Choice Models with Stochastic Variables and Random Coefficients

Mehek Biswas
Department of Civil Engineering
Indian Institute of Science (IISc)
Bengaluru 560012, India
Tel: +91-80-2293-2043 s
Email: mehekbiswas@iisc.ac.in

Abdul R. Pinjari ***(Corresponding author)***
Department of Civil Engineering
Centre for infrastructure, Sustainable Transportation, and Urban Planning (C*i*STUP)
Indian Institute of Science (IISc)
Bengaluru 560012, India
Tel: +91-80-2293-2043
Email: abdul@iisc.ac.in

Chandra R. Bhat
The University of Texas at Austin
Department of Civil, Architectural and Environmental Engineering
301 E. Dean Keeton St. Stop C1761
Austin, TX 78712, USA
Tel: 1-512-471-4535
Email: bhat@mail.utexas.edu

Sulagna Ghosh
Former M. Stat. Student
Indian Statistical Institute
Kolkata 700108, India
Email: ghoshsulagna97@gmail.com

ABSTRACT

In travel choice models, variables describing alternative attributes such as travel time can be inherently stochastic due to variability in travel conditions. Such variability in stochastic variables is different from unobserved heterogeneity in travellers' response to those variables. Specifying only one of these as random while keeping the other fixed can potentially result in biased parameter estimates, inferior goodness-of-fit, and distorted information for policy and welfare analyses. Simultaneous identification of these two separate sources of variability is important for understanding travel behaviour in complex environments characterized by variable travel conditions. Therefore, in this study, we propose an integrated choice and stochastic variable modelling framework with random coefficients (i.e., an *ICSV-RC* framework) that allows the analyst to simultaneously accommodate stochasticity in explanatory variables and random coefficients on such variables. In addition, we show that ignoring either source of stochasticity – stochasticity in alternative attributes or unobserved heterogeneity in response to the attributes – results in models with inferior goodness-of-fit and a systematic bias toward zero in all parameter estimates. We demonstrate this using simulation experiments for two different travel choice settings, one involving labelled mode choice alternatives and the other involving unlabelled route choice alternatives. In addition, we present an empirical analysis in the context of route choice of trucks to highlight the importance of accommodating both sources of variability – stochasticity in travel times and random heterogeneity in response to travel times.

Keywords: discrete choice, stochastic variables, random coefficients, identification, integrated choice and latent variable (ICLV) models

1 INTRODUCTION

In random utility maximization (RUM)-based travel choice models, it is common to assume that exogenous variables entering the utility functions are deterministic. In many situations, however, it may be more appropriate to specify the exogenous variables entering the utility functions as stochastic. This is due to at least three reasons: (1) inherent stochasticity in the variables, (2) analyst's errors in measuring the true value of the variables, and (3) travellers' perceptions of the values of the variables (that may be different from the measurements that the analyst may possess). Each of these sources of stochastic variability is briefly discussed next in the specific context of travel time, which is a level-of-service variable used in travel choice models such as those for mode choice, route choice, and departure time choice.

The first reason for travel time stochasticity may be attributed to day-to-day and intra-day variability in travel conditions on the network (Chen *et al*., 2011; Srinivasan *et al*., 2014; Biswas *et al*., 2019). The resulting travel time distribution has been studied extensively in the existing literature (Rakha *et al*., 2006; Srinivasan *et al*., 2014; Polus, 1979; Al-Deek and Emam, 2006; Taylor, 2012; Aron *et al*., 2014; Guo *et al*., 2012; Zang *et al*., 2018, etc.). The second reason for travel time stochasticity may be attributed to measurement errors by the analyst (Bhatta and Larsen, 2011*;* Ortúzar and Ivelic, 1987; Train, 1978). Such errors may arise due to: (a) the use of spatially aggregate (zone-to-zone) measures of level-of-service attributes instead of disaggregate (point-to-point) measurements, and (b) calculating travel times based on free flow speed assumptions instead of measuring actual travel times or speeds. The third reason for travel time stochasticity may originate from travellers perceiving travel times to be different from what may be objective travel times and the errors in travellers' perceptions (Daly and Ortúzar, 1990).

As discussed in Diaz *et al*. (2015) and Ortúzar and Willumsen (2011), ignoring any of the above-mentioned sources of stochasticity, if present, will, in general, lead to biased parameter estimates and distorted marginal rates of substitution (e.g., willingness to pay) during estimation. Important to note also is that some or all of the above-discussed stochasticity sources, while invoked in the specific context of travel time, can also apply to several other exogenous variables used in travel choice models. For example, crowding levels in transit modes can be stochastic in mode choice settings because of day-to-day variability and measurement errors. Travel costs perceived by travellers may be different from the actual costs they pay due to the different time

scales in which the different costs occur (e.g., fuel costs are paid regularly, whereas maintenance and insurance costs are paid once in a few months or a year).

Among the approaches used to accommodate stochasticity in variables in discrete choice models is the classic errors-in-variables (EIV) approach widely used in regression models (Carroll and Spiegelman, 1984; Stefansky and Carroll, 1985; Durbin, 1954; Gleser, 1981, etc.). Some choice modelling studies have adopted the EIV method through Rubin's multiple imputation (Rubin, 1987) for cases when data on exogenous variables are missing or unknown beyond certain interval bounds, such as for travel time (Steinmetz and Brownstone, 2005). Alternatively, studies such as Conniffe and O'Neil (2008) propose analytic expressions for estimators in the presence of missing data. Diaz *et al.* (2015) use the mixed logit approach to specify errors in variables as additional error components in the utility functions, while keeping the coefficients on those variables as deterministic. A second method is the Integrated Choice and Latent Variable (ICLV) modelling approach (Ben Akiva *et al*., 2002; Alvarez-Daziano and Bolduc, 2013; Bhat and Dubey, 2014; Vij and Walker, 2016). As the name suggests, this approach allows the explanatory variables in a choice model as latent and stochastic. Doing so helps in recognizing measurement errors in variables (Walker *et al*., 2010), perception errors by individuals (Varotto *et al*., 2017), and even missing data (Sanko *et al*., 2014). Recently, Biswas *et al*. (2019) demonstrated the use of this approach to accommodate inherent stochasticity (due to day-to-day variability) in the travel time variables used in choice models.

In a separate and rather large stream of literature, unobserved taste heterogeneity of individuals (that is, variations in the sensitivity to exogenous variables due to unobserved factors) has been modelled in a multitude of choice contexts. These studies specify random coefficients on alternative attributes through frameworks such as the mixed multinomial logit (Bhat, 2001; Bhat, 2003; Hensher and Greene, 2003; Greene and Hensher, 2003; Batley *et al*., 2004; Hess and Polak, 2005; Mc Fadden and Train, 2000; Revelt and Train, 1998; Brownstone *et al*., 2000; Swait, 2022) and the mixed multinomial probit (Bhat, 2011; Bhat and Sidharthan, 2012; Patil *et al*., 2017; Dubey *et al*., 2022). Regardless of the approach used to accommodate unobserved heterogeneity in response to exogenous variables, this stream of literature does not consider stochasticity in the exogenous variables themselves. In fact, to the best of our knowledge, no study has attempted to simultaneously recognize and disentangle the two sources of variability – stochasticity in explanatory variables and unobserved heterogeneity in response to those variables. This is because

typical mixed logit/probit and ICLV model formulations do not allow the simultaneous estimability or identifiability of both sources of variability.

The objective of the current research is to formulate a choice modelling framework that allows the analyst to simultaneously accommodate stochasticity in explanatory variables and random coefficients on such variables. In addition, the study aims at applying the proposed framework to disentangle travel time variability from unobserved heterogeneity in response to travel time in travel choice models. To this end, we formulate an integrated choice and stochastic variable (ICSV) modelling framework with random coefficients (RC) in its choice model.[1] We show that the proposed *ICSV-RC* framework allows the simultaneous identification of stochasticity in travel time as well as random heterogeneity in response to travel time – due to its ability to bring together two (or more) different data sources such as travel time measurements and traveller choices. In addition, we show that ignoring either source of stochasticity – variability in travel time or heterogeneity in response to travel time – results in models with inferior fit to data and a systematic bias toward zero in all parameter estimates. Using simulation experiments, we demonstrate this in two distinct choice settings – one involving labelled mode choice alternatives and the other involving unlabelled route choice alternatives.

While our proposed formulation is generic and applicable to any choice context, the empirical application in this study pertains to truck route choice and travel time measurements derived from a large truck-GPS dataset in Florida. Using this empirical data, we demonstrate the applicability of the *ICSV-RC* framework for simultaneously identifying variability in travel time and a random coefficient on travel time. We then compare the proposed model with simpler versions of it – one without random coefficients and one without variability in travel time – to highlight the importance of accounting for both sources of variability.

[1] In the field of choice modelling, the incorporation of latent psychological constructs such as attitudes and perceptions as explanatory variables within the random utility maximization framework assumes the form of a hybrid model that is typically referred to as the Integrated Choice and Latent Variable (ICLV) model (for further reading, one may refer to Ben-Akiva *et al*. (2002), Bhat and Dubey (2014), Alvarez-Daziano and Bolduc (2013) and Vij and Walker (2016)). In the ICLV framework, stochastic variables are typically used to represent latent psychological constructs such as attitudes and perceptions of the individuals making choices. In this paper, we use the more general label of Integrated Choice and "Stochastic" variable (ICSV) framework to recognize that individual latent attitudes/perceptions are but only one form of a stochastic variable within an integrated choice context; the stochasticity in the variable can also derive from inherent variability or measurement error or other sources.

The rest of this paper is structured as follows. Section 2 discusses the proposed *ICSV-RC* modelling framework in the context of an integrated model of traveller choice and stochastic travel time. Section 3 discusses the reason for bias in parameter estimates (and the direction of bias) in models that do not incorporate variability in stochastic variables. Section 4 presents simulation experiments. Section 5 presents the empirical results and findings in the context of truck route choice in Florida, USA. Section 6 concludes the study and identifies directions for future research.

2 MODEL FRAMEWORK

2.1. Model Formulation

The proposed *ICSV-RC* framework is formulated to jointly model the distribution of travel time for the choice alternatives available to the traveller and the traveller's choice[2] in a setting characterized by stochastic travel times. Simultaneous to the estimation of the travel time distributions, the stochastic travel time variable is used as an explanatory variable in a mixed multinomial logit-based model of traveller choice (route choice or mode choice) with a random coefficient specified on it. Note that the distributional forms of both travel time and its coefficient are assumed to be known *a priori*, but the parameters of those distributions must be estimated.

Here, we present the notational preliminaries for the proposed model. Denote $\mathbf{J} = \{1,2, \dots, i, \dots, j, \dots J\}$ as the set of all alternatives available to a traveller, where J is the total number of alternatives available to a traveller. In a route-choice setting **J** represents route alternatives, and, in a mode-choice setting **J** represents travel mode alternatives. For each such alternative i, we define a set of travel time measurements, $\mathbf{M_i} = \{OTT_{i1}, OTT_{i2}, \dots, OTT_{im}, \dots OTT_{iM_i}\}$, where OTT_{im} is the m^{th} measurement (or observation) of travel time associated with alternative i. The number of measurements $|\mathbf{M_i}|$ may vary across alternatives, while some alternatives may not have any measurements available. Further, define y_i as an indicator whether alternative i was chosen by the individual (or by the trip); y_i assumes the value 1 if the alternative *i* is chosen and is zero otherwise. Note that we suppress the index for the decision-maker (or traveller) for ease in developing the notation for the structural and measurement equations of the proposed framework.

[2] We refer to route choice of trips and mode choice of an individual by the common term *traveller's choice*. Such terminology is adopted to view the framework as a tool relevant to a broad range of choice settings. Further, in the same vein, the decision-maker is referred to as the *traveller*.

2.1.1 Structural equations for the ICSV-RC model

Define the utility (U_i) associated with choosing an alternative *i* as:

$$U_i = \gamma_{TT^*} TT_i^* + \boldsymbol{\theta}' \boldsymbol{x_i} + \varepsilon_i \quad (1)$$

In the above equation, TT_i^* is the stochastic travel time variable for alternative *i* and γ_{TT^*} is its coefficient (which is specified as a random parameter). In the current study, we assume $\gamma_{TT^*} = \mu_{\gamma_{TT^*}} + \sigma_{\gamma_{TT^*}} z$; $z \sim N(0,1)$, albeit one can explore several other distributions such as lognormal or truncated normal. Further, $\boldsymbol{x_i}$ is a *W*x1 vector of other, deterministic attributes of alternative *i* and $\boldsymbol{\theta}$ is the corresponding *W*x1 vector of coefficients (some or all of which may be assumed as random parameters, as in the case of typical mixed multinomial choice models); and ε_i is a standard Gumbel distributed error term assumed to be independent and identically distributed (IID) across all choice alternatives and travellers[3].

Note that the stochastic travel time variable (TT_i^*) can be specified using different functional forms depending on the model setup. The simplest approach is to assume that travel time for any trip follows an *a priori* distribution with the same parameters. For example, Walker *et al*. (2010) represent mode-specific travel times in a mode choice model using a normal distribution whose parameters are estimated using available measurements of travel times. A second approach is to specify the mode-specific travel time distribution as a function of mode-specific inverse speed (i.e., the time it takes to travel unit distance) and travel distance. That is, for a mode *i*, the travel time distribution may be expressed as below:

$$TT_i^* = \theta_i d_i \quad (2)$$

In the above equation, θ_i can be interpreted as the inverse speed of mode *i*, where d_i denotes the travel distance on mode *i* between the origin and destination of the trip. Here, θ_i may be considered random to allow variability in inverse speeds (and travel times) for mode *i*.

A third approach, applicable in route choice models, is to represent the travel time distribution using a structural equation that specifies route-level travel time as a function of the

[3] Besides other exogenous variables in $\boldsymbol{x_i}$ although not discussed in Equation (1), the empirical specification in route choice models typically includes error components to capture inter-route correlations (Frejinger and Bierlaire, 2007). While we do not expand the model specification here to include the error components (for simplicity in the notation), we do include error components in both our simulation experiments and the empirical application for route choice.

underlying route structure, as in Equation (3). Doing so helps capture stochasticity due to variability in travel conditions (and allocating variability to different elements) on the network.

$$TT_i^* = \sum_{l=1}^{L} \beta_l d_{il} + \sum_{q=1}^{Q} \gamma_q n_{iq} \quad (3)$$

Here, d_{il} denotes the length of roadway links of type *l* on route *i* (length of interstates, length of arterials, length of local roads, etc.), n_{iq} denotes the number of nodes of type *q* on route *i* (no. of left turns, no. of right turns, etc.) and *L* and *Q* denote the total number of roadway link types and stop types, respectively. Further, β_l is the coefficient on d_{il}, which may be interpreted as the inverse speed on roadway link of type *l* (i.e., the time it takes to traverse unit length of a roadway of type *l*) on route *i*. β_l is considered random to allow variability in inverse speeds for a link of type *l*. γ_q is the coefficient on n_{iq}, which may be interpreted as the time it takes a vehicle to cross a node of type *q*. γ_q is considered random to allow for variability in node-crossing times. The random coefficients $\beta_l (l = 1, \dots, L)$ and $\gamma_q (q = 1, \dots, Q)$ help capture variability in TT_i^* due to variability in travel conditions. In vector form, the structural equation for TT_i^* may be written as:

$$TT_i^* = \boldsymbol{B}'\boldsymbol{D_i} + \boldsymbol{\Gamma}'\boldsymbol{N_i} \quad (4)$$

where, $\boldsymbol{D_i} = [\, d_{i1}, d_{i2}, \dots, d_{iL}]'$ is the vector of lengths of roadway links of each of *L* types on route *i* and $\boldsymbol{N_i} = [\, n_{i1}, n_{i2}, \dots, n_{iQ}]'$ is the vector of number of nodes of each of Q types.

2.1.2 Measurement equations for the ICSV-RC model

Equation (1) for the multinomial choice model involves both the stochastic travel time variable (TT_i^*) and its coefficient (γ_{TT^*}), which is also random. This necessitates two separate sources of data to identify each of these distributions through the proposed integrated model. The measurements used to identify taste heterogeneity of travellers (i.e., the distribution of γ_{TT^*}) are the observed choices (y_i) for each trip in a route choice setting or of each individual in a mode choice setting. Invoking the utility maximization theory, the measurement equation for the observed choices can be written as:

$$y_i = 1, \text{ if } U_i > U_j \forall\, J \in \mathbf{J}, j \neq i$$
$$= 0, \text{ otherwise} \quad (5)$$

The measurements used to identify travel time variability (that is, the distribution of the random parameters specified in Equation (2) or Equation (3) for TT_i^*) are the observed travel times (OTT_{im}) obtained from GPS data or other such data sources. Specifically, the travel time measurements (OTT_{im}) may be specified as realizations from the stochastic travel time function (TT_i^*) along with an additive noise term to represent measurement errors, as below:

$$OTT_{im} = TT_i^* + \xi_{im}, \forall\, m \in \mathbf{M_i} \tag{6}$$

Here, ξ_{im} is a noise term capturing the measurement error in OTT_{im} and assumed to be normally distributed, $\xi_{im} \sim N(0, \rho)$, with variance ρ to be estimated.

As discussed in Biswas *et al*. (2019), the stochastic travel time function in Equation (2), identified due to available travel time measurements for some observations in the data, can be simultaneously used to *impute* the travel time distribution for observations without travel time measurements. Doing so helps in utilizing partial measurement data, where travel time measurements may not be available for all observations or choice alternatives, for estimating the proposed integrated model.

2.2. Model System Estimation

Equation (1) along with the equation for the travel time function (Equation (2) or (4), depending on the functional form of the stochastic variable under consideration), and Equations (5) and (6) are brought together into an *ICSV-RC* framework for deriving the joint likelihood of travel time measurements and traveller choices in the observed data. Furthermore, distributional assumptions are made on the stochastic components of the formulation to derive the likelihood function for estimating model parameters.

In the *ICSV-RC* model, let $\boldsymbol{\Theta} = \{\mu_{\gamma_{TT^*}}, \sigma_{\gamma_{TT^*}}, \boldsymbol{\theta}, Vech(\boldsymbol{B}), Vech(\boldsymbol{\Gamma}), \rho\}$ denote the full set of parameters to be estimated in the integrated model system, where $Vech$ (.) is an operator used to represent the vector of the parameters inside the parentheses. For later use, define $\widetilde{\boldsymbol{\Theta}} = \{Vech(\boldsymbol{B}), Vech(\boldsymbol{\Gamma}), \rho\}$ and $\breve{\boldsymbol{\Theta}} = \{\mu_{\gamma_{TT^*}}, \sigma_{\gamma_{TT^*}}, \boldsymbol{\theta}, Vech(\boldsymbol{B}), Vech(\boldsymbol{\Gamma})\}$. Let $\boldsymbol{D}$ = [$\boldsymbol{D_1}'$,$\boldsymbol{D_2}'$,…,$\boldsymbol{D_J}'$]′, $\boldsymbol{N}$ = [$\boldsymbol{N_1}'$,$\boldsymbol{N_2}'$,…,$\boldsymbol{N_J}'$]′, and $\mathbf{X}$ = [$\boldsymbol{x_1}$,$\boldsymbol{x_2}$,…,$\boldsymbol{x_J}$]′; and let $\boldsymbol{TT}^*$= [TT_1^*,TT_2^*,…,TT_J^*]′ be the J x1 vector of stochastic travel times for all alternatives in the choice set. Next, recall that $\mathbf{M_i}$ is a vector of all travel time measurements for alternative *i*. Let $\boldsymbol{OTT}$ be a vector that stacks the $\mathbf{M_i'}$ vectors of all J alternatives into a column vector of size $\left[\sum_{i=1}^{J}|\mathbf{M_i}|\right]$ x 1. Next, to define the probability density of

the stochastic components, let $f_{jm}(.)$ be the PDF of OTT_{jm} given TT_j^*, let $g_j(.)$ be the PDF of TT_j^*, and let $h(.)$ be the PDF of γ_{TT^*} .

2.2.1 The ICSV-RC model

For the proposed model with a random coefficient specified on the stochastic travel time variable in the choice utility function, the likelihood that the i^{th} alternative is chosen along with the available observed measurements of route-level travel time is given by:

$$P\ (y_i=1, \boldsymbol{OTT} \mid \boldsymbol{D}, \boldsymbol{N}, \boldsymbol{X}, \boldsymbol{\Theta}) =$$

$$\int_{\gamma_{TT^*}} \int_{\boldsymbol{TT}^*} \frac{\exp(\gamma_{TT^*} TT_i^* + \boldsymbol{\theta}' \boldsymbol{x_i})}{\sum_{j=1}^{J} \exp\left(\gamma_{TT^*} TT_j^* + \boldsymbol{\theta}' \boldsymbol{x_j}\right)} \prod_{j=1}^{J} \prod_{m=1}^{M_i} f_{jm}(OTT_{jm} | TT_j^*, \boldsymbol{D}, \boldsymbol{N}, \widetilde{\boldsymbol{\Theta}}) \prod_{j=1}^{J} g_j(TT_j^* | \boldsymbol{D}, \boldsymbol{N}, \widetilde{\boldsymbol{\Theta}}) h(\gamma_{TT^*})\, d\boldsymbol{TT}^*\, d\gamma_{TT^*} \tag{7}$$

The dimensionality of integration in the above equation is the sum of the number of unique random parameters in U_i (except the IID Gumbel error terms). This includes the number of random coefficients in TT_i^*, one random coefficient for γ_{TT^*}, and any random coefficients for error components, etc. One can use the maximum simulated likelihood method to optimize the above likelihood function and estimate the parameters of the proposed model.

For observations without any measurements of travel time, Equation (7) becomes:

$$P\ (y_i=1 \mid \boldsymbol{D}, \boldsymbol{N}, \boldsymbol{X}, \boldsymbol{\Theta}) =$$

$$\int_{\gamma_{TT^*}} \int_{\boldsymbol{TT}^*} \frac{\exp(\gamma_{TT^*} TT_i^* + \boldsymbol{\theta}' \boldsymbol{x_i})}{\sum_{j=1}^{J} \exp\left(\gamma_{TT^*} TT_j^* + \boldsymbol{\theta}' \boldsymbol{x_j}\right)} \prod_{j=1}^{J} g_j(TT_j^* | \boldsymbol{D}, \boldsymbol{N}, \widetilde{\boldsymbol{\Theta}}) h(\gamma_{TT^*})\, d\boldsymbol{TT}^*\, d\gamma_{TT^*} \tag{8}$$

In this context, the travel time measurement data in Equation (7) helps estimate the parameters describing the distribution of $\boldsymbol{TT}^*$. These parameters are, in turn, used simultaneously in Equation (8) to inform or *impute* the stochastic travel time function for observations without travel time measurements. The same happens in Equation (7) as well, for observations with travel time measurements for some alternatives and no measurements for other alternatives.

2.2.2 Integrated model with stochastic travel time (ICSV model)

A restricted form of the proposed *ICSV-RC* model is obtained by specifying the coefficient γ_{TT^*} as fixed (denoted by $\bar{\gamma}_{TT^*}$ in the current model) instead of random. For this *ICSV* model, the likelihood expression in Equation (7) reduces to the following:

$$P(y_i{=}1, \boldsymbol{OTT} \mid \boldsymbol{D}, \boldsymbol{N}, \boldsymbol{X}, \boldsymbol{\Psi}) =$$

$$\int_{\boldsymbol{TT^*}} \frac{\exp(\bar{\gamma}_{TT^*} TT_i^* + \boldsymbol{\theta}' \boldsymbol{x_i})}{\sum_{j=1}^{J} \exp\left(\bar{\gamma}_{TT^*} TT_j^* + \boldsymbol{\theta}' \boldsymbol{x_j}\right)} \prod_{j=1}^{J} \prod_{m=1}^{M_i} f_{jm}(OTT_{jm} | TT_j^*, \boldsymbol{D}, \boldsymbol{N}, \widetilde{\boldsymbol{\Psi}}) \prod_{j=1}^{J} g_j(TT_j^* | \boldsymbol{D}, \boldsymbol{N}, \breve{\boldsymbol{\Psi}})\, d\boldsymbol{TT}^* \quad (9)$$

In the above expression, the full set of parameters to be estimated is denoted by $\boldsymbol{\Psi} = \{\bar{\gamma}_{TT^*}, \boldsymbol{\theta}, Vech(\boldsymbol{B}), Vech(\boldsymbol{\Gamma}), \rho\}$, where $\bar{\gamma}_{TT^*}$ is the deterministic coefficient on travel time, $\widetilde{\boldsymbol{\Psi}} = \{Vech(\boldsymbol{B}), Vech(\boldsymbol{\Gamma}), \rho\}$, and $\breve{\boldsymbol{\Psi}} = \{m_{\gamma_{TT^*}}, \boldsymbol{\theta}, Vech(\boldsymbol{B}), Vech(\boldsymbol{\Gamma})\}$. The notation for all other components remains the same as that of the vectors in $\boldsymbol{\Theta}$ used in the *ICSV-RC* formulation.

2.2.3 Mixed logit model with expected travel time and random coefficient on travel time (ML-RC)

This model (*ML-RC*) involves using the expected travel time, $E(TT_i^*)$, obtained from the parameter estimates of the stochastic travel time function (instead of using the entire stochastic distribution for travel time) in the choice utility function. The coefficient γ_{TT^*} is assumed to be random to capture traveller taste heterogeneity. This specification leads to the standard mixed logit model that is long established in the existing literature. In this case, the likelihood that alternative i is chosen is given by:

$$P(y_i{=}1 | \boldsymbol{D}, \boldsymbol{N}, \boldsymbol{X}, \breve{\boldsymbol{\Theta}}) = \int_{\gamma_{TT^*}} \frac{\exp\left(\gamma_{TT^*} E(TT_i^*) + \boldsymbol{\theta}' \boldsymbol{x_i}\right)}{\sum_{j=1}^{J} \exp\left(\gamma_{TT^*} E(TT_j^*) + \boldsymbol{\theta}' \boldsymbol{x_j}\right)} h(\gamma_{TT^*}) d\gamma_{TT^*} \quad (10)$$

2.2.4 Multinomial logit (MNL) and mixed logit with error components (ML-EC)

Simplifying the above model further by treating the travel time coefficient as deterministic results in the multinomial logit (*MNL*) model with expected trave time ($E(TT_i^*)$) as one of the variables in the utility function.

In situations with correlated utility functions, one can also specify a mixed logit model with error components (*ML-EC*). For example, in route choice settings, error components may be included to consider inter-route correlations due to unobserved factors common to routes sharing

the same major roads (Frejinger and Bierlaire, 2007). Such error components may be included in all the above-discussed model structures.

3 ESTIMATION BIAS DUE TO IGNORING STOCHASTICITY IN VARIABLES

In this section, following the approach by Diaz *et al*. (2015), we analytically show that in models that ignore stochasticity in explanatory variables (for example, travel time), the estimates of both the mean and standard deviation of the coefficient on such variables would be biased toward zero.

Consider the true model below where the utility associated with alternative *i* is given as:

$$\widetilde{U_i} = (b + \sigma_b z_b)\left(\mu_{X_i} + \sigma_{X_i} z_{X_i}\right) + \varepsilon_i \tag{11}$$

In the above model, the explanatory variable $X_i = \mu_{X_i} + \sigma_{X_i} z_{X_i}$ is random and assumed to be normally distributed, $z_{X_i} \sim N(0,1)$. Also, assume that the estimated coefficient on X_i is normally distributed with mean b and standard deviation σ_b.

Now, assume that the stochasticity in the exogenous variable X_i is ignored, although it is present in the true model. The ignored variability gets lumped with the stochastic term of the utility function to result in a new stochastic term δ_i, as below:

$$\begin{aligned} \widetilde{U_i} &= b\mu_{X_i} + \sigma_b z_b \mu_{X_i} + b\sigma_{X_i} z_{X_i} + \sigma_b z_b \sigma_{X_i} z_{X_i} + \varepsilon_i \\ &= b\mu_{X_i} + \sigma_b z_b \mu_{X_i} + \delta_i \end{aligned} \tag{12}$$

where, $\delta_i = b\sigma_{X_i} z_{X_i} + \sigma_b z_b \sigma_{X_i} z_{X_i} + \varepsilon_i$.

In the above utility function, δ_i has a standard deviation $\sigma_\delta = \sqrt{b^2 {\sigma_{X_i}}^2 + {\sigma_b}^2 {\sigma_{X_i}}^2 + {\sigma_\varepsilon}^2}$. Assuming an IID Gumbel distribution for ε_i, IID across individuals and across alternatives, the estimated parameters for b and σ_b can be expressed as:

$$\tilde{b} \cong \frac{\pi}{\sqrt{6}} \frac{b}{\sigma_\delta} \text{ and } \widetilde{\sigma_b} \cong \frac{\pi}{\sqrt{6}} \frac{\sigma_b}{\sigma_\delta} \tag{13}$$

where, b and σ_b are the true parameters, respectively, and $\tilde{b}$ and $\widetilde{\sigma_b}$ are the parameter estimates. This follows from the basic knowledge of RUM-based discrete choice models that the parameter estimates are confounded by the unknown standard deviation of the random utility function.

Next, let's assume that the stochasticity in exogenous variable X_i is not ignored. The corresponding utility associated with alternative i may be written as:

$$U_i = (b + \sigma_b z_b)(X_i) + \varepsilon_i$$
$$= bX_i + \sigma_b z_b X_i + \varepsilon_i \quad (14)$$

That is, the true model and the estimated model are the same. Since the stochasticity of the exogenous variable X_i is explicitly recognized (along with that in the coefficient), the remaining random utility component would be left with only ε_i. Again, assuming an IID Gumbel distribution for ε_i, the parameter estimates for b and σ_b can be expressed as:

$$\hat{b} \cong \frac{\pi}{\sqrt{6}}\frac{b}{\sigma_\varepsilon} \text{ and } \widehat{\sigma_b} \cong \frac{\pi}{\sqrt{6}}\frac{\sigma_b}{\sigma_\varepsilon}, \quad (15)$$

where b and σ_b are the true parameters, respectively, and $\hat{b}$ and $\widehat{\sigma_b}$ are the parameter estimates.

Comparing the parameter estimates $\{\tilde{b}, \widetilde{\sigma_b}\}$ and $\{\hat{b}, \widehat{\sigma_b}\}$ in Equations (13) and (15), it is easy to see that $\tilde{b} < \hat{b}$ and $\widetilde{\sigma_b} < \widehat{\sigma_b}$ because $\sigma_\delta > \sigma_\varepsilon$. That is, one can expect the parameter estimates of both the mean and standard deviation of the random coefficient to be biased toward zero when stochasticity in explanatory variables is ignored. For the same reason discussed above, a bias toward zero can be expected for other parameter estimates in linear utility functions.[4]

To be sure, Diaz *et al*. (2015) discuss that models ignoring stochasticity in explanatory variables would lead to a downward bias in the magnitude of parameter estimates. However, they did not consider random coefficients along with stochastic variables (they considered deterministic coefficients). Therefore, it was not clear that the variability of random coefficients would get underestimated if the stochasticity in explanatory variables is ignored. Further, they do not suggest a way to identify both sources of stochasticity, which this study aims to address.

4 SIMULATION EXPERIMENTS

Simulation experiments were conducted for two distinct choice settings. The first is a travel mode choice context with three labelled alternatives, for which data were generated to reflect travel conditions akin to those in Bengaluru, India. The second context is that of route choice with unlabelled alternatives, for which synthetic data were generated to mimic the empirical dataset we used for the empirical analysis on route choice from Florida, USA. The current section discusses

[4] Note that this trend in bias (toward zero) is similar to the bias one can expect when a random coefficient in the utility function is incorrectly specified as deterministic, which is an established result in the literature (Brownstone *et al*., 2000; Cherchi and Ortúzar, 2008; Swait and Bernardino, 2000). In sum, ignoring any source of variability in the utility functions of RUM-based choice models can lead to parameter estimates with a bias toward zero.

the simulation design and evaluation results for the mode choice setting, while those for the route choice context are presented in Appendix A.

4.1 Simulation design

The mode choice simulation experiments were conducted using 200 simulated datasets, each dataset with a sample size of 5,000 individuals. Three modes – bus, car, and walk – were considered. Of these, bus and car were assumed to be available for all individuals, while the walk mode was assumed to be available for travel distances of 10 km or less. Two mode-specific attributes were considered in the utility functions: travel time and travel cost, as shown below:

$$U_b = \beta_{0b} + \gamma_{TT^*} TT_b^* + \beta_C TC_b + \varepsilon_b \quad (16)$$

$$U_c = \gamma_{TT^*} TT_c + \beta_C TC_c + \varepsilon_c \quad (17)$$

$$U_w = \beta_{0w} + \gamma_{TT^*} TT_w + \varepsilon_w \quad (18)$$

In the above utility functions, travel times of the bus mode (TT_b^*) were considered stochastic ($TT_b^* \sim N(\mu_{TT_b^*}, \sigma_{TT_b^*})$) while those of other modes (TT_c and TT_w) were considered deterministic. To simulate travel times for the bus mode, the equation $TT_b^* = \theta_b d_b$ was used, where θ_b is the inverse speed for the bus mode and d_b is the trip distance on bus. θ_b was assumed to follow a left-truncated normal distribution, whose underlying normal distribution mean is 1.50 min/km (which corresponds to a speed of 40 kmph) and standard deviation (SD) is 0.15. The left-truncation value for θ_b is 1.33 min/km (i.e., a maximum bus speed of 45 kmph in the city), which was assumed to be known while the afore-mentioned mean and standard deviation are parameters to be estimated. Trip distance (d_b) is an exogenous variable representing distances travelled in Bengaluru. TC_b and TC_c are travel costs of the bus and car modes, respectively, which were assumed to be deterministic. Data for TC_b were generated to reflect distance-based bus ticket prices in the city, and data for TC_c were generated based on fuel price in Bengaluru and average mileage of a hatchback car model. The travel time coefficient (γ_{TT^*}) was assumed to follow a normal distribution with mean -1.00 and standard deviation (SD) 0.19. The travel cost coefficient (β_C) was assumed to be -0.25. Finally, the error terms (ε_b, ε_c, ε_w) were assumed to be IID standard Gumbel distributed.

Using the above utility functions and the utility maximization principle, 200 mode choice datasets, each comprising 5,000 trips were simulated. For each of the 5,000 trips from each of the 200 datasets, a single measurement of bus travel time (i.e., observed travel time or OTT_{im}, where $m = 1$) was simulated by adding a normal distributed measurement error to the simulated value of ${TT_b}^*$. The standard deviation of this measurement error was assumed to be 0.95.

4.2 Evaluation and discussion

The above-discussed simulated data of mode choices and observed travel times (y_i, OTT_{im}) along with the simulated exogenous variables ($d_b, TC_b, TC_c, TT_c, TT_w$) were used to estimate the models discussed in Section 2. Parameter recovery across the 200 simulated datasets was examined using the metrics summarized below:

(1) *Absolute Percentage Bias* (APB): For a given parameter in the model, APB is the absolute value of the difference between the true parameter value and the mean of the parameter estimates across the 200 simulated datasets – expressed as a percentage of the true parameter value.

(2) *Asymptotic Standard Error* (ASE): ASE for a given parameter is the mean (across the 200 simulated datasets) of the standard errors of the parameter's estimated values.

(3) *Finite Sample Standard Error* (FSSE): FSSE for a given parameter is the standard deviation of the parameter's estimated values across the 200 datasets.

Table 1 presents the above evaluation metrics for different models estimated in this study. The true parameter values used for simulating the data are shown in the second column of the table. The next set of columns, under the title "*ICSV-RC model parameter estimates*", shows the parameter recovery metrics for the proposed *ICSV-RC* model. As can be observed from these columns, the parameters of both the travel time model and the mode choice model are recovered accurately (i.e., with low APB) and precisely (i.e., with low standard errors). In addition, the closeness of the FSSE and ASE values suggests that the estimator of the standard error serves as a good approximation to the finite sample efficiency for the sample size considered in the study. Further, it is worth noting from Table 1 that the APB values of the *ICSV-RC* model are lower than those of all other models – *ICSV*, *ML-RC*, and *MNL* models. Furthermore, although not reported in the table, in most of the 200 datasets, the data fit of the mode choice component of the *ICSV-RC* model was statistically superior (as tested by the log-likelihood ratio test) to that of other models that ignore either the randomness in travel time (*ML-RC* model), or randomness in the

coefficient of travel time (*ICSV* model), or both sources of stochasticity (*MNL*). These results highlight the importance of incorporating stochasticity in explanatory variables and the coefficients on such variables. Ignoring any of these two sources of variability, when present, can potentially lead to inferior parameter recovery and inferior model fit.

Next, we turn to the direction of bias in the parameter estimates for the *ML-RC* model that ignores travel time variability when compared to the *ICSV-RC* model that incorporates travel time variability. In Table 1, the column titled "*t-stat. for* H_0: $\hat{\beta}_{ICSV-RC} = \hat{\beta}_{ML-RC}$" presents the t-test statistics for the null hypothesis that the magnitude of the parameter estimates from the *ICSV-RC* model are statistically the same as those from the *ML-RC* model. As can be observed from the parameter estimates of the two models (*ICSV-RC* and *ML-RC*) and the t-test statistics, both the estimated mean and standard deviation of the coefficient on travel time are biased towards zero in the *ML-RC* model. Further, the other parameters in the mode choice utility functions also demonstrate a similar trend – a bias towards zero – in the *ML-RC* model when compared to models that incorporate both travel time variability and random coefficient on travel time. These results are in line with the discussion in Section 3 on the repercussions of ignoring stochasticity in explanatory variables. In addition, the parameter estimates in the mode choice model component of the *ICSV* model are also biased towards zero when compared to those in the choice component of the *ICSV-RC* model. The bias increases further in the *MNL* model that ignores both sources of variability – stochastic variables and random coefficients on those variables. These results highlight the importance of incorporating both sources of variability, ignoring either of which would result in parameter estimates with a systematic bias toward zero.

Finally, during the estimation of the *ICSV-RC* model on each of the 200 datasets, we explored different sets of starting values for the parameters. For each dataset, the *ICSV-RC* model converged to the same maximum likelihood parameter estimates regardless of the starting parameter values employed in estimation. This pattern indicates that the proposed *ICSV-RC* model did not encounter a flat likelihood surface at the maximum likelihood values of the parameters. This observation, and that the *ICSV-RC* model provided a better fit and lower APB than other models, indicates that the *ICSV-RC* model can be used to simultaneously identify stochasticity in alternative attributes and random heterogeneity in response to those attributes – conditional on the availability of data on attribute measurements and traveller choices.

Table 1 Simulation evaluation results for the mode choice setting

	True value	*ICSV-RC* model parameter estimates				*ICSV* model parameter estimates					*ML-RC* model parameter estimates					*MNL* model parameter estimates			
		Mean	APB	FSSE	ASE	Mean	APB	FSSE	ASE	t-stat. for H_0: $\hat{\beta}_{ICSV-RC} = \hat{\beta}_{ICSV}$	Mean	APB	FSSE	ASE	t-stat. for H_0: $\hat{\beta}_{ICSV-RC} = \hat{\beta}_{ML-RC}$	Mean	APB	FSSE	ASE
Bus travel time model																			
Mean inverse speed θ_b	1.50	1.49	0.7	0.002	0.004	1.47	1.8	0.002	0.004	4.04	--	--	--	--	--	--	--	--	--
SD of θ_b	0.15	0.14	9.3	0.002	0.003	0.13	14.2	0.002	0.003	2.28	--	--	--	--	--	--	--	--	--
SD of measurement error	0.95	0.93	2.1	0.045	0.043	0.94	1.3	0.045	0.043	0.16	--	--	--	--	--	--	--	--	--
Mean APB, FSSE, ASE	--	--	4.1	0.016	0.017	--	5.7	0.017	0.017	--	--	--	--	--	--	--	--	--	--
Mode choice model																			
ASC for transit	-0.56	-0.53	5.4	0.066	0.069	-0.61	8.9	0.056	0.067	0.83	-0.22	60.0	0.053	0.057	2.89	-0.24	56.9	0.053	0.056
ASC for walk	1.56	1.47	5.9	0.120	0.114	1.17	24.8	0.093	0.096	2.02	1.02	34.6	0.097	0.095	3.03	0.96	38.2	0.090	0.092
Mean of travel time coefficient γ_{TT^*}	-1.00	-0.91	8.9	0.048	0.045	-0.76	23.5	0.032	0.028	2.85	-0.67	32.7	0.029	0.028	4.58	-0.61	38.9	0.020	0.019
SD of γ_{TT^*}	0.19	0.19	0.5	0.020	0.016	--	--	--	--	--	0.09	50.2	0.016	0.014	4.69	--	--	--	--
Cost coefficient (β_C)	-0.25	-0.23	6.1	0.013	0.014	-0.21	15.5	0.011	0.011	1.15	-0.19	25.6	0.009	0.009	2.38	-0.17	31.3	0.009	0.008
Mean APB, FSSE, ASE	--	--	7.0	0.053	0.051	--	18.2	0.048	0.050	--	--	40.6	0.041	0.041	--	--	41.3	0.043	0.044

5 EMPIRICAL ANALYSIS

In this section, we present an empirical analysis for a joint analysis of route-level travel time and route choice while considering both stochasticity in network travel times and random heterogeneity in sensitivity to travel time. This empirical analysis is focused more on corroborating the findings from the earlier sections than on the substantive aspect of route choice analysis itself.

5.1 Empirical data

The main source of empirical data for this analysis, provided by the American Transportation Research Institute (ATRI) is a large truck-GPS dataset of about 96 million GPS traces in the state of Florida, USA (Pinjari *et al*., 2015). The raw data were first converted into a database of truck trips by Thakur *et al*. (2015) using GPS-to-trip conversion algorithms. For these trips, the travelled routes were not readily observable in the form of network links and nodes traversed between the OD locations. The raw GPS data was map-matched to the roadway network to derive the travelled routes using a high-resolution roadway network obtained from the Florida Department of Transportation (FDOT) (Tahlyan *et al*., 2017). Such truck route choice data were generated for a total of 8211 truck trips in the state of Florida.

For all the 8211 trucks trips used in this study, route choice sets were generated by Tahlyan and Pinjari (2020) using the Breadth First Search-Link Elimination (BFS-LE) algorithm proposed by Rieser-Schüssler *et al.* (2013). [5] For each trip, the BFS-LE algorithm was run to generate up to 16 unique route choice alternatives. For some of these trips, the chosen route was included as an additional choice alternative since the choice set did not include the chosen route completely. Next, for all route alternatives of each of the 8211 truck trips, route attributes such as the total route length, lengths on different types of roads, number of intersections, and the proportion of toll road length were derived. In addition, to account for the degree of overlap of a route with other routes in the choice set for that same OD pair, a path-size attribute (Ben-Akiva and Bierlaire, 1999) was

[5] The BFS-LE is a deterministic link elimination approach based on a repeated least cost path search, where links on the current shortest path are eliminated, one by one, to find subsequent least cost paths. Hence, it is well-suited for extracting routes from large-scale, high-resolution networks. The primary difference between this algorithm when compared against other link-elimination approaches is that it uses a tree structure in which each node is a network. Starting initially with the original network (which is the root node of the tree), any unique network obtained after the elimination of a link from a current least cost path is a node of the tree, given that the network offers at least one feasible route for the OD pair under consideration.

computed, where a greater path-size value indicates a smaller extent of overlap and a path-size value of one indicates no overlap.[6]

The same GPS data were used to extract measurements of travel times for each of the 8211 chosen routes. Among the chosen routes, the number of available travel time measurements varied from one to as many as ten or more, although more than 70% of the chosen routes had only one or two measurements. Non-chosen route alternatives did not have travel time measurements.[7]

5.2 Empirical results and findings

Table 2 presents the parameter estimates for the empirical route choice models estimated in this study – *ICSV-RC*, *ICSV*, *ML-RC*, and *ML-EC*. In all these models, we included error components to consider correlations among route-specific utility functions due to unobserved factors. Due to the importance of such error correlations in route choice settings, we did not consider a simple *MNL* model for this analysis. Next, we briefly discuss the empirical results from the *ICSV-RC* model and compare them with those of other models to evaluate the importance of accommodating both sources of stochasticity discussed earlier.

5.2.1 Empirical results from the ICSV-RC model

The columns in Table 2 under the title "*ICSV-RC* model" presents the parameter estimates for the proposed *ICSV-RC* model where both the travel time and its coefficient are specified as random. In this model, the random coefficients in the stochastic travel time function (Equation (3)) were specified as normally distributed. Other distributional assumptions could be made in this regard, such as a truncated normal or a shifted lognormal; however, we used the normal distribution specification as an initial effort to disentangle variability in travel time from that in its coefficient (which is also assumed to be normally distributed).

---

[6] The path-size variable for a route i is defined as: $PS_i = \sum_{a \epsilon \Gamma_i} \left(\frac{l_a}{L_i}\right) \frac{1}{\sum_{j \epsilon C_n} \delta_{aj}}$, where Γ_i is the set of all links in path/route i between the OD pair n, l_a is the length of link a, L_i is the length of path i, C_n is the choice set of route alternatives between the OD pair n. δ_{aj} is equal to 1 if a route $j \epsilon C_n$ uses link a and 0 otherwise.

[7] The same empirical dataset was used by Biswas *et al*. (2019) for estimating an *ICSV* model of truck route choice and route-level travel time. However, in that paper, there is no recognition that the *ICSV* framework can be extended to the *ICSV-RC* framework to simultaneously identify both travel time variability and random heterogeneity in response to travel time. Also, there is no discussion of the nature of bias in parameters when either of these sources of stochasticity is ignored. Besides, the *ICSV* formulation of Biswas *et al*. (2019) is based on a multinomial probit kernel for the route choice model, which is not easy to use when both travel time and the coefficient on travel time in the model are stochastic. In this paper, we use the logit-based kernel for the route choice component, which makes it easier to consider both stochastic travel time and a random coefficient on it in an *ICSV-RC* framework.

Table 2 Empirical results for the route choice setting

	ICSV-RC model		***ICSV model***			***ML-RC model***			***ML-EC model***	
	Parameter estimates	Std. error	Parameter estimates	Std. error	t-stat for H_0: $\hat{\beta}_{ICSV-RC} = \hat{\beta}_{ICSV}$	Parameter estimates	Std. error	t-stat for H_0: $\hat{\beta}_{ICSV-RC} = \hat{\beta}_{ML-RC}$	Parameter estimates	Std. error
Structural eqn. for stochastic travel time										
Interstate highway length - mean parameter	0.955	0.0017	0.945	0.0017	--	--	--	--	--	--
Major arterial length - mean parameter	1.284	0.0050	1.322	0.0056	--	--	--	--	--	--
Minor arterial length - mean parameter	1.599	0.0120	1.709	0.0156	--	--	--	--	--	--
Collector street length - mean parameter	1.924	0.0199	2.180	0.0267	--	--	--	--	--	--
Local road length - mean parameter	2.784	0.0398	2.851	0.0458	--	--	--	--	--	--
Total number of junctions - mean parameter	0.194	0.0148	0.067	0.0155	--	--	--	--	--	--
Interstate highway length - SD parameter	0.069	0.0006	0.060	0.0006	--	--	--	--	--	--
Major arterial length - SD parameter	0.212	0.0042	0.235	0.0041	--	--	--	--	--	--
Minor arterial length - SD parameter	0.510	0.0038	0.573	0.0050	--	--	--	--	--	--
Collector street length - SD parameter	0.407	0.0153	0.559	0.0176	--	--	--	--	--	--
Measurement eqn. for travel time										
SD of measurement error in GPS data	3.603	0.0024	3.597	0.0025	--	--	--	--	--	--
Route choice utility functions										
Mean of route-level travel time coefficient	-1.243	0.0470	-0.475	0.0090	16.05	-1.072	0.0346	2.94	-0.449	0.0064
SD of route-level travel time coefficient	0.871	0.0338	0.000	--	--	0.678	0.0253	4.56	--	--
Natural logarithm of path size	-2.340	0.0880	-2.048	0.0597	2.75	-1.119	0.6331	1.91	-1.014	0.5114
Proportion of tolled portion on the route	-7.644	1.0444	-7.214	0.7804	0.33	-5.245	0.0640	2.29	-3.637	0.0892
Error components in utility functions										
SD of error component on square root of route length on interstate 75 in Florida	5.558	0.2522	2.945	0.1196	9.35	4.257	0.2172	3.91	2.994	0.0923
SD of error component on square root of route length on Polk Parkway in Florida	3.403	0.2813	2.724	0.2548	1.79	3.127	0.3568	0.61	2.386	0.2092

As can be observed from the parameter estimates of the stochastic travel time equation, the estimates for mean inverse speeds (in minutes per mile) and the corresponding standard deviations are in increasing order from interstates to local roads. This is intuitive given that interstates figure at the top in the hierarchy of functional classification of roadways. Further, the probability of zero or negative values of the mean inverse speeds is zero for all practical purposes (less than 8.6×10^{-3} for minor arterials and of the order of 10^{-6} or lesser for other roadway functional classes). The value of mean junction-crossing time at turns (0.194 minutes per turn) also turned out to be statistically significant.

As discussed in Section 2, the measurement equation for the travel time model includes a measurement error term. The standard deviation estimate in the measurement equation for travel time suggests a significant error in the measurement or extraction of travel time using GPS data.

Moving on to the route-choice model component, it is notable that in addition to allowing the estimation of random coefficients in the stochastic travel time equation (i.e., stochasticity in travel time), the model allows the estimation of a random coefficient on travel time in its route choice utility component. Specifically, the mean and standard deviation parameter estimates of γ_{TT^*} (see the coefficient on route-level travel time distribution and its standard deviation in Table 2) are statistically significant and reasonable, with more than 92% of the population having a negative value for γ_{TT^*}.

In the remainder of the route choice utility function, the coefficient for the natural logarithm of path size has a negative sign, which is expected because routes with higher overlap would each have a lower probability of being chosen than the probability of all of them being chosen. Further, the coefficient on the proportion of tolled roads on a route is negative, indicating lower utilities for routes having greater proportions of tolled lengths, *ceteris paribus*. In addition, the error components specified in the utility functions to capture inter-route correlations are statistically significant.

5.2.2 Empirical results from the ICSV model

Now, we turn to the set of parameter estimates for the *ICSV* model in Table 2, where the travel time coefficient was estimated as a fixed parameter. The parameter estimates for the choice model component in this model demonstrate a bias towards zero when compared to those in the *ICSV*-

RC model. That is, as discussed in Section 3, ignoring randomness in the coefficient on the stochastic travel time leads to biased (toward zero) parameter estimates in the choice model.

5.2.3 Empirical results from mixed logit models

Next, let us examine the results for the *ML-RC* model, which ignores the stochasticity in travel time. As can be observed from Table 2, the *ML-RC* route choice model parameter estimates for the mean and standard deviation coefficients on route-level travel time are biased toward zero when compared to the corresponding coefficients from the *ICSV-RC*. This finding, once again, corroborates our claims from Section 3. In particular, the statistically significant underestimation of the two primary model parameters under scrutiny (the mean and standard deviation of the coefficient on travel time) when travel time stochasticity is ignored highlights the drawbacks of using conventional mixed logit models in settings that involve random variables such as travel time as well as randomness in the sensitivity to such variables.

The *ML-EC* model's coefficient on route-level travel time also shows a bias toward zero. Since this model ignores both sources of variability, the bias in its parameter estimates is greater than that in the *ICSV* model or the *ML-RC* model.

5.2.4 Goodness-of-fit in estimation and validation samples

To assess the goodness-of-fit of the various empirical models estimated in this study, we conducted a five-fold validation. That is, from the full dataset of 8,211 trips available for the empirical study, we randomly drew five estimation samples of 6,453 trips and estimated all the above-discussed models. For each of the five estimation samples of 6,453 trips, the remaining 1,758 trips were kept aside for validation purposes. Subsequently, we computed the goodness-of-fit metrics shown in Table 3 for all five sets of estimation and validation samples (for each set of estimation and validation samples, the parameter estimates from the corresponding estimation sample were used). These metrics include log-likelihood at convergence for the integrated models, log-likelihood at convergence for only the route choice model component, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values for the route choice model component, and adjusted McFadden's Rho-square for the route choice model component. Average values of these metrics (averaged across the five sample datasets) are reported in Table 3 – separately for the estimation samples (in the first set of rows) and the validation samples (in the second set of rows).

As can be observed from the table, the choice model component of the *ICSV-RC* model provides the best goodness-of-fit measures, followed by that of the *ICSV*, the *ML-RC*, and the *ML-EC* (in that order). The same trend can be observed in both estimation and validation samples. It is an expected result that the *ICSV-RC* that incorporates stochastic travel time and a random coefficient on travel time will provide a better fit than other models that ignore one or both sources of variability.

Table 3 Goodness-of-fit metrics in estimation and validation samples

Goodness-of-fit measures in estimation samples	***ICSV-RC model***	***ICSV model***	***ML-RC model***	***ML-EC Model***
Log-likelihood at convergence	-2,15,846.07	-2,16,424.19	--	--
Log-likelihood of route choice component (LL_β)	-6,464.24	-6,935.97	-6,554.99	-7,178.41
No. of parameters estimated in the choice model (k)	6	6	6	6
AIC for choice model component = - 2 (LL_β) +2k	12,960.47	13,901.93	13,121.97	14,366.81
BIC for choice model component = k ln (n) - 2 (LL_β)	13,068.82	14,003.52	13,162.61	14,400.67
McFadden's Rho-square for the choice model	0.597	0.568	0.592	0.553
Goodness-of-fit measures in validation samples	***ICSV-RC model***	***ICSV model***	***ML-RC Model***	***ML-EC Model***
Log-likelihood at convergence	-62,750.10	-62,889.24	--	--
Log-likelihood of route choice component (LL_β)	-1,687.70	-1,835.08	-1,701.32	-1,896.91
No. of parameters estimated in the choice model (k)	6	6	6	6
AIC for choice model component = - 2 (LL_β) +2k	3,407.39	3,711.89	3,414.65	3,803.82
BIC for choice model component = k ln (n) - 2 (LL_β)	3,494.95	3,752.60	3,447.48	3,825.78
McFadden's Rho-square for the choice model	0.617	0.583	0.614	0.569

6 SUMMARY AND CONCLUSIONS

In this study, we formulate a choice modelling framework that allows the analyst to simultaneously accommodate stochasticity in explanatory variables and random coefficients on such variables. Specifically, we develop an integrated choice and stochastic variable modelling framework with random coefficients (i.e., an *ICSV-RC* framework) to disentangle travel time variability from unobserved heterogeneity in response to travel time in travel choice models. The proposed *ICSV-RC* model allows the simultaneous identification of both the sources of variability – stochastic explanatory variables and random coefficients on those variables – due to its ability to bring together travel choice data and measurement data for the stochastic variables. In addition, we show that ignoring either source of stochasticity – stochasticity in alternative attributes or heterogeneity

in response to the attributes – results in models with inferior goodness-of-fit and a systematic bias toward zero in all parameter estimates. We demonstrate this using simulation experiments for two different travel choice settings – one involving labelled mode choice alternatives and the other involving unlabelled route choice alternatives. Furthermore, we applied the proposed *ICSV-RC* model to an empirical analysis of truck route choice in Florida, USA. The integrated model was found to successfully disentangle stochasticity in route-level travel time from heterogeneity in response to travel time. Simpler versions of the model that ignore either stochasticity in travel time or impose a deterministic coefficient on travel time had inferior goodness-of-fit and showed a bias toward zero in the parameter estimates. These results highlight the importance of accounting for both sources of variability.

The methodology proposed in this paper overcomes the limitation of mixed logit/probit models used to accommodate random coefficients on deterministic explanatory variables and the limitation of ICLV models used to accommodate latent (stochastic) variables with deterministic coefficients. It has hitherto been believed that identification of both these sources of variability – stochastic attributes and random coefficients on those attributes – is very difficult, if not impossible (Diaz *et al.*, 2015). Going forward, we hope that the proposed method will help increase the simultaneous recognition of stochastic variables and random coefficients in choice models.

Some limitations of this study offer scope for further research. First, we used the normal distributional assumption for travel time as a first step to address the core idea that the paper puts forth – the identification of the two sources of variability. The authors recognize that a normal (or truncated normal) distribution for travel time would imply, for a given distance, a reciprocal normal (reciprocal of truncated normal) distribution for travel speed, which leads to theoretically undefined mean and variance parameters. Therefore, it is important to explore alternative distributional forms for the stochastic travel time variable. Second, we used the maximum simulated likelihood estimation method in this study. The estimation time for each dataset was about two hours on a workstation-grade computer. The exploration of alternative estimation methods for the proposed model is a fruitful research avenue. Finally, the current study focuses on the stochasticity of alternative attributes, which vary across choice alternatives. It will be useful to explore avenues to identify stochasticity of choice environment variables that do not vary across choice alternatives. A recent study by Nirmale and Pinjari (2022) is a step forward in this direction.

ACKNOWLEDGMENTS

This research was undertaken as part of a SPARC project funded by the Indian Ministry of Education. The authors are thankful to the American Transportation Research Institute (ATRI) for providing the truck-GPS data and to Divyakant Tahlyan for providing the processed data for the empirical analysis of route choice presented in this study.

APPENDIX A: SIMULATION EVALUATION FOR THE ROUTE CHOICE SETTING

In addition to the mode choice simulation experiments presented in Section 4 of the paper, we conducted a second set of simulation experiments for a route choice setting with unlabelled alternatives. To keep the synthetic data realistic (akin to that from a real road network), we drew data on exogenous variables and choice sets from the empirical route choice data used in Section 5 of the study. The true values of the parameters assumed to simulate the data are taken from the parameter estimates reported for the empirical *ICSV-RC* model in Table 2. Using these parameters and the exogenous variable data, we generated 200 route choice datasets, each with a sample size of 2000 trips for the proposed *ICSV-RC* model. For each of the 2000 trips in the 200 datasets, we generated 10 travel time measurements for the simulated chosen route. We assumed that the non-chosen routes would not have travel time measurements.

We estimated all the models discussed in Section 2 on all the 200 simulated datasets and then computed the parameter recovery metrics discussed in Section 4 (i.e., APB, ASE, and FSSE). These metrics are reported in Table A1 below. As can be observed from this table, the trends in overall parameter recovery and bias in parameter estimates are similar to those observed in Section 4 for the mode choice context. These findings thus underscore the need for a model framework such as the *ICSV-RC* model to accommodate both stochastic variables and random coefficients when compared to the other models typically used in the literature.

Table A1 Simulation evaluation results for the route choice setting

Variable description	True value	*ICSV-RC* model parameter estimates				*ICSV* model parameter estimates					*ML-RC* model parameter estimates					*ML-EC* model parameter estimates			
		Mean	APB	FSSE	ASE	Mean	APB	FSSE	ASE	t-stat. for H_0: $\hat{\beta}_{ICSV-RC} = \hat{\beta}_{ICSV}$	Mean	APB	FSSE	ASE	t-stat. for H_0: $\hat{\beta}_{ICSV-RC} = \hat{\beta}_{ML-RC}$	Mean	APB	FSSE	ASE
Structural eqn. for stochastic travel time																			
Interstate highway length – mean parameter	0.955	0.955	0.06	0.0034	0.0026	0.953	0.16	0.0038	0.0026	--	--	--	--	--	--	--	--	--	--
Major arterial length – mean parameter	1.284	1.284	0.02	0.0172	0.0149	1.283	0.07	0.0164	0.0071	--	--	--	--	--	--	--	--	--	--
Minor arterial length - mean parameter	1.599	1.596	0.17	0.0337	0.0238	1.597	0.11	0.0331	0.0139	--	--	--	--	--	--	--	--	--	--
Collector street length - mean parameter	1.924	1.919	0.25	0.0475	0.0402	1.907	0.89	0.0513	0.0201	--	--	--	--	--	--	--	--	--	--
Local road length - mean parameter	2.784	2.802	0.63	0.1065	0.0919	2.816	1.13	0.1085	0.0909	--	--	--	--	--	--	--	--	--	--
Total number of turns - mean parameter	0.194	0.198	2.20	0.0320	0.0271	0.203	4.95	0.0301	0.0270	--	--	--	--	--	--	--	--	--	--
Interstate highway length - SD parameter	0.069	0.064	7.85	0.0162	0.0170	0.065	5.86	0.0036	0.0017	--	--	--	--	--	--	--	--	--	--
Major arterial length - SD parameter	0.212	0.214	0.99	0.0148	0.0110	0.228	3.20	0.0151	0.0045	--	--	--	--	--	--	--	--	--	--
Minor arterial length - SD parameter	0.510	0.521	2.22	0.0293	0.0208	0.513	0.55	0.0251	0.0075	--	--	--	--	--	--	--	--	--	--
Collector street length - SD parameter	0.407	0.398	2.13	0.0369	0.0324	0.391	3.96	0.0309	0.0117	--	--	--	--	--	--	--	--	--	--
Measurement eqn. for travel time																--	--	--	--
SD of measurement error in GPS data	3.603	3.790	5.19	0.0256	0.0266	3.783	4.99	0.0018	0.0240	--	--	--	--	--	--	--	--	--	--
Mean of APB, FSSE, ASE values	--	--	1.97	0.0330	0.0280	--	2.09	0.0318	0.0187	--	--	--	--	--	--	--	--	--	--
Model (utility functions) for route choice																			
Coefficient on travel time - mean parameter	-1.243	-1.088	12.50	0.0652	0.0625	-0.875	29.61	0.0026	0.0063	3.39	-0.977	21.45	0.0515	0.0506	1.38	-0.446	64.11	0.0089	0.0102
Coefficient on travel time - SD parameter	0.871	0.790	9.31	0.0528	0.0490	--	--	--	--	--	0.658	24.50	0.0400	0.0407	2.07	--	--	--	--
Natural logarithm of path size	-2.340	-1.984	15.23	0.1523	0.1478	-1.569	32.96	0.1080	0.1500	1.97	-0.894	61.82	0.1520	0.1013	6.08	-1.727	26.21	0.4893	0.6653
Proportion of tolled portion on the route	-7.644	-6.222	18.60	1.3051	1.3013	-4.599	39.84	1.0560	0.9860	0.99	-4.496	41.19	0.9553	0.9319	1.07	-4.230	44.66	0.0269	0.5660
Error components for inter-route correlations																			
SD of error component on square root of route length on interstate 75 in Florida	5.558	4.718	15.10	0.5196	0.3796	2.869	48.38	0.3300	0.5680	2.71	4.372	21.33	0.4884	0.373	0.65	2.107	62.10	0.1025	0.1996
SD of error component on square root of route length on interstate 75 in Florida	3.403	2.972	12.66	0.6399	0.5124	1.587	53.37	0.6690	0.8190	1.43	2.447	28.09	0.5262	0.4806	0.75	2.569	24.48	0.3669	0.2056
Mean of APB, FSSE, ASE values	--	--	13.90	0.4560	0.4090		40.83	0.4331	0.5059	--	--	33.06	0.3689	0.3296	--	--	44.31	0.1989	0.3293

References

Al-Deek, H., Emam, E. B., 2006. New methodology for estimating reliability in transportation networks with degraded link capacities. *Journal of Intelligent Transportation Systems*, *10*(3), 117-129.

Alvarez-Daziano, R., Bolduc, D., 2013. Incorporating pro-environmental preferences towards green automobile technologies through a Bayesian hybrid choice model. *Transportmetrica A: Transport Science* 9, 74–106.

Aron, M., Bhouri, N., Guessous, Y., 2014. Estimating travel time distribution for reliability analysis. *Transportation Research Arena*, TRA2014, paper, 19638.

Batley, R. P., Toner, J. P., & Knight, M. J., 2004. A mixed logit model of U.K. household demand for alternative-fuel vehicles. *International Journal of Transport Economics*, 31(1), 55–77.

Ben-Akiva, M., Walker, J., Bernardino, A. T., Gopinath, D. A., Morikawa, T., Polydoropoulou, A., 2002. Integration of choice and latent variable models. In Perpetual motion: Travel behaviour research opportunities and application challenges, 431-470. Pergamon, Amsterdam.

Ben-Akiva, M., Bierlaire, M., 1999. Discrete choice methods and their applications to short term travel decisions. Handbook of Transportation Science, Springer, 5-33.

Bhat, C. R., 2001. Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. *Transportation Research Part B: Methodological*, 35(7), 677-693.

Bhat, C. R., 2003. Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences. *Transportation Research Part B: Methodological*, 37(9), 837-855.

Bhat, C. R., 2011. The maximum approximate composite marginal likelihood (MACML) estimation of multinomial probit-based unordered response choice models. *Transportation Research Part B: Methodological*, 45(7), 923-939.

Bhat, C. R., Sidharthan, R., 2012. A new approach to specify and estimate non-normally mixed multinomial probit models. *Transportation Research Part B: Methodological*, 46(7), 817-833.

Bhat, C. R., Dubey, S. K., 2014. A new estimation approach to integrate latent psychological constructs in choice modeling. *Transportation Research Part B: Methodological*, 67, 68-85.

Bhatta, B. P., Larsen, O. I., 2011. Errors in variables in multinomial choice modeling: A simulation study applied to a multinomial logit model of travel mode choice. *Transport Policy*, 18(2), 326-335.

Biswas, M., Pinjari, A. R., Dubey, S. K., 2019, January. Travel Time Variability and Route Choice: An Integrated Modelling Framework. In 2019 11th International Conference on Communication Systems & Networks (COMSNETS), 737-742, IEEE.

Brownstone, D., Bunch, D. S., Train, K., 2000. Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles. *Transportation Research Part B: Methodological*, 34(5), 315-338.

Carroll, R. J., Spiegelman, C. H., Lan, K. G., Bailey, K. T., Abbott, R. D., 1984. On errors-in-variables for binary regression models. *Biometrika*, 71(1), 19-25.

Chen, A., Zhou, Z., Lam, W. H., 2011. Modeling stochastic perception error in the mean-excess traffic equilibrium model. *Transportation Research Part B: Methodological*, 45(10), 1619-1640.

Cherchi, E., Ortúzar, J. de D., 2008. Predicting best with mixed logit models: understanding some confounding effects. In: Inweldi, P.O. (Ed.), *Transportation Research Trends*. Nova Science Publishers, Inc., New York, pp. 215–235.

Conniffe, D., O'Neill, D., 2008. An efficient estimator for dealing with missing data on explanatory variables in a probit choice model. Available at SSRN 1284362.

Coifman, B., 1998. Vehicle re-identification and travel time measurement in real-time on freeways using existing loop detector infrastructure. *Transportation Research Record*, 1643(1), 181-191.

Daly, A. J., Ortúzar, J. D., 1990. Forecasting and data aggregation: theory and practice. *Traffic Engineering and Control*, 31(12), 632-643.

Díaz, F., Cantillo, V., Arellana, J., de Dios Ortúzar, J., 2015. Accounting for stochastic variables in discrete choice models. *Transportation Research Part B: Methodological*, 78, 222-237.

Dubey, S., Cats, O., Hoogendoorn, S., Bansal, P., 2022. A multinomial probit model with Choquet integral and attribute cut-offs. *Transportation Research Part B: Methodological*, 158, 140-163.

Durbin, J., 1954. Errors in variables. *Revue de l'institut International de Statistique*, 23-32.

Frejinger, E., Bierlaire, M., 2007. Capturing correlation with subnetworks in route choice models. *Transportation Research Part B: Methodological* 41, 363-378.

Gleser, L. J., 1981. Estimation in a multivariate "errors in variables" regression model: large sample results. *The Annals of Statistics*, 24-44.

Greene, W. H., & Hensher, D. A., 2003. A latent class model for discrete choice analysis: contrasts with mixed logit. *Transportation Research Part B: Methodological*, 37(8), 681-698.

Guo, F., Li, Q., Rakha, H., 2012. Multistate travel time reliability models with skewed component distributions. *Transportation Research Record*, 2315(1), 47-53.

Hensher, D. A., Greene, W. H., 2003. The mixed logit model: the state of practice. *Transportation*, 30(2), 133-176.

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

McFadden, D., Train, K., 2000. Mixed MNL models for discrete response. *Journal of applied Econometrics*, 15(5), 447-470.

Nirmale, S. K. and Pinjari, A. R., 2022. Discrete Choice Models with Multiplicative Stochasticity in Choice Environment Variables: Application to Accommodating Perception Errors in Driver Behaviour Models. Working paper. Indian Institute of Science.

Ortúzar, J. de D., Ivelic, A. M., 1987. Effects of using more accurately measured level-of-service variables on the specification and stability of mode choice models. In: Proceedings 15th PTRC Summer Annual Meeting, University of Bath, September 1987, England.

Ortúzar, J. de D., Willumsen, L. G., 2011. Modelling transport. John Wiley & Sons.

Patil, P. N., Dubey, S. K., Pinjari, A. R., Cherchi, E., Daziano, R., Bhat, C. R., 2017. Simulation evaluation of emerging estimation techniques for multinomial probit models. *Journal of choice modelling*, 23, 9-20.

Pinjari, A.R., Kamali, M., Luong, T., Ozkul, S., 2015. *GPS Data for Truck-Route Choice Analysis of Port Everglades Petroleum Commodity Flow*. Report BDV25-977-17. Florida Department of Transportation

Polus, A., 1979. A study of travel time and reliability on arterial routes. *Transportation*, 8(2), 141-151.

Rakha, H. A., El-Shawarby, I., Arafeh, M., Dion, F., 2006, September. Estimating path travel-time reliability. In 2006 IEEE Intelligent Transportation Systems Conference (pp. 236-241). IEEE.

Revelt, D., Train, K., 1998. Mixed logit with repeated choices: households' choices of appliance efficiency level. *Review of economics and statistics*, 80(4), 647-657.

Rieser-Schüssler, N., Balmer, M., Axhausen, K.W., 2013. Route choice sets for very high-resolution data. *Transportmetrica A: Transport Science* 9, 825-845.

Rubin, D., 1987. Multiple Imputation for Nonresponse in Surveys. Wiley, New York.

Sanko, N., Hess, S., Dumont, J., Daly, A., 2014. Contrasting imputation with a latent variable approach to dealing with missing income in choice models. *Journal of Choice Modelling* 12, 47-57.

Srinivasan, K. K., Prakash, A. A., Seshadri, R., 2014. Finding most reliable paths on networks with correlated and shifted log–normal travel times. *Transportation Research Part B: Methodological*, 66, 110-128.

Stefanski, L. A., Carroll, R. J., 1985. Covariate measurement error in logistic regression. *The Annals of Statistics*, 13(4), 1335-1351.

Steinmetz, S.S., Brownstone, D., 2005. Estimating commuters' "value of time" with noisy data: a multiple imputation approach. *Transportation Research Part B: Methodological* 39, 865-889.

Swait, J.D., Bernardino, A., 2000. Distinguishing taste variation from error structure in discrete choice data. *Transportation Research Part B: Methodological* 34, 1–15.

Swait, J., 2022. Distribution-free estimation of individual parameter logit (IPL) models using combined evolutionary and optimization algorithms. *Journal of Choice Modelling.*

Tahlyan, D., Luong, T., Pinjari, A., Ozkul, S., 2017. Development and Analysis of Truck Route Choice Data for the Tampa Bay Region using GPS Data. Report BDK25-730-3. Florida Department of Transportation.

Tahlyan, D, Pinjari, A., 2020. Performance evaluation of choice set generation algorithms for analyzing truck route choice: insights from spatial aggregation for the breadth first search link elimination (BFS-LE) algorithm. *Transportmetrica A: Transport Science*, 16 (3). 1030-1061.

Taylor, M. A., 2012. Modelling travel time reliability with the Burr distribution. Procedia-Social and Behavioral Sciences, 54, 75-83.

Thakur, A., Pinjari, A.R., Zanjani, A.B., Short, J., Mysore, V., Tabatabaee, S.F., 2015. Development of Algorithms to Convert Large Streams of Truck GPS Data into Truck Trips. *Transportation Research Record: Journal of the Transportation Research Board* 2529, 66-73.

Train, K., 1978. A validation test of a disaggregate mode choice model. *Transportation Research*, 12(3), 167-174.

Varotto, S. F., Glerum, A., Stathopoulos, A., Bierlaire, M., Longo, G., 2017. Mitigating the impact of errors in travel time reporting on mode choice modelling. *Journal of Transport Geography*, 62, 236-246.

Vij, A., Walker, J. L., 2016. How, when and why integrated choice and latent variable models are latently useful. *Transportation Research Part B: Methodological*, 90, 192-217.

Walker, J., Li, J., Srinivasan, S., Bolduc, D., 2010. Travel demand models in the developing world: Correcting for measurement errors. *Transportation Letters*, 2(4), 231-243.

Zang, Z., Xu, X., Yang, C., Chen, A., 2018. A distribution-fitting-free approach to calculating travel time reliability ratio. *Transportation Research Part C: Emerging Technologies*, 89, 83-95.