Transportation Research Part A 38 (2004) 435–453

TRANSPORTATION RESEARCH PART A

www.elsevier.com/locate/tra

Uncovering the contribution of travel time reliability to dynamic route choice using real-time loop data

Henry X. Liu [a,*], Will Recker [b], Anthony Chen [a]

[a] Department of Civil and Environmental Engineering, Utah State University, Logan, UT 84322, USA

[b] Department of Civil and Environmental Engineering, Institute of Transportation Studies University of California, Irvine, CA 92697, USA

Abstract

Travel time reliability has generally been surmised to be an important attribute of transportation systems. In this paper, we study the contribution of travel time reliability in travelers' route choice decisions. Traveler's route choice is formulated as a mixed-logit model, with the coefficients in the model representing individual traveler's preferences or tastes towards travel time, reliability and cost. Unlike the traditional approach involving the use of traveler surveys to estimate model coefficients and thereby uncover the contribution of travel time reliability, we instead apply the methodology to real-time loop detector data, and use genetic algorithm to identify the parameter set that results in the best match between the aggregated results from traveler's route choice model and the observed time-dependent traffic volume data from loop detectors. Based on freeway loop data from California State Route 91, we find that the estimated median value of travel-time reliability is significantly higher than that of travel-time, and that the estimated median value of degree of risk aversion indicates that travelers value a reduction in travel time variability more highly than a corresponding reduction in the travel time for that journey. Moreover, travelers' attitudes towards congestion are not homogeneous; substantial heterogeneity exists in travelers' preference of travel time and reliability. Our results validate results from previous studies involving the California State Route 91 value-pricing project that were based on traditional traveler surveys and demonstrate the applicability of the approach in travelers' behavioral studies.

Keywords: Travel time reliability; Route choice; Mixed logit; Random utility; Genetic algorithm

* Corresponding author.
E-mail addresses: xliu@cc.usu.edu (H.X. Liu), wwrecker@uci.edu (W. Recker), achen@cc.usu.edu (A. Chen).

0965-8564/$ - see front matter
doi:10.1016/j.tra.2004.03.003

1. Introduction

It is accepted that a wide range of factors influences the route choice of individual travelers. In addition to such factors as perceived travel time, monetary cost, comfort and safety, the reliability of travel time has been generally conceded to be an important factor, particularly for trips, such as journey-to-work, where time constraints (e.g. arrival time) may impose significant penalties on an individual. Reliability, by its nature, implies something about the certainty or stability of travel time of any particular trip under repetition. As such, reliability is closely associated with the statistical concept of variability. Variability could result from the differences in the mix of vehicle types on the network for the same flow rates, differences in driver reactions under various weather and driving conditions, differences in delays experienced by different vehicles at intersections, and such random incidents as vehicle breakdown and signal failure, etc. Variability in network travel times introduces uncertainty for travelers in that they do not know with certainty when they will arrive at their respective destinations. This risk (or added cost) to a traveler making a trip may be manifest in a willingness to pay a premium (e.g., through use of toll roads or HOV lane) to avoid congestion and to achieve greater reliability in travel times.

Although travel time reliability ostensibly plays an important role in the traveler's route choice behavior (Abdel-Aty et al., 1995; Bates et al., 2001; Lam and Small, 2001; Small et al., 2002), many questions related to an understanding of the effects of reliability on the traveler's route choice decision making remain unanswered. How do travelers value travel time and its reliability, how much does the travel time reliability contribute to travelers' route choice, and how much variation is there in travelers' preferences regarding the potential tradeoff between reliability and travel time itself? Answering these questions can help in the design and evaluation of transportation planning and operation strategies, but requires that this attribute be accounted for explicitly in the modeling of travelers' choice.

The above questions could be studied either through direct or indirect methods (Jackson and Jucker, 1981). The direct method involves posing a series of questions to a sample of travelers in a certain population, usually either as a revealed preference (RP) survey, in which the actual behavioral response to the traffic condition is reported, or as a stated preference (SP) survey, in which a traveler's behavior in hypothetical scenarios is reported. The indirect method involves inferring the answers to these behavioral questions from observed data describing the flows on alternative routes connecting an origin–destination pair.

Most previous research that has attempted to address issues of reliability using direct methods has analyzed RP data and/or SP data. Abdel-Aty et al. (1995) conducted a study to investigate effect of travel time variability on route choice using repeated measurement SP data. Their results indicated the significance of both the degree of travel time variation and traffic information on route choice. Bates et al. (2001) provided a comprehensive overview of the theory underlying the valuation of reliability, and discussed the empirical issues in data collection. Because of the difficulties of finding real choice situations with sufficient variation to allow statistically reliable estimates to be obtained for RP data, they acknowledged the value of SP data, and applied SP data in the study of valuation of reliability in passenger rail services. Although SP is usually de facto the only realistic method for data collection, Lam and Small (2001) leveraged the opportunity of a road pricing project and measured values of travel time and reliability from 1998 RP data on actual behavior of commuters on State Route 91 (SR91) in Orange County, California.

Recently, Small et al. (2002) continued their previous studies by combining both RP and SP data on SR91 to empirically identify the varied nature of traveler preference for travel time and reliability. They found that highway users exhibit substantial heterogeneity in their valuation of travel time and reliability.

However, both RP and SP data have drawbacks. Although a greater level of detailed data could be obtained from the survey of individual travelers, and these data may lead to a higher degree of accuracy for the estimation, data collection for the survey of large sample size and its following analysis are time-consuming and expensive. For SP data, it is not sufficient to say that the response to hypothetical situations really reflects traveler's behavioral choice to actual situations. For instance, because people tend to overstate the time delays they actually experienced, it is common that they respond more to a given actual time saving than to a hypothetical time saving of the same amount. Therefore, the estimated value of travel time may be lower than the actual. Although there is a general econometric tradition for favoring RP data, there are often serious problems in achieving the level of detail in data that is ideally required. In the study of travel time reliability, as noted by Bates et al. (2001), it is virtually impossible to find RP situations where there is sufficient perceived variation to allow statistically reliable estimates—notable exceptions to this are the studies conducted by Lam and Small (2001) and Small et al. (2002) with the California State Route 91 value-pricing project.

Alternatively, advancements in traffic surveillance and monitoring technologies, including real-time data from inductive loop detectors, can provide valuable aggregated information that ostensibly resulted from the disaggregated individual travel route choices. Instead of surveying motorists on their choices in the direct method, we propose an indirect method to answer some behavioral questions in such a way that the estimated choice probability resulting from route choice model matches the revealed probability exhibited in the real-time loop detector data. To the best of our knowledge, there are but a very few research works that study behavioral issues using the indirect method, and this paper aims to fill this gap.

In this paper, traveler's route choice is formulated as a mixed-logit model (also known as random coefficient logit and logit kernel), which generalizes the standard logit model by allowing the coefficient associated with each observed variable to vary randomly across individuals (McFadden and Train, 2000; Bhat, 2001; Bhat and Castelar, 2002). The coefficients in the model represent individual traveler's preferences or tastes toward travel time, reliability and cost. To find the distribution of the model coefficients and thereby uncover the contribution of travel time reliability in the dynamic route choice, we use genetic algorithm (GA) to identify the parameter set that results in the best match between the aggregated results from traveler's route choice model and the observed time-dependent traffic volume data from loop detectors.

We apply the proposed approach to measure value of time (VOT), value of reliability (VOR), and degree of risk aversion (DORA) simultaneously using data on actual travel behavior drawn from a real pricing context. A recent value pricing project on a major commuting highway, State Route 91 (SR91) in Orange County, California, gives travelers the option to travel free on the regular lanes or to pay a time-varying price for express travel on toll lanes situated along the median of the highway. Based on their respective choices of whether or not to pay a toll for the congestion-free travel, we observe the outcome from the choice probability between the two parallel routes in the form of loop detector data on 30-s averages of count and occupancy. We find that travel time reliability plays an important role in traveler's decision making for route choice.

Moreover, the results indicate that travelers value travel-time reliability substantially higher than they do travel-time savings. Our results validate results from such previous studies as Lam and Small (2001) and Small et al. (2002), and demonstrate the applicability of our approach in travelers' behavioral studies.

The rest of the paper is organized as follows. The next section discusses the route choice structure and formulation using mixed-logit model. Section 3 presents estimation procedure and solution algorithm to identify the unknown parameters in the route choice model. We then apply the proposed methodology to the SR91 data and describe empirical results in Sections 4 and 5. The final section provides a summary of research findings.

2. Route choice structure and formulation

Travelers' route choice among the available options will reflect their perception of the costs and benefits associated with each option. If the costs or benefits are perceived to be uncertain, the choice will be influenced by the travelers' attitude to that uncertainty. We incorporate the stochasticity of route travel time as a measure of the risk associated with the selection of specific routes. On the basis of the perceived distribution of network travel times, travelers are assumed to behave differently when considering routes for which their perceived travel times have a probabilistic component. Some are risk averse, choosing routes with longer expected travel times but smaller variations. Others, the risk takers, may choose routes with shorter expected travel times but greater variations in travel time reliability.

In modeling this behavior, we assume that the individual traveler has a subjective perception of the probability distribution of travel time for each available route. Additionally, we assume that there exists an objective distribution of travel time based on actual measurement over a suitably defined time period. This stochastic travel time reflects the intrinsic fluctuations in the transportation network, as noted, due to particular weather conditions, unpredictable lane closures, and traffic accidents, etc. It is not necessarily the case that the subjective and objective probability distributions over route travel time are identical. (They may differ, and when this is the case, the existence of perception errors is witnessed.) However, for travelers in peak-hour commuting trips it seems reasonable to assume that these two distributions are same, i.e., drivers have no misperceptions of either travel time or travel time variation. This is the case for the model considered here; we leave the incorporation of traveler's perception errors for future studies.

We assume that travelers consider travel time, travel-time reliability (i.e., risk), and out-of-pocket monetary cost (such as toll) in their choices of routes. Moreover, we assume that they value any tradeoff between travel time and travel-time reliability differently depending on individual tastes, and that such tastes are distributed in the population in a manner that covers a spectrum in terms of the degree of risk aversion. We further assume that drivers are rational and are maximizing some utility measure, and suggest a common disutility functional form, but with coefficients (disutility weights) that reflect individual's preference. Traveler's perceived disutility is a function of the route travel time, travel time variability, monetary cost, and the individual traveler's attitude toward these three variables at each time. Specifically, we assume that the traveler is faced with a choice among P_{rs} alternative freeway routes between a freeway on-ramp origin, r, and a corresponding freeway off-ramp, s. (In this formulation, we assume that surface street travel is irrelevant to the choice and that the freeway origin and destination are fixed.) Under these assumptions, the disutility to traveler n of travel, commencing at time t, along path p linking on-ramp origin r and off-ramp destination s is specified as:

$$U_{np}(t) = \beta_n' x_{np}(t) + \varepsilon_{np} \tag{1}$$

where $x_{np}(t)$ is a vector of observed variables (including alternative specific constants), β_n is a corresponding coefficient vector that may vary over individuals but does not vary across alternatives or time, and ε_{np} is an unobserved extreme value random term that captures the idiosyncratic effect of all omitted variables that are not individual specific. ε_{np} is assumed to be identically and independently distributed across all choice occasions and independent of $x_{np}(t)$ and β_n.

Since only commuting trips are considered and we assume that the traveler's subjective distribution is identical to the objective distribution of route travel time, we omit n from the subscripts of $x_{np}(t)$. Therefore, $x_p(t)$ is a vector of observed variables based on the actual measurements from the field. Specifically, $x_p(t) = [T_p(t), R_p(t), C_p(t)]'$, where $T_p(t)$ measures route travel time, $R_p(t)$ measures travel-time reliability, and $C_p(t)$ is the monetary toll cost. Therefore, the value of travel time (VOT), value of travel-time reliability (VOR), and the degree of risk aversion (DORA) are defined as:

$$\mathrm{VOT}_n = \frac{\partial U_{np}(t)/\partial T_p(t)}{\partial U_{np}(t)/\partial C_p(t)} = \frac{\beta_n^{\mathrm{T}}}{\beta_n^{\mathrm{C}}}; \tag{2}$$

$$\mathrm{VOR}_n = \frac{\partial U_{np}(t)/\partial R_p(t)}{\partial U_{np}(t)/\partial C_p(t)} = \frac{\beta_n^{\mathrm{R}}}{\beta_n^{\mathrm{C}}}; \tag{3}$$

$$\mathrm{DORA}_n = \frac{\partial U_{np}(t)/\partial R_p(t)}{\partial U_{np}(t)/\partial T_p(t)} = \frac{\beta_n^{\mathrm{R}}}{\beta_n^{\mathrm{T}}}. \tag{4}$$

where $\beta_n = [\beta_n^{\mathrm{T}}, \beta_n^{\mathrm{R}}, \beta_n^{\mathrm{C}}]'$, a vector of coefficients reflecting individual n's particular tastes toward travel time, reliability, and monetary cost. As the notation indicates, the models we consider are specified so that VOT, VOR, and DORA depend on the individual traveler n but not on the choice instant t. DORA_n reflects the degree of risk aversion, i.e., the extent to which travel time variability is undesirable to traveler n. The larger the value of DORA, the higher the perceived cost of uncertainty, and the more risk averse the traveler.

To account for taste variation across individuals, preference heterogeneity is introduced by assuming that the coefficients β_n are realizations of random variables $\boldsymbol{\beta}$. These coefficients are assumed to vary over travelers based on individual characteristics in the population with density $f(\boldsymbol{\beta})$. This density is a function of parameters $\boldsymbol{\Theta}$ that represent, for example, the mean and the covariance of the $\boldsymbol{\beta}$ in the population. This specification, known as "mixed" logit or "random coefficients" logit, is identical to standard logit except that β_n varies over decision makers rather than being fixed. As such, the probability that traveler n will select path p, conditioned on β_n, is given by:

$$L_{np}(\beta_n; t) = \frac{\mathrm{e}^{\beta_n' x_{np}(t)}}{\sum_{\forall j \in P^{rs}} \mathrm{e}^{\beta_n' x_{nj}(t)}} \tag{5}$$

The unconditional probability is the integral of $L_{np}(\beta_n; t)$ over the distribution of all possible values of β_n, i.e.,

$$P_{np}(t) = \int \frac{e^{\boldsymbol{\beta}' x_{np}(t)}}{\sum_{\forall j \in P^{rs}} e^{\boldsymbol{\beta}' x_{nj}(t)}} \cdot f(\boldsymbol{\beta})\, d\boldsymbol{\beta} \tag{6}$$

Eq. (6) is the general form for the so-called mixed logit probability. Train (2002) provides an excellent overview of the properties of such models and procedures for their estimation. Choice probabilities can be estimated using Monte-Carlo simulation to integrate the computational difficult parts of the preference distribution. Assuming that the parameters of $f(\boldsymbol{\beta})$ are $\boldsymbol{\Theta}$, the unbiased estimator $\breve{P}_{np}$ for P_{np} can be obtained as follows:

1. Select trial values for $\boldsymbol{\Theta}$
2. Draw Q values of $\boldsymbol{\beta}$ from $f(\boldsymbol{\beta}|\boldsymbol{\Theta})$; label the qth such value $\boldsymbol{\beta}^q$
3. Calculate $L_{np}(\boldsymbol{\beta}^q; t)$; $q = 1, 2, \ldots, Q$
4. Compute average simulated probability as:

$$\breve{P}_{np}(t) = \frac{1}{Q} \sum_{q=1}^{Q} L_{np}(\boldsymbol{\beta}^q; t) \tag{7}$$

An issue of terminology arises here since there are two sets of parameters in a mixed logit model (Train, 2002). First, we have the parameters $\boldsymbol{\beta}$, which enter the logit formula. These parameters have density $f(\boldsymbol{\beta})$. The second set of parameters that describe this density. For example, if we assume that $\boldsymbol{\beta}$ is normally distributed with mean b and covariance W, then b and W are parameters that describe the density $f(\boldsymbol{\beta})$.

In this study, we assume that the parameters $\boldsymbol{\beta}$ have independent normal distributions. If we further place the traveler in one of the M groups defined by their access to information (the assumption being that more risk adverse travelers may be more inclined to search out travel-time information), we would specify $\boldsymbol{\beta} \sim N(b_m, W_m)$; $m \in M$.

With the assumption that the random parameters $\boldsymbol{\beta}^{\mathrm{T}}$, $\boldsymbol{\beta}^{\mathrm{V}}$, $\boldsymbol{\beta}^{\mathrm{C}}$ for time, reliability and cost have normal distributions, i.e., $\beta^{\mathrm{T}} \sim N(b^{\mathrm{T}}, W^{\mathrm{T}})$, $\beta^{\mathrm{V}} \sim N(b^{\mathrm{V}}, W^{\mathrm{V}})$, $\beta^{\mathrm{C}} \sim N(b^{\mathrm{C}}, W^{\mathrm{C}})$, the parameter sets that need to be estimated are $\{b^{\mathrm{T}}, W^{\mathrm{T}}, b^{\mathrm{V}}, W^{\mathrm{V}}, b^{\mathrm{C}}, W^{\mathrm{C}}\} \subset \boldsymbol{\Theta}$. The estimation of $\boldsymbol{\Theta}$ will define VOT, VOR, and DORA, and thereby uncover the relative roles of travel time and travel time reliability in route choice, as well as identify the distribution of the population along the risk dimension. We let $\boldsymbol{\Omega}$, called the parameter space, denote the set of all possible values that parameters $\boldsymbol{\Theta}$ could assume. The estimation of $\boldsymbol{\Theta}$ is to search $\boldsymbol{\Omega}$ and find the best $\boldsymbol{\Theta}$ satisfying certain criteria. We provide the estimation procedure in the next section.

3. Estimation procedure and solution

Traditionally, the parameters in the mixed logit model are estimated based on RP and/or SP data by simulated maximum likelihood estimation (SMLE), as described in Train (2002). However, since our observed data are aggregated responses in the form of loop counts rather than the individual responses of each traveler, the essential concept of our estimation procedure is to find the set of parameter values that results in the best match between the aggregated results from traveler's route choice model and the observed time-dependent traffic volume data from the loop detectors. Therefore, the problem considered here is a minimization program of the difference between the volume data generated from the mixed logit route choice model (we assume the dynamic origin–destination matrix is given) and the observed loop counts. Because the route choice probability in the minimization program does not admit a closed form, as shown in Eq. (7), gradient-based optimization methods require expensive computational effort to calculate the derivatives numerically and often result in finding a local optimal solution. We therefore adopt genetic algorithm to solve the minimization program.

3.1. Estimation procedure

Consider freeway trips originating from origin O located at on-ramp r during time interval $t_0 - \Delta t_0$. The total number of such trips is evident from the 30-s loop counts Q at r as:

$$\overline{Q}_r(t_0) = \sum_{j=0}^{\Delta t_0/30} Q_r(t_0 - j) \tag{8}$$

Since the dynamic Origin–Destination matrix $O(t) \rightarrow D(t)$ is presumed to be known, the total number of trips originating at $O = r$, and bound for destination $D = s$, during time interval $t_0 - \Delta t_0$ is given as:

$$\overline{Q}_{rs}(t_0) = \frac{\sum_{j=0}^{\Delta t_0/30} O_{rs}(t_0 - j)}{\sum_{\forall k \in S} \sum_{j=0}^{\Delta t_0/30} O_{rk}(t_0 - j)} \cdot \overline{Q}_r(t_0); \quad S = \{k \in \text{set of all off ramps}\} \tag{9}$$

From Eq. (6), the expected number of these trips to use any path p is given by:

$$\overline{Q}_{rs}^p(t_0) = P_{rs}^p(t_0) \cdot \overline{Q}_{rs}(t_0) \tag{10}$$

where

$$P_{rs}^p(t_0) = \int \frac{e^{\boldsymbol{\beta}'\mathbf{x}_p(t_0)}}{\sum_{\forall j \in P_{rs}} e^{\boldsymbol{\beta}'\mathbf{x}_j(t_0)}} \cdot f(\boldsymbol{\beta}) \mathrm{d}\boldsymbol{\beta} \tag{11}$$

For any loop station i on the path $p = \{r, 1, 2, \ldots, i, \ldots, s\}$, we denote the travel time from origin r to loop station i for trips starting at time t_0 as $t_i^*(t_0)$. Then the expected time at which the flow contribution from $\overline{Q}_{rs}^p(t_0)$ first will be counted (i.e., show up in the loop count poll) is $t_i^*(t_0)$; the expected end of the contribution from $\overline{Q}_{rs}^p(t_0)$ will occur at $t_i^*(t_0 + \Delta t_0)$, or $\Delta t_0/30$ polls later. For purposes of identification, we expand the notation on $t_i^*(t_0)$ to include reference to path p from r to s, i.e.,

$$t_i^*(t_0) \rightarrow t_i^*(t_0; p_{rs})$$

where p_{rs} denotes path p from on-ramp r to off-ramp s. Consider the observed loop count at some station i on the path p_{rs} over the time interval $t - \Delta t_0$, i.e.,

$$\widehat{Q}_i(t) = \sum_{j=0}^{\Delta t_0/30} Q_i(t-j) \tag{12}$$

Let

$$H(t_i^*(t_0,p_{rs}),\Delta t_0) = \begin{cases} 1, & t_i^*(t_0,p_{rs}) - \Delta t_0 \leqslant t \leqslant t_i^*(t_0,p_{rs}) \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

Then, an estimate of $\widehat{Q}_i(t)$ is given by:

$$\widehat{\widehat{Q}}_i(t) = \sum_{\forall t_0<t} \sum_{\substack{\forall p_{rs} \\ i\in p_{rs}}} H(t_i^*(t_0,p_{rs}),\Delta t_0) \cdot \overline{Q}_{rs}^p(t_0) \tag{14}$$

or,

$$\widehat{\widehat{Q}}_i(t) = \sum_{\forall t_0<t} \sum_{\substack{\forall p_{rs} \\ i\in p_{rs}}} H(t_i^*(t_0,p_{rs}),\Delta t_0) \cdot P_{rs}^p(t_0) \cdot \overline{Q}_{rs}(t_0) \tag{15}$$

or, using the estimate for $P_{rs}^p(t_0)$ given by Eq. (7),

$$\widehat{\widehat{Q}}_i(t) = \sum_{\forall t_0<t} \sum_{\substack{\forall p_{rs} \\ i\in p_{rs}}} H(t_i^*(t_0,p_{rs}),\Delta t_0) \cdot \breve{P}_{rs}^p(t_0) \cdot \overline{Q}_{rs}(t_0) \tag{16}$$

$\widehat{\widehat{Q}}_i(t)$ is a function only of known values and the unknown parameters of $f(\boldsymbol{\beta}|\boldsymbol{\Theta})$. A standard approach to selecting the values of $\boldsymbol{\Theta}$ is to minimize the mean square error (MSE) between the estimate $\widehat{\widehat{Q}}_i(t)$ and its true (observed) value $\widehat{Q}_i(t)$ over some specified time period $t_1 \leqslant t \leqslant t_2$, i.e.,

$$\min_{\boldsymbol{\Theta}}(\text{MSE}) = \int_{t_1}^{t_2} \sum_{\forall i} \left[\widehat{Q}_i(t) - \widehat{\widehat{Q}}_i(t)\right]^2 \mathrm{d}t \tag{17}$$

or, for distinct 30-s counting intervals,

$$\min_{\boldsymbol{\Theta}}(\text{MSE}) = \sum_{t=1}^{t=\frac{t_2-t_1}{30}} \sum_{\forall i} \left[\widehat{Q}_i(t) - \widehat{\widehat{Q}}_i(t)\right]^2 \tag{18}$$

3.2. Solution method

Our solution method to Eq. (18) is based on genetic algorithms (GAs). GAs are heuristic search algorithms that attempt to search the solution space in a "smart" manner on the basis of natural selection and natural genetics. When using GAs to solve an optimization problem, each solution is encoded in a string (called chromosome), which is the concatenation of sub-strings corresponding to the set of decision variables. The entire population of such strings (solutions) is called a generation. Operators such as selection, crossover, and mutation are applied to parent chromosomes to create child chromosomes. The performance of each chromosome is evaluated by a fitness function, which corresponds to the objective function of the optimization program. Chromosomes that have high fitness values have high opportunities to reproduce, by crossbreeding with other chromosomes in the population. Detailed discussions of general GAs are available in Goldberg (1989).

The convergence of GAs has been proven by Holland (1975). Although not guaranteed to find the optimal solution, GAs often are successful in finding a solution with high fitness. GAs are also considered robust because at any time step of a search (or generation), GA progresses towards the optimal solution from a population of points, instead of starting the search at a single point, which increases the likelihood that the global, rather than a local, optimum will be found (Gen and Cheng, 2000). The population-based search procedure, together with stochastic operators used in GAs for reproducing child chromosomes in the next generation, are essential concepts for GAs to locate better solutions for complex and noisy objective functions than do such conventional techniques as gradient-based search methods.

The GA procedure used in the estimation of parameter set $\mathbf{\Theta}$ that satisfies the condition defined by Eq. (18) is summarized in Fig. 1. The first step in the estimation procedure is to generate a

GA (fit_threshold, max_generation, p, r, m)

fit_threshold: a threshold specifying the termination criteria.
max_generation: maximal generations will be performed in the procedure.
p: the number of chromosomes (solutions) to be included in the population.
r: the fraction of population to be replaced by crossover at each generation (step).
m: the mutation rate.

1. Initialize population G: Generate p chromosomes at random.
2. Decode each chromosome in the population G.
3. Evaluate: Compute fitness for each chromosome in the population G.
4. While (fitness of best chromosome is less than fit_threshold or the number of generations is less than max_generation) do

   Create a new generation of chromosomes, G_{new}:

   - Select: Probabilistically select (1-r)*p members from the current population G, and add to the new generation G_{new}.
   - Crossover: Probabilistically select (r*p)/2 pairs of chromosomes from the current population G. For each pair, produce two offspring by applying the crossover operators. Add all offspring to the new generation G_{new}.
   - Mutate: Choose m percent of the new generation G_{new} with uniform probability. Apply the mutation operator.
   - Update: $G = G_{new.}$
   - Evaluate: Compute fitness for each chromosome in the new generation.
   - Find the highest fitness chromosome.
5. Decode the chromosome with the best fitness and obtain the best solution to this problem.

Fig. 1. GA-based estimation procedure.

number of random individuals as the initial population, each carrying a chromosome that represents a feasible solution. From the initial population, each chromosome is first decoded into the actual parameter values and fed into the mixed-logit dynamic route choice model described in Section 2. Since the O–D matrices are given, traffic assignment based on the mixed-logit route choice model can be preformed. The fitness function, in which the objective function of Eq. (18) is embedded, is then evaluated using the estimated volumes from the route choice model and the observed volumes from the field data. If the stopping criterion is not met, a set of GA operators (including selection, crossover and mutation) are applied to the chromosomes in the current population to produce offspring. The reproduction cycle including decoding of chromosomes and fitness evaluation is repeated until the stopping criterion is met or the predetermined number of generations is reached. The major components in this genetic-algorithm-based estimation procedure, including encoding and decoding genetic chromosomes, evaluating the fitness of each chromosome, and reproducing child chromosomes by selection, crossover and mutation, are described in the following.

3.2.1. Encoding and decoding chromosomes

To apply GA to a given problem, a suitable encoding scheme for the parameter set must first be determined. Of the various encoding methods that have been proposed, the most popular representation structures are binary vector and floating vector (Gen and Cheng, 2000). Because of its ease of implementation, a binary encoding method was applied to represent the parameter set in this research.

In binary encoding, each decision variable is represented by a binary substring, and these substrings are concatenated to form a longer string, i.e., the chromosome, which represents the set of decision variables. The length of the substring is determined by the range of the decision variable and the level of precision. Let x be the real-valued decision variable and let it have a domain $[x_{\min}, x_{\max}]$. The length of the binary string used to represent x can be obtained from:

$$L = \mathrm{INT}\left(\log_2 \left(\frac{x_{\max} - x_{\min}}{D} + 1 \right) + 1 \right) \tag{19}$$

where L is the length of binary string, D is the desired precision of variable x, and INT is the truncate operation to convert a real number into an integer. Then the decoding from a binary string to a real variable x is computed as:

$$x = x_{\min} + (x_{\max} - x_{\min}) * \frac{A}{2^L - 1} \tag{20}$$

where A is the value of binary string base 10.

To identify $x_{\min}$, $x_{\max}$ and D, a preliminary analysis and understanding of variable x is required. Specifically, if variable x is very sensitive, the desired precision needs to be higher. As a result, a longer string would be used in the representation scheme for variable x.

In our study, the parameter sets that need to be estimated are $\{b^{\mathrm{T}}, W^{\mathrm{T}}, b^{\mathrm{V}}, W^{\mathrm{V}}, b^{\mathrm{C}}, W^{\mathrm{C}}\} \subset \boldsymbol{\Theta}$, which contain six decision variables. Each parameter is represented by 6 binary bits, so the chromosome which represents the parameter set will have 36 binary bits. An example of genetic representation is shown in Fig. 2.

100010	001001	010101	100001	000001	111000
b^T	W^T	b^V	W^V	b^C	W^C

Fig. 2. An example of genetic representation of six decision variables with 36 bits.

3.2.2. Fitness function

A fitness function is required to measure the "goodness" of each chromosome. The fitness function used here is the linear scaling of the objective function in the minimization program of Eq. (18), which represents how well the estimate volume $\widehat{\widehat{Q}}_i(t)$ and its true (observed) value $\widehat{Q}_i(t)$ match, as shown in Eq. (21):

$$f_k = a \sum_{t=1}^{t=\frac{t_2-t_1}{30}} \sum_{\forall i} \left[\widehat{Q}_i(t) - \widehat{\widehat{Q}}_i(t) \right]^2 + b \tag{21}$$

where f_k is the fitness value of the kth chromosome and a and b are the scaling factors. Linear scaling is introduced to avoid two significant difficulties in the fitness proportionate selection process: premature convergence termination at early generations, and stalling at late generations (Goldberg, 1989). Parameters a and b are selected so that the average fitness is mapped to itself and the best fitness is increased by a designed multiple of the average fitness.

3.2.3. Selection process

Selection is an operation through which chromosomes are picked for reproduction with a probability proportional to their fitness. In this study, a combination of fitness proportionate selection and elitism strategy is adopted for the reproduction process. In the fitness proportionate selection, also called roulette wheel selection, the probability that a chromosome will be selected is given by the ratio of its fitness to the fitness of the entire population, as shown in the following:

$$q_k = \frac{f_k}{\sum_{m=1}^{pop\_size} f_m}, \tag{22}$$

where q_k is the probability of selecting chromosome k to produce offspring, f_k is the fitness value of the kth chromosome in the current generation and *pop_size* is the population size.

The elitism strategy keeps a certain number of the top chromosomes that have the highest fitness values and propagates to the next generation. This procedure ensures that the best solution in the next generation is not worse than the one in the current generation.

3.2.4. Crossover operation

The crossover operator produces two new offspring from two parent strings, by copying selected bits from each parent. The bit at position i in each offspring is copied from the bit at position i in one of the two parents. The choice of which parent contributes the bit for position i is determined by an additional string called the crossover mask. With the crossover mask, crossover operation can be performed at single-point or two-point, as illustrated in Fig. 3. In our study, both single-point crossover and two-point crossover are used.

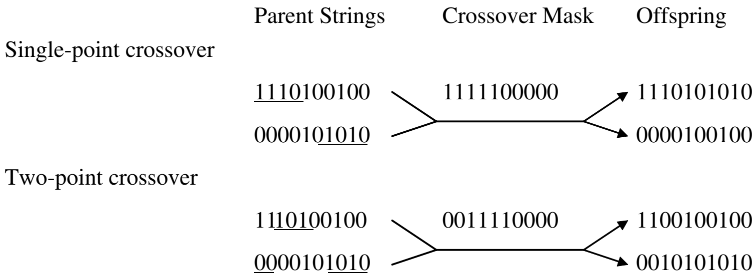

Fig. 3. Illustration of crossover operation (10 bits string is used as an example).

3.2.5. Mutation operation

Mutation is a process to overcome the local optimum problem. Each bit of a selected string is allowed to mutate according to a predetermined mutation probability, thereby reducing the likelihood that the search process will get stuck in a local optimum.

4. Empirical data collection and processing

4.1. Study site

We apply the proposed method to newly collected data concerning route choice in the California State Route 91 value-pricing project. The SR91 toll lanes, located between the SR91/SR55 junction in Anaheim, CA and the Orange/Riverside County Line, are the world's first fully automated privately operated toll lanes (Sullivan, 2000). The express lanes extend about 10 miles along the former median of the Riverside Freeway (SR91), connecting rapidly growing residential areas in Riverside and San Bernardino Counties to job centers in Orange and Los Angeles Counties to the west. SR91 in eastern Orange County includes four regular freeway lanes (91F) and two express lanes (91X) in each direction. Motorists who wish to use express lanes must register and carry identifying electronic transponder (the so-called *FasTrak*) to pay a toll that varies hourly according to a preset schedule. Tolls in the express lanes vary hour-by-hour to control demand and maintain free flow traffic, in contrast to often congested traffic conditions in the adjacent free lanes. Tolls on westbound traffic during morning commute hours ranged from \$1.65 (at 4–5 a.m.) to \$3.30 (at 7–8 a.m., Monday–Thursday). Within the SR 91 corridor, the Eastern Toll Road (ETR) competes with the 91X for trips to Irvine and vicinity. However, since the 91X has no entrance or exit between its starting and ending points, ETR users must use the highly congested 91F for access.

We regard the SR 91 toll road portion as a two-route network. One route is the 91X, and the other is 91F, both 10 miles in length. This gives motorists the option to travel free on regular roads or to pay a time-varying price for congestion-free express travel on a limited part of their journey. Because of the toll pricing structure, observation is that traffic consistently moves at a free-flow speed of approximately 75 mph even during peak hours on this facility. Consequently, we assume that travel time on 91X is deterministic (reliable) and equal to 8 min, corresponding to a speed of 75 mph. However, the free lanes are often congested during the morning peak hours

(5–9 a.m.), and travel time on the 91F is rather stochastic and unreliable, presenting a relatively "clean" real-world experimental environment to study the relative contributions of travel time and travel-time reliability in the route choice decision process.

4.2. Travel time and reliability data collection and processing

Obtaining accurate measures of travel conditions, especially the appropriate measurement of travel time reliability, is a formidable task. We use actual field measurements (floating cars) of travel time on 91F taken at different times during morning peak period. Our data was obtained from the study of Small et al. (2002). The data consist of peak period travel time on 91F for 11 days: first on October 28, 1999, and then on July 10–14 and September 18–22, 2000. Data were collected from 4–10 a.m. on each day, and include a total of 210 observations of travel time along the 10-miles stretch of 91F at different times of day encompassing the morning peak period. Interested readers may refer to their paper for more details on the travel time data collection and processing techniques.

In order to construct measures of travel time and its reliability, we consider both the central tendency and the dispersion of the travel time distribution. Measures of central tendency include the mean and the median, and measures of dispersion include the standard deviation, the inter-quartile difference such as the 90th–50th or 80th–50th, ratio of standard deviation to mean, and percent of observations that exceed the mean by some specific threshold, etc. The nature of these measures is that they are positive, monotonically increasing functions of variability. We assume that motorists, especially commuters in the morning peak hours, are concerned with the probability of significant delay, and are likely to pay particular attention to the upper tail of the distribution of travel times. Among the candidate measures that capture this effect, we then use the difference between the upper quartile and the median. To make our results comparable to Small et al. (2002), we use the same measures of central tendency and dispersion, i.e., median and the 80th–50th percentile differences.

Fig. 4 shows the raw field observations of travel time savings (i.e., the difference between the 91F and 91X travel times over the 10-miles stretch). The non-parametric estimates of mean, median, and 80th percentile are calculated and displayed. Median time savings reach a peak of 5.6 min around 7:15 a.m. Fig. 5 shows the median travel time savings and the 80th–50th percentile differences. The latter reaches a peak around 8:10 a.m. Correlations between these two measures are insignificant.

4.3. Traffic volume data collection and processing

Along the stretch of SR 91 under consideration, loop detector stations are spaced at a distance of every mile. Each loop detector station includes 6 loops covering all lanes of 91F and 91X. Volume data were collected using 30-s loop detector data and aggregated into 5-min interval. The data consist of volumes on 91X, 91F, and ETR for 30 weekdays from September 17 to November 16, 2001. Since our study is concerned with the traveler's choice probability between 91X and 91F, the volume data of ETR was subtracted from 91F because ETR users have no option but to use 91F. Fig. 6 shows the traffic flow on both 91X and 91F. As shown in Fig. 7, the percentage of travelers taking 91X reaches a peak at around 8:00 a.m.

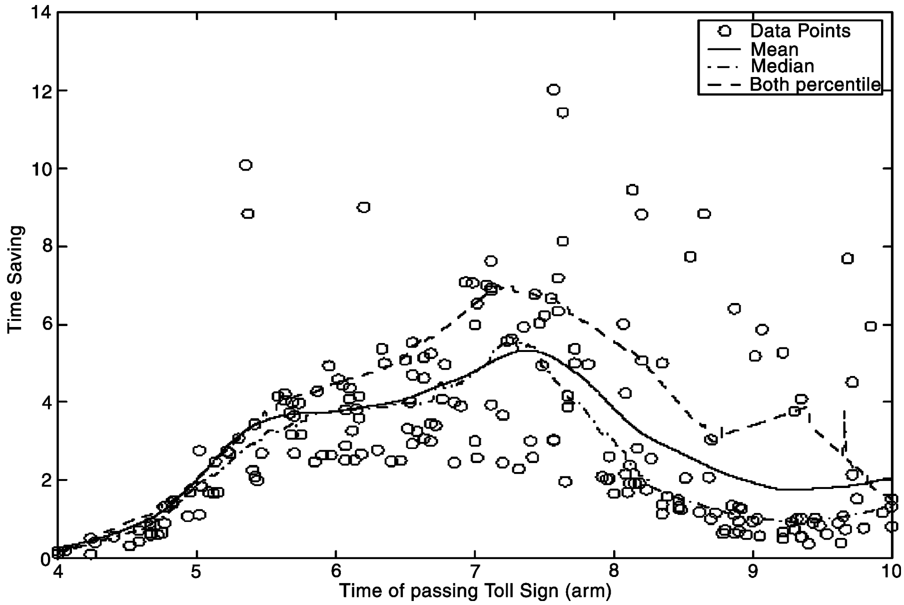

Fig. 4. Travel time saving (from the study of Small et al. (2002)).

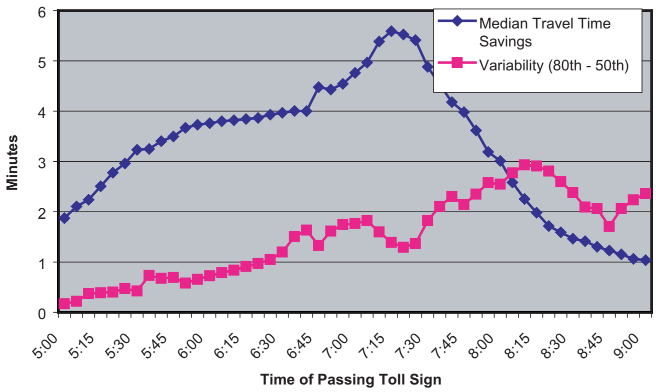

Fig. 5. Measurements of travel time and its variability.

5. Results and analysis

After data collection and processing, we applied the GA algorithm to identify the parameter set in the mixed-logit route choice model that produced the best match to the volume data revealed from the loop detectors. The parameters used in the GA algorithms are shown in Table 1. To evaluate each chromosome, the choice probability estimated from mixed-logit was calculated using 2000 random draws from a normal distribution of the components of $\boldsymbol{\beta}$ for the Monte-Carlo simulations. The convergence of the GA is shown in Fig. 8, which displays the decreasing mean absolute error ratio (MAER) values with the number of generations.

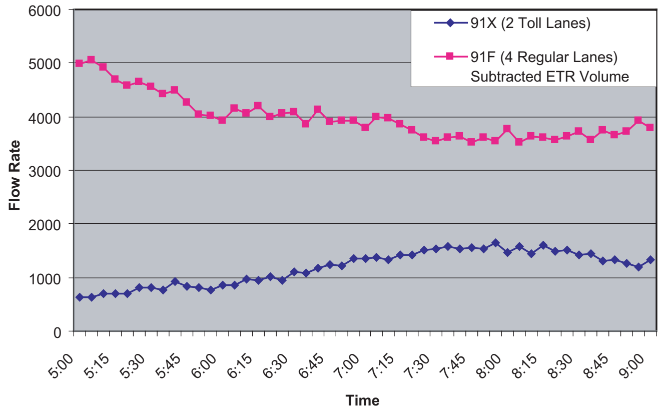

Fig. 6. Traffic flow on SR91 (September 18, 2001).

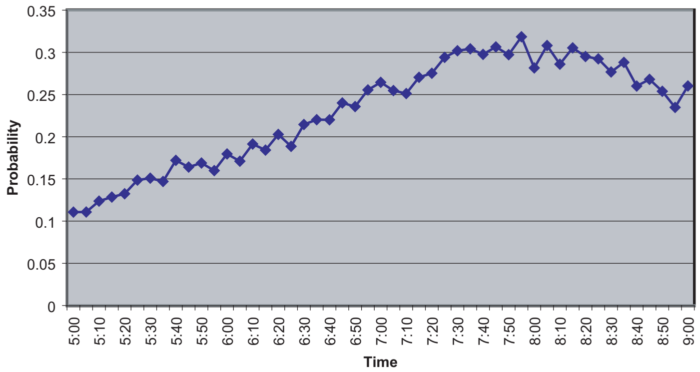

Fig. 7. Percentage of travelers taking toll lanes from loop detector data (September 18, 2001).

Table 1
Control parameters for the GA

Population size	20
Maximal number of generations	50
Substring length per parameter	6
Crossover probability	0.6
Mutation probability	0.033
Elitism flag	1

To estimate the parameters Θ with certain confidence level, we performed 30 GA runs with the volume data from 30 days, with each GA run corresponding to the volume data from one particular day. Fig. 9 shows the identified values for parameters Θ with 30 GA runs (The standard deviation Sd is given instead of variance W). The statistical estimates of parameters Θ from these 30 GA runs are shown in Table 2.

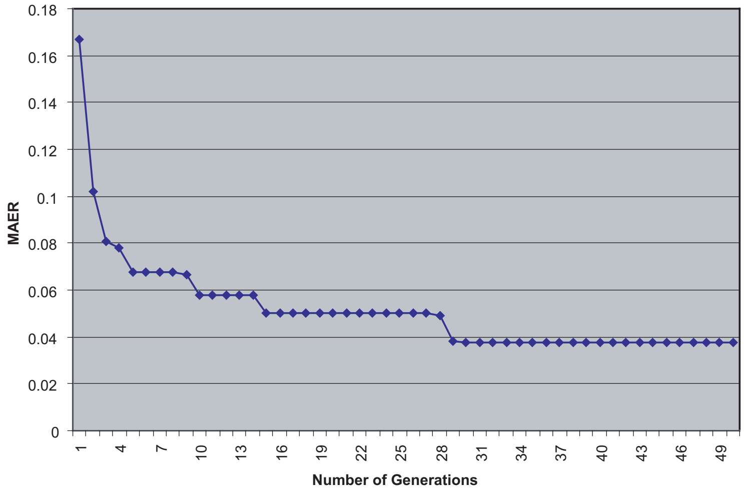

Fig. 8. GA convergence curve with best fitness values.

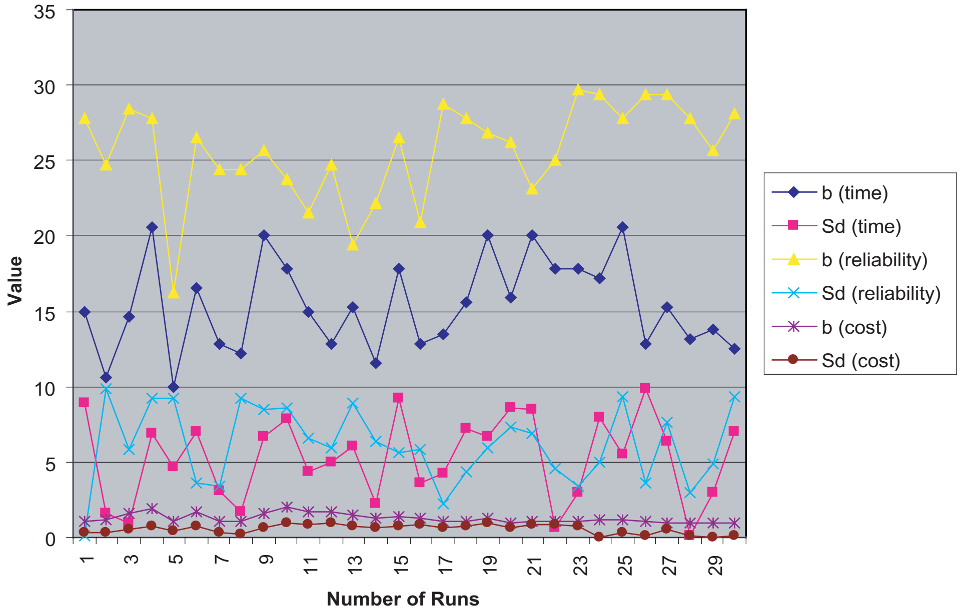

Fig. 9. Best estimated parameters with 30 GA runs.

From the estimated statistical distributions of the parameter estimates, travelers' implied VOT, VOR and DORA, and the extent of their heterogeneity can be determined by Monte-Carlo simulations. We performed 2000 random draws from normal distributions of $\boldsymbol{\beta}$ ($\beta^{\mathrm{T}} \sim N(b^{\mathrm{T}}, W^{\mathrm{T}})$, $\beta^{\mathrm{V}} \sim N(b^{\mathrm{V}}, W^{\mathrm{V}})$, $\beta^{\mathrm{C}} \sim N(b^{\mathrm{C}}, W^{\mathrm{C}})$), and calculated VOT, VOR and DORA using Eqs. (2)–(4). Percentile values, including 25%-ile, 50%-ile (median), and 75%-ile, were then obtained from the

Table 2
Estimated parameter values

Parameters	Predefined range	Estimated value (median/[5%-ile, 95%-ile])
b^T	$[10, 40]$	$15.39/[11.05, 20.34]$
Sd^T	$[0, 10]$	$5.28/[0.77, 9.07]$
b^V	$[10, 40]$	$25.66/[20.08, 29.38]$
Sd^V	$[0, 10]$	$6.14/[2.54, 9.38]$
b^C	$[1, 2]$	$1.26/[1.0, 1.81]$
Sd^C	$[0, 1]$	$0.55/[0.04, 0.94]$

Table 3
Estimated value of time and reliability

	Median estimate	90% confidence interval [5%-ile, 95%-ile]
Value of time ($/h)		
Median	12.81	$[8.66, 16.08]$
Heterogeneity (75th–25th)	8.72	$[4.16, 16.73]$
Value of reliability ($/h)		
Median	20.63	$[12.81, 28.47]$
Heterogeneity (75th–25th)	13.06	$[5.62, 23.44]$
Degree of risk aversion (DORA)		
Median	1.73	$[1.26, 2.13]$
Heterogeneity (75th–25th)	0.97	$[0.41, 1.75]$

2000 values of VOT, VOR and DORA. Travelers' heterogeneity is measured as the inter-quartile difference, i.e., the difference between the 75th and 25th percentile values, because it is unaffected by high upper-tail values occasionally found in the calculation of ratios. We repeated this process for every parameter set $\{b^T, W^T, b^V, W^V, b^C, W^C\}$ identified by the 30 GA runs. The estimates of the median and heterogeneity of VOT, VOR, and DORA are shown in Table 3. In Table 3, we note that the confidence interval represents uncertainty due to statistical error, not heterogeneity. A positive 5th percentile value means the quantity is significantly greater than zero according to a conventional one-sided hypothesis test at a 5% significance level.

As shown in Table 3, the median value of time is $12.81, and the median value of reliability is $20.63. Since median time savings in our data peaks at 5.6 min in the rush hour and unreliability peaks at 3 min, the average commuter would pay $1.20 to realize time savings and pay $1.03 to avoid this possibility of unanticipated delay. In other words, travelers with the median VOT and VOR would save $2.23 from travel time and its reliability if they use 91X, but they need to pay $3.30 for the toll. So less than half travelers will choose to use the express lanes, and this is confirmed from the loop detector data, as shown in Fig. 7. Regarding travelers' heterogeneity towards travel time and its reliability, both measures of heterogeneity in the cases of VOT and VOR are more than 60% of their median values, indicating that commuters exhibit a wide distribution of preferences for speed and reliability. By recognizing the heterogeneity in travelers' preference and offering choices that caters to their preferences, road pricing policies can increase transportation efficiency.

Similar results for the median values of VOT and VOR are found in the study of Small et al. (2002). In their study, they estimated VOT and VOR using the combination of RP and SP data. The median value of VOT estimated from RP data is \$20.20, and \$9.46 from SP data. Our result in this regard falls between these two values. In terms of VOR, the estimated median value from their study using RP data is \$19.56, which is very similar with our result. Their estimate of VOR from SP data is not comparable to ours since different measures of (un)reliability are used. Our results validate the analysis from their study and demonstrate the applicability of an approach based on conventional loop data in the study of travelers' behavior.

Table 3 also shows the estimated median value of DORA is 1.73, i.e., the disutility caused by certain amount of travel time unreliability is 1.73 times more than that caused by travel time of the same amount. For instance, assuming the commuting alternative is a 20-min commute with essentially no possibility of significant delays and a commuting alternative that normally takes 10 min but has a variability about 6 min. If the traveler has a DORA equal to 1.73, he/she will be almost indifferent between these two choices ($10 + 1.73 * 6 \approx 20$). A traveler with a DORA greater than 1.73 is more risk averse and will choose the first alternative. In other words, travelers with DORA greater than 1.0 value more highly a reduction in variability than a comparable reduction in the travel time. These travelers are willing to go out of their way to decrease the possibility of a delay, either because they dislike the risk of being delayed or dislike the discomfort normally associated with a delay, such as stop and go traffic. From the operator's point of view, this finding implies that traffic management strategies aimed at reducing travel time variability, such as incident management, deserve serious attention.

6. Conclusions

It is generally accepted that travel time reliability can have significant influence on traveler's route choice behavior and that it cannot be ignored in any model which purports to predict behavior or provide a basis for evaluation. With respect to the valuation of reliability, such direct methods as revealed preference surveys and/or state preference surveys have been used extensively in previous studies. In this research, we proposed an indirect method to study the contribution of travel time reliability in traveler's route choice behavior. We formulated traveler's route choice as a mixed-logit model, with the coefficients in the model representing individual traveler's preferences or tastes to travel time, reliability and cost. Unlike the traditional approach to estimate these coefficients with RP and/or SP data by simulated maximum likelihood estimation, we adopt genetic algorithm to identify the coefficients that enable the flows resulting from route choice model to best match the time-dependent traffic volume data obtained from loop detectors. Such an approach eliminates both the cost and biases inherent to RP and SP survey techniques.

We applied the proposed method to newly collected data concerning route choice in the California State Route 91 value-pricing project. Based on travelers' choice of whether or not to pay a congestion-based toll in order to use express lanes, we are able to estimate how travelers value travel time and travel-time reliability. We find that the estimated median value of travel-time reliability is substantially greater than that of travel-time, and the median value of degree of risk aversion is significantly greater than 1, indicating that travelers value more highly a reduction in variability than in the travel time saving for that journey. Moreover, travelers' attitude towards congestion is not homogeneous; in fact, substantial heterogeneity exists in travelers' preference of travel time and reliability. The results of our study yield important insights into commuters' route choice in general and the tradeoffs among travel time, reliability, and monetary cost. Our results validate the analysis from some previous studies and demonstrate the applicability of the approach in the study of travelers' behavior.

Acknowledgements

The authors thank Professor Ken Small and Dr. Jia Yan of the University of California, Irvine for providing travel time data used in this work. This study was supported, in part, by the California PATH Program; their support is gratefully acknowledged. The views expressed herein are the authors' own.

References

Abdel-Aty, M., Kitamura, R., Jovanis, P., 1995. Investigating effect of travel time variability on route choice using repeated measurement stated preference data. Transportation Research Record 1493, 39–45.

Bates, J., Polak, J., Jones, P., Cook, A., 2001. The valuation of reliability for personal travel. Transportation Research Part E 37, 191–229.

Bhat, C.R., 2001. Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. Transportation Research Part B 35, 677–693.

Bhat, C.R., Castelar, R., 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, 593–616.

Gen, M., Cheng, R., 2000. Genetic Algorithms and Engineering Optimization. John Wiley & Sons, Inc., New York.

Goldberg, D., 1989. Genetic algorithm in search, optimization and machine learning. Addison-Wesley, Reading, MA.

Holland, J.H., 1975. Adaptation in Natural and Artificial Systems, University of Michigan Press (reprinted in 1992 by MIT Press, Cambridge, MA).

Jackson, W.B., Jucker, J.V., 1981. An empirical study of travel time variability and travel choice behavior. Transportation Science 16 (4), 460–475.

Lam, T.C., Small, K.A., 2001. The value of time and reliability: measurement from a value pricing experiment. Transportation Research Part E 37, 231–251.

McFadden, D., Train, K., 2000. Mixed MNL models for discrete response. Journal of Applied Econometrics 15, 447–470.

Small, K.A., Winston, C., Yan, J., 2002. Uncovering the distribution of motorists' preferences. Working paper 01-02-10, Department of Economics, University of California, Irvine.

Sullivan, E., 2000. Continuation study to evaluate the impacts of the SR91 value-priced express lanes, Final Report Submitted to the Department of Transportation of the State of California. See http://gridlock.calpoly.edu/sr91/sr91.htm.

Train, K., 2002. Discrete Choice Methods with Simulation. Cambridge University Press.