| Adjusted rho-squared: | 0.5628 | 0.5487 | R²    | 0.4528 |
| Log-likelihood at optimal | -6,674 | -6,894 | | -6,356 |
| Log-likelihood at zero | -15,295 | | | -6,713 |

*pStatistically significant at 5% significance level. Parenthesis indicates standard deviation.*
7.1 Results of multinomial logit model
Results obtained with MNL indicate that disutility is associated with advance booking and that the lack of flexibility to change travel plan is negatively perceived. Consumers 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 5% significance level in all the three models estimated. The decreasing magnitude of price sensitivity as booking period approaches departure is in line with the expectation. Passengers are believed to be most sensitive to fare price at the beginning of the sale horizon and less concerned about seat availability. As time approaches departure, passengers become less sensitive to fare price especially on the departure day and more concerned about seats availability.

The weekend dummies show the expected pattern; the value decreases as the booking period approaches departure indicating that passenger traveling on weekend generally purchase ticket earlier in advance compared to passenger traveling on weekday.

7.2 Results of mixed logit model
The mixed logit model accounting for heterogeneity in price sensitivity shows a better model fit when compared to MNL, except for the short distance market. The model results show the expected sign for advance booking, fare price and weekend dummies as observed in the multinomial logit models. The estimates are all statistically significant at the 5% significance level except for the weekend dummy at booking period 4 (long distance market) and booking period 3 (short distance market).

7.3 Results of latent class model
In this context, the main advantage of the latent class specification over multinomial logit model is the ability to identify distinct group of passengers with respect to preferences for ticket booking time. More specifically, the six departure time of day and seven day of week define a set of 42 underlying scenarios of booking, also called covariate pattern. For each scenario, the likelihood that a booking belongs to a latent class can be calculated as the logit probability
associated with the parameter estimates of the class membership model. Figure 3 to Figure 5 represent the probability of belonging to class 1 for the 42 different covariate patterns.

The results obtained with the latent class model are coherent with those given by the multinomial logit and mixed logit models except that in the medium distance market, passenger in class 1 are shown to be insensitive to the price coefficient, that results to be positive. Given that the magnitude of the coefficients within the same model could not be compared between different classes due to scale parameter (16), the ratio of advance booking coefficient to the price coefficient is calculated for each choice model. This ratio can be viewed as the willingness to pay (WTP) to delay the ticket purchase for one day. The two latent classes specified in the class membership model are assumed to segment passengers into business oriented and leisure oriented passenger. Given that the leisure oriented passenger generally know their travel plan earlier in advance while the business traveler generally book the ticket closer to departure date, we assume that passenger with higher WTP for purchase delay is likely to be business oriented passenger.

For the long distance market results in Table 1, the results show higher WTP for purchase delay in class 1 (ranging from $13.90 to $99.29 per day) than class 2 (ranging from $11.03 to $18.38 per day) for the majority of the booking periods, indicating that passengers from class 1 are willing to pay more for the possibility to change their travel plans. Based on our assumption, passengers in class 1 are believed to be business oriented travelers which accounts for 61.9 % of this market (see Table 1). More specifically, the class membership model indicates that passengers departing from early morning to AM peak are predominantly class 1 passenger (see Figure 3).

For the medium distance market results in Table 2, WTP for class 1 (price insensitive) is higher than for class 2 (ranging from $5.18 to $46.67 per day) for most of the booking periods considered. The results indicate that passenger in class 1 are predominantly business oriented travelers which accounts for 56.6 % of this market (see Table 2). More specifically, the class membership model indicates that passenger class 1 (business oriented travelers) predominantly depart from early morning to AM off peak especially on Sunday as shown in Figure 4.
Based on the short distance market results in Table 3, WTP for purchase delay of class 1 (ranging from $1.92 to $3.08 per day) is lower than for class 2 (ranging from $3.64 to $4.99 per day). The result indicates that passengers in class 1 are predominantly leisure oriented travelers which accounts for 59.5% of this market (see Table 3). More specifically, the class membership model indicates that passenger class 1 (leisure oriented travelers) predominantly depart from AM off peak until evening as shown in Figure 5.

![Figure 3 Probability of passenger belonging to class 1 in long distance model](image-url)
7.4 Results of multinomial logit model with socioeconomic information

Finally we report the results obtained from the estimation of a multinomial logit model accounting for deterministic taste heterogeneity (Table 1 to Table 3). The data set contains a limited number of observations including passenger types. This makes impossible the estimation
of price sensitivity for a certain number of class and results into statistically insignificant coefficients for some of the classes considered (i.e. child and unaccompanied child for the long distance market).

For the long distance market, the results are in line with the expectation. The price sensitivity in the last booking period shows relatively low magnitude (-0.019) indicating that passenger become less sensitive to fare price. Unidentified passenger type shows the highest price sensitivity (-0.043); passengers in this group are subjected to special discount, thus it is reasonable that they are highly price sensitive. Military adult is the passenger type which is the second most price sensitive (-0.028). The adult passenger type subjected to full fare amount shows price sensitivity equal to disabled adult (-0.024). Passengers subjected to student advantage discount and adult passenger with AAA membership have equal price sensitivity (-0.010) which is the lowest price sensitivity among all the passenger types.

For the medium distance market, the price sensitivity in the last booking period shows relatively low magnitude (-0.058) as expected. The unidentified passenger type shows the largest price sensitivity (-0.088) among all passenger types. Child is the passenger type which is the second most price sensitive (-0.071) followed by adult passenger (-0.064). Senior (-0.063), unaccompanied child (-0.063), and disabled adult (-0.063) are slightly less price sensitive than adult. Passengers associated with student advantage (-0.054) and adult with AAA membership (-0.057) are the least price sensitive.

For the short distance market, the estimation results show slightly different pattern compared to long and medium distance market. The result indicates that the military group is the most price sensitive (-1.084). They are followed by unidentified passenger (-0.238), disabled adult (-0.227), adult (-0.215), senior (-0.209), and adult with AAA membership (-0.158) respectively. The passenger subjected to student advantage discount is the least price sensitive (-0.099).

The weekend dummies indicate that passenger departing on weekend generally purchase their ticket ahead of time.
7.5 Comparison of Model Fit

To conclude, for the long distance market, the mixed logit provides the best fit with the rho-square slightly lower than the one obtained with the latent class model. Both mixed logit and latent class models outperform multinomial logit models (with and without socioeconomics) as expected. Mixed logit provide the best statistical fit for the medium distance market; it outperforms both multinomial logit specifications (with and without socioeconomics). Latent class provides the best statistical fit in short distance market while in this case the multinomial logit models (with and without socioeconomics) outperform mixed logit model.

The results also indicate that segmenting passengers by booking period appears to be more appropriate than segmenting passengers by socioeconomic information for all the three markets.

8. CONCLUSIONS

In this paper advanced demand modeling approaches are proposed to study intercity passenger booking decision and to segment preferences; all the models are calibrated on internet booking data.

Modeling formulations considered include multinomial logit, mixed logit, and latent class models; markets are segmented on trip distances: long, medium, and short. The results show that the following variables: fare price, advance booking (number of day before departure), and departure day of week, can be used as determinants affecting ticket booking. Mixed logit and latent class models are then applied to account for taste heterogeneity. The results indicate that mixed logit model provides the best statistical fit for the long distance and medium distance markets, while the latent class model provides the best statistical fit for the short distance market. Results also indicate that segmenting passengers by booking period provides better fit than segmenting passengers by socioeconomic information.

Results from this study prove that advanced demand models can be estimated on internet booking data and that market segmentation can be obtained even with limited knowledge of socio-demographic characteristics of the population. A number of directions could be considered for future research. First, the proposed choice model could be integrated into an optimization algorithm for maximizing revenues. Second, the proposed model is based on a static approach
where decisions are made at a single point in time. Dynamic discrete choice models are expected to provide a significant improvement in prediction accuracy by offering the possibility to account for the evolving characteristics of the market over time and for sequential purchasing decisions. Third, the modeling approach applied here to railway revenue management could be applied to other problems for which just internet booking data are available (i.e. shippers, or couriers selling their services online).
REFERENCES


2. Berry J. Why web sites fall short—according to CSC study, top management and IT are out of synch. Presented at Internet Week, 1999, 765:27–8 (May 17).


