JOURNAL OF ADVANCED TRANSPORTATION
J. Adv. Transp. 2014; **48**:48–65
Published online 28 February 2012 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/atr.201

Bayesian imputation of non-chosen attribute values in revealed preference surveys

Simon Washington[1]*, Srinath Ravulaparthy[2], John M. Rose[3], David Hensher[4] and Ram Pendyala[5]

[1]*School of Urban Development, Faculty of Science and Engineering, Queensland University of Technology, 2 George Street, Brisbane, Queensland 4001, Australia*
[2]*Department of Geography, University of California, Santa Barbara, Santa Barbara, CA, U.S.A.*
[3]*Institute of Transport and Logistics Studies, University of Sydney, 144 Burren Street Newtown, Sydney, New South Wales 2042, Australia*
[4]*Institute of Transport and Logistics Studies, University of Sydney, 144 Burren Street Newtown, Sydney, New South Wales 2006, Australia*
[5]*School of Sustainable Engineering and the Built Environment, Arizona State University, Tempe, AZ, U.S.A.*

SUMMARY

Obtaining attribute values of non-chosen alternatives in a revealed preference context is challenging because non-chosen alternative attributes are unobserved by choosers, chooser perceptions of attribute values may not reflect reality, existing methods for imputing these values suffer from shortcomings, and obtaining non-chosen attribute values is resource intensive. This paper presents a unique Bayesian (multiple) Imputation Multinomial Logit model that imputes unobserved travel times and distances of non-chosen travel modes based on random draws from the conditional posterior distribution of missing values. The calibrated Bayesian (multiple) Imputation Multinomial Logit model imputes non-chosen time and distance values that convincingly replicate observed choice behavior. Although network skims were used for calibration, more realistic data such as supplemental geographically referenced surveys or stated preference data may be preferred. The model is ideally suited for imputing variation in intrazonal non-chosen mode attributes and for assessing the marginal impacts of travel policies, programs, or prices within traffic analysis zones. Copyright © 2012 John Wiley & Sons, Ltd.

KEY WORDS: multinomial logit; choice models; imputation; synthesized data; Bayesian methods; missing data analysis; unobserved choice attributes

1. INTRODUCTION

Obtaining reliable information on travel mode attribute values in revealed preference surveys is time consuming, resource intensive, yet necessary for estimating useful choice models. Obtaining attribute values of non-chosen travel modes is particularly challenging. First, non-chosen mode attributes are often unobserved by choosers over the course of 1-day to 3-day travel surveys. Second, chooser perceptions of mode attribute values, such as travel time or distance, may not correspond with reality. Finally, the methods for obtaining or imputing these values have varying degrees of inaccuracy, resource demands, and defensibility.

In practice, five approaches have been used for obtaining information on non-chosen alternative attribute values (adapted from [1]). Perhaps the most commonly practiced approach is to use network skim values obtained from transportation planning models and use these forecasted attribute values (by origin-destination [O-D] pair) to impute unobserved attributes in the survey. Skims are generated to determine the cost of travel (e.g., time, distance) on the "cheapest" route from zone i to zone j. The skims are based on the computation of the minimum time paths between zones based on free-flow link

*Correspondence to: Simon Washington, Faculty of Built Environment and Engineering, Queensland University of Technology, 2 George Street, Brisbane, Queensland 4001, Australia. E-mail: simon.washington@qut.edu.au

speeds. The skims generated are then blended using weighted averages to replicate the actual costs or travel times. The problems with this approach include lack of sufficient sample sizes for certain O-D pairs by mode (e.g., [2]), lack of resolution within a Traffic Analysis Zone (TAZ) (e.g., the bus travel time within a TAZ is assumed constant, whereas there can be tremendous variability in walk times within a TAZ, especially in suburban and rural TAZs), and lack of dynamic realism regarding congestion effects on travel time by time of day (e.g., traditional peak and off-peak assignments capture two time of day congestion effects).

A second approach involves imputing unobserved attribute values with the average attribute values of observed alternatives. A major limitation of this approach is that sample sizes may be quite small or even non-existent for calculating certain O-D pairs—a network with 500 zone centroids for example, has 998 000 one-way peak period travel times to estimate in a modest four mode choice context (500*499*4). Moreover, imputed values are constant within an O-D pair, as travel times or distances within a TAZ are not differentiated. This method also does not preserve the variance in the underlying variable and as such produces inconsistent estimates [3].

A third method involves sampling across the distribution of choosers to identify respondents having chosen multiple alternatives and substituting the means of the observed sample attributes for non-observed alternatives. For example, a person observed to take both auto and bus (on say different days of a multiday travel diary) will inform bus attributes for those who only took auto but not bus in the same O-D pairs. Again, samples may be small or not exist for many O-D pairs, thus leaving the analyst to amend the approach with exogenous information. A fourth approach involves the use of stated non-chosen attribute values, based on the notion that capturing perceptual data yields realistic behavioral models that represent the choice context in which choices are made. The limitations of the stated attribute approach include increased respondent load, and the significant challenges associated with forecasting future perceptions needed to support transportation planning activities. For example, how does one forecast what high occupancy vehicle travel time will be perceived to be across individuals in the future?

A proposed fifth approach presented in this paper is to synthesize the data using known exogenous information such as travel distances or other sociodemographic characteristics and to condition the synthesized data on these constraints. The Bayesian Imputation Multinomial Logit (BI-MNL) method, described in this paper, incorporates elements of Bayes' theorem, the multinomial logit choice model, and sampling-based estimation to synthesize or impute unobserved data. The approach is motivated by a desire to (1) reduce the reliance of model calibration on network skim values and their noted limitations; (2) to obtain within-zone variation in unobserved attributes, such as travel times by walking, which may be important to assess policies and programs that may differentially impact travelers within a zone; and (3) to provide a robust analytical alternative to imputation-based approaches. The decision to use this approach, compared with say using network skims, would be influenced by the intended use of such data, the technical capabilities of modeling staff, and the benefit to the user of within-zone accuracy improvements.

Thus, the aim of this research is to develop and demonstrate the feasibility and advantages of imputing non-chosen travel mode attribute values using a relatively small calibration sample of complete data. An explicit advantage of this approach over the previously described approaches is the ability to impute within-zone variation in unobserved mode attributes. To estimate the BI-MNL model, a calibration sample is exploited to determine priors and constraints. Three types of information derived from the calibration sample are exploited: descriptive statistics of chosen and non-chosen modes; classical MNL coefficients and marginal rates of substitution, and observed correlation among the mode choice attributes. The proposed approach offers several advantages over alternative approaches, which serve as motives for the approach: (1) it captures within TAZ variation in unobserved mode attributes; (2) unobserved mode attributes can be forecasted; (3) it can be accomplished using a relatively small calibration data; (4) it does not require an increase in survey respondent load; (5) the approach is relatively inexpensive compared with costs of collecting data; and (6) it may be more accurate than alternative approaches, as it takes into account the non-ignorable nature of the missing data (see following section) and captures within-zone attribute value variation.

Figure 1 provides a flowchart of the BI-MNL model to serve the reader throughout the ensuing discussion. The solid boxes on the left of the figure represent the major activities that would be

DOI: 10.1002/atr

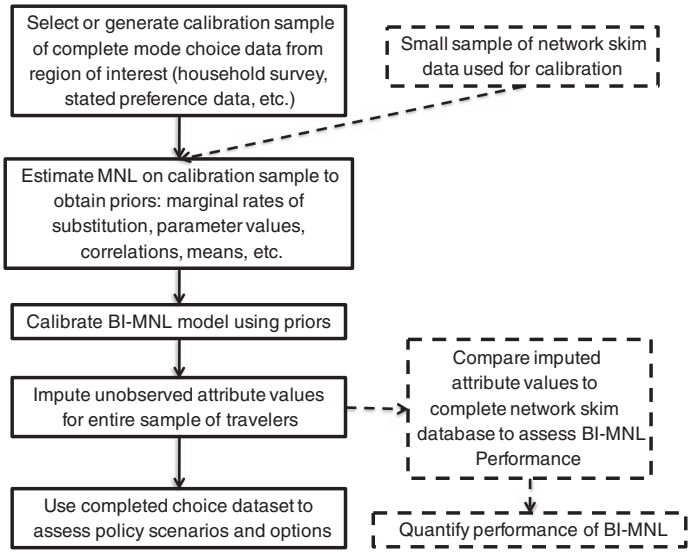

Figure 1. Flowchart of Bayesian Imputation Multinomial Logit (BI-MNL) model implementation steps for practice (solid boxes and arrows) and BI-MNL model development and testing steps (dashed boxes and arrows).

undertaken in practice to implement the BI-MNL for forecasting purposes, whereas the supplemental dashed boxes on the right identify activities described in the paper to develop, demonstrate, and test the BI-MNL. Thus, the activities described on the left represent intended end-user activities, whereas the activities on the right represent additional activities the authors undertook to demonstrate and test the viability of the BI-MNL model.

In addition, six key concepts are presented throughout the remainder of this paper with the intent to highlight fundamental ideas and objectives of the BI-MNL model and to increase the accessibility of the material to a broad audience.

2. MISSING DATA PROBLEMS AND IMPUTATION

Missing data problems are not new and have captured the attention of statisticians for some time. In general, data are said to be missing completely at random (MCAR) when "missingness" (a word adopted by statisticians to denote the mechanism responsible for an observation's omission) is not a function of observed or unobserved variables, missing at random (MAR) when missing data are functions of observed variables (exogenous), and non-ignorable when missing data are functions of unobserved (endogenous) variables [4]. When data are MAR, Rubin [4] showed that likelihood-based inference does not *require* a model for the missing data mechanism. With several continuous independent variables, one or more subject to incomplete observation, a commonly used imputation model is the multivariate normal under a MAR assumption [5].

Imputing or synthesizing values of non-chosen mode choice attributes is a missing data problem. In contrast to MCAR and MAR, missing travel mode attributes are a function of their unobserved status and non-ignorable. For instance, unobserved bus travel time may be because of an unobserved preference for the auto, the lack of knowledge of a nearby bus stop, or an unobserved need to link trip purposes—all of which are unobserved (presumably). For non-ignorable missingness mechanisms, it is necessary to set up a model for the response mechanism as well as for the data themselves [6]. For example, in a mode choice data set, a missingness model of non-ignorable missing travel time for a non-chosen mode (i') may be a function of the chosen mode travel time, ratio of travel time of non-chosen mode to chosen mode for other travelers, etc. It is quite reasonable to expect that this missingness model may produce data that are consistent with perceived attribute values, especially if a defensible model of the missingness mechanism is established. In keeping, the primary objective of this research is to establish a multivariate model of missingness for unobserved mode choice attribute values, with the intent to reliably estimate missing values.

DOI: 10.1002/atr

Research on missing data problems in travel time studies and travel behavior is relatively rare. Datla *et al*. [7] applied k-nearest neighbor nonparametric regression to impute holiday traffic, whereas Wang *et al*. [8] applied multiple imputation and nearest-neighbor methods to estimate travel times. Zhao and Chung [9] emphasize the need to develop data imputation techniques to support a variety of land use and travel behavior forecasting models. Jou *et al*. [2] applied the Deming-Stephan proportional fitting procedure to overcome problems with missing data in O-D matrices—an application that emphasizes the problem with using network skims from O-D matrices to impute non-chosen attribute values, as discussed previously. Steimetz and Browstone [10] illustrated the value of multiple imputation compared with single imputation approaches by examining the value of commuters' travel time savings (VTTS). They used a conditional logit mode choice model to examine the VTTS on mode choice; however, their approach applied the choice model to examine VTTS, not to synthesize or impute data, as is the objective of this research. Brownstone [3], using a choice model of occupational code (job changes), illustrates an example of multiple imputation of job changes. Finally, Brownstone and Golob [11] used multiple imputation to synthesize the number of employees at worksites (which was 30% unreported in the sample) needed to estimate reliable choice models of drive alone and carpool behavior. As such, merging the concepts of multiple imputation, missing data, and choice modeling has been applied by relatively few researchers, albeit for solving important missing data problems, leaving considerable room for advances in this growing area of interest.

3. THE BAYESIAN IMPUTATION MULTINOMIAL LOGIT MODEL

The BI-MNL model developed in this section is unique and directly addresses a missing data problem in travel behavior research.

Key Concept #1: The justification or motivation for the BI-MNL model is to obtain reliable (posterior) estimates of the values of the "missing" or unobserved mode attributes by exploiting *a priori* information on a subset of travelers regarding mode choices and characteristics.

The BI-MNL is implemented using Bayes' theorem (also Bayes' rule), which combines current sample data with prior information regarding parameter values to obtain posterior probabilities of parameter values. In mathematical form, Bayes' theorem relates the posterior probability of parameters taking on specific values given data $P(\boldsymbol{\beta}|\text{data})$ to the product of the likelihood $P(\text{data}|\boldsymbol{\beta})$ and prior $P(\boldsymbol{\beta})$ divided by the marginal probability of the data.

$$P(\boldsymbol{\beta}|\text{data}) = P(\text{data}|\boldsymbol{\beta})P(\boldsymbol{\beta})/P(\text{data}) \tag{1}$$

In the BI-MNL model, the *data* refer to observed and unobserved mode attributes and choices, whereas the parameters $\boldsymbol{\beta}$ correspond with a variety of parameters of interest such as mean travel times and marginal rates of substitution. Prior to describing the BI-MNL in greater detail, it is first necessary to examine the classical MNL, which lies at its core.

Predictors of choice for mode choice set $i = 1, 2, \ldots, J$ in the MNL model may include social and demographic characteristics of individuals $n = 1, 2, 3, \ldots, M$, $\boldsymbol{X}_n$ (such as age, gender, household income), attributes of alternative i, $\boldsymbol{A}_i$ (such as the seating capacity of a bus), and attributes of an alternative specific to an individual $\boldsymbol{C}_{in}$ (such as travel cost for mode i and traveler n). Traveler utility functions consist of deterministic and random components of utility or attractiveness of alternatives,

$$U_{in} = V_{in} + \varepsilon_{in} = \boldsymbol{X}_n\boldsymbol{\beta} + \boldsymbol{A}_i\varphi + \boldsymbol{C}_{in}\boldsymbol{\gamma} + \varepsilon_{in} \tag{2}$$

where U_{in} is the total utility of alternative i and chooser n, and $V_{in} = \boldsymbol{X}_n\boldsymbol{\beta} + \boldsymbol{A}_i\boldsymbol{\varphi} + \boldsymbol{C}_{in}\boldsymbol{\gamma}$ reflects the deterministic aspects of utility measured by vectors $\boldsymbol{X}_n$, $\boldsymbol{C}_{in}$, and $\boldsymbol{A}_i$. A traveler's or chooser's

DOI: 10.1002/atr

selection of travel mode, $\mathbf{Y}_{in}$, is said to correspond with a maximum utility decision with $U_{in} = \max(U_{1n}, U_{2n}, \ldots, U_{Jn})$, whereas the ε_{in} represent the random unobserved components of utility.

The BI-MNL model is developed by exploiting the choice structure of the MNL model, by placing priors on the estimated coefficient vectors $\boldsymbol{\beta}$, $\boldsymbol{\varphi}$, and $\boldsymbol{\gamma}$ during estimation, and by drawing multiple samples (imputations) using Markov chain Monte Carlo (MCMC) sampling. Important in BI-MNL model specification, recall that mode choice alternatives are indexed $i = 1, 2, \ldots, J$. Suppose chooser n selects mode $i = k$ (say bus), then a subset of all possible modes consisting of non-chosen modes exists such that $I' = 1, \ldots, k-1, k+1, \ldots, J$, where i' represents the index over which the chosen travel mode is excluded, $i \neq k$ (for convenience throughout, it is assumed that chooser n makes one mode choice $i = k$; however, it is recognized that more than one mode can be chosen and easily accommodated with enhancements to notation). As such, $\boldsymbol{A}_{i'}$ and $\boldsymbol{C}_{i'n}$ are vectors of unobserved alternative and individual-specific alternative attributes, respectively.

With the necessary notation and fundamentals in place, a stark contrast can be drawn between the traditional MNL model used to model choice outcomes and the BI-MNL model used to impute missing data values in a choice data set.

Key Concept #2: The classical MNL model is used to predict or explain choice outcome U_{in} by making inferences on parameter vectors $\boldsymbol{\beta}$, $\boldsymbol{\varphi}$, and $\boldsymbol{\gamma}$ conditioned on chooser attributes $\boldsymbol{X}_n$, alternative attributes $\boldsymbol{A}_i$, and chooser-alternative attributes $\boldsymbol{C}_{in}$. Interest is focused on the roles of the $\boldsymbol{X}_n$, $\boldsymbol{A}_i$, and $\boldsymbol{C}_{in}$ (through their respective coefficients) in explaining or predicting observed choice behavior.The BI-MNL model is used to impute non-chosen and thus unobserved attribute values $\boldsymbol{A}_{i'n}$ and $\boldsymbol{C}_{i'n}$ conditioned on parameter vectors $\boldsymbol{\beta}$, $\boldsymbol{\varphi}$, and $\boldsymbol{\gamma}$, which are assigned informative priors based on observed data $\boldsymbol{X}_n$, $\boldsymbol{A}_i$, $\boldsymbol{C}_{in}$, and mode choice $\boldsymbol{Y}_{in}$. Inference is focused on imputing missing values of the unobserved data in the choice process.

Unobserved attributes in the BI-MNL model are obtained by assigning informative priors, which are derived from exogenous data. The model is estimated using MCMC methods or sampling-based estimation (see [12], 2002, [13]). Sampling-based estimation requires that the model and priors are structured in terms of well-defined sampling distributions.

The generic structure of the MCMC BI-MNL model is given as follows:

$$Y_{in} \sim \text{Multinomial}(P_{1n}, P_{2n}, \cdots, P_{Jn}) \tag{3}$$

$$P_{in} = \frac{e^{V_{in}}}{\sum_{t=1}^{J} e^{V_{in}}} \tag{4}$$

$$\sum_{t=1}^{J} e^{V_{in}} = e^{V_{kn}} + \sum_{t=1}^{J(i\neq k)} e^{V_{in}} \tag{5}$$

$$V_{in} = \boldsymbol{X}_n\boldsymbol{\beta} + \boldsymbol{A}_i\boldsymbol{\varphi} + \boldsymbol{C}_{in}\boldsymbol{\gamma} \tag{6}$$

where Y_{in} is multinomial distributed and P_{in} is the probability that choice i is selected by chooser n, and all other terms are as defined previously. Deterministic utility is a function of observed attributes of the chosen mode and non-observed attributes, as shown in Equation (5). Parameter vectors $\boldsymbol{\varphi}$, and $\boldsymbol{\gamma}$ can be estimated either by mode (e.g., a cost coefficient for J-1 modes) or as generic (see [1,14]).

Equations (3) through (5) represent the general BI-MNL model form. That is, data generated by the BI-MNL model are consistent with a MNL-modeled choice process. The use of Bayesian priors, importantly, is key to both BI-MNL model formulation and understanding, and thus requires careful explanation and illustration.

DOI: 10.1002/atr

Key Concept #3: Bayesian priors are used to constrain and influence imputed choice attribute values. For the BI-MNL model to be useful in practice, priors are extracted from statistics and modeling results from a calibration sample, which may be a prior survey (e.g., a prior household travel survey in the region), or supplemental traveler interviews, surveys, etc. The priors obtained from the calibration sample include marginal rates of substitution, estimated coefficients, and simple descriptive statistics such as means and ratios. In the proof of concept provided herein, the calibration sample is a subsample of network skim data—suboptimal in practice because of the averaging within zones and large variances as a function of zone size—but sufficient to demonstrate the BI-MNLs ability to replicate calibration data. This research leaves as future discussion the relative merits of various data sets for BI-MNL model calibration.

Priors needed for the BI-MNL model include marginal rates of substitution and estimated coefficients from a classical MNL model estimated on the calibration sample, in addition to descriptive statistics on the calibration sample. The marginal rate of substitution π_{RS} between variables R and S is estimated using the MNL model and used as a prior in the BI-MNL model, such that

$$\pi_{RS} \sim N\left[\frac{\widehat{\theta_R}}{\theta_S}, (\hat{\sigma}_{\theta_{RS}})\right] \tag{7}$$

where N denotes the normal distribution, $\frac{\widehat{\theta_R}}{\theta_S}$ is the mean ratio of estimated parameters of variables R and S, and $\hat{\sigma}_{\theta_{RS}}$ is the estimated standard error of $\frac{\theta_R}{\theta_S}$. Marginal rates of substitution represent average utility tradeoffs among predictor variables found in $\boldsymbol{X}_n$, $\boldsymbol{A}_i$, and $\boldsymbol{C}_{in}$. Their use is predicated on the notion that marginal rates of substitution reflect the average willingness of consumers to trade off one attribute for another in the sample, and that preserving these utility tradeoffs is also important in replicating data appropriately. For example, if the average traveler in the random calibration sample is willing to trade \$2.50 in order to avoid each 10 minutes of bus delay, then this information is important for imputing realistic bus travel times and fares and provides useful information on the relative utility of travel cost and bus wait time.

It is critical to note, however, that the normality of the ratio $\frac{\widehat{\theta_R}}{\theta_S}$ is an assumption that is true only in limited circumstances, and often, this ratio is ill-defined within infinite variance. For example, Marsaglia [15] showed that despite that no theoretical moments exist, many ratios of normal variates in practice can themselves be taken as approximately normally distributed. He further shows that when the denominator in the ratio is not expected to approach zero and certain reasonable conditions are imposed, the ratios are indeed approximately normal. Daly *et al*. [16] and Marsaglia [15] highlight that when the dominator is expected to approach zero, the distribution of $\frac{\widehat{\theta_R}}{\theta_S}$ is not well behaved and has no moments. One solution Daly *et al*. propose, although not optimal for various reasons described in their paper, is to arbitrarily truncate or censor the denominator to exclude zero such that a variance can be defined. This is essentially the approach taken in this research, whereby the variances of the marginal rates of substitution are arbitrarily constrained to be finite.

Parameters from the MNL model are also used as priors in the BI-MNL model, such that a prior is placed on the parameter for variable R such that

$$\theta_R \sim N\left[\hat{\theta}_R, (\hat{\sigma}_{\theta_R})\right] \tag{8}$$

where $\hat{\theta}_R$ is the estimated mean parameter value associated with variable R and $\hat{\sigma}_{\theta_R}$ is the estimated standard deviation of θ_R. As is explained later, judicious use of constraints on absolute values of parameters is necessary because of the ambiguity of absolute parameter values but necessary for BI-MNL model calibration.

DOI: 10.1002/atr

Priors for imputed attribute values are also derived from descriptive statistics on the calibration sample. To illustrate, suppose chooser n is observed taking auto and not observed taking the bus, and so bus travel time $\left(TT_{i=2',n}\right)$ is imputed (note the prime here refers to a single non-chosen mode, in contrast to the prime on i applied previously that refers to the index over all non-chosen travel modes). Priors derived from the calibration sample are placed on the unobserved auto travel time sampling distribution by sampling from a normal distribution with mean equal to the observed auto travel time ($TT_{i=1,n}$) or traveler n multiplied by the ratio of average travel times for non-chosen bus and chosen auto across travelers in the calibration sample. In equation form,

$$TT_{i=2',n} \sim N\left[TT_{i=1,n} \cdot \hat{\phi}_{TT_{z':1}}, \left(\hat{\tau}_{TT_z}\right)\right] \tag{9}$$

where,

$$\hat{\phi}_{TT_{z':1}} = \frac{\sum_{n=1}^{M} TT_{i=z',n}/M}{\sum_{n=1}^{M} TT_{1,n}/M} \tag{10}$$

$$\hat{\tau}_{TT_z} = \sqrt{\frac{\sum_{n=1}^{M}\left(TT_{i=z',n} - \hat{\mu}_{TT_{i=z'}}\right)^z}{M-1}} \tag{11}$$

$$\hat{\mu}_{TT_{i=z'}} = \sum_{n=1}^{M}\left(TT_{i=2',n}\right)/M \tag{12}$$

The statistics $\hat{\phi}_{TT_{z':1}}$, $\hat{\tau}_{TT_z}$ and $\hat{\mu}_{TT_z}$ are derived from the calibration sample, where both observed and unobserved attribute values are known. To illustrate numerically, suppose a traveler is observed taking 60 minutes to drive an auto to work. His or her bus travel time is unknown and imputed. Suppose further that in a random calibration sample of travelers in the region, the average travel time for observed auto users was 30 minutes, whereas the average travel time for people declining to take the bus was 35 minutes, or $\hat{\phi}_{TT_{z':1}} = 35/30$. Suppose that the standard deviation of bus travel times of people in the calibration sample who declined to take the bus was $\hat{\tau}_{TT_z} = 10$. From this, a prior is placed on imputed bus travel time for this traveler such that $TT_{i=2',n} \sim N[60 \cdot 35/30, (10)]$.

The final prior in the BI-MNL model is used to influence the correlation between imputed attributes because attribute values in practice are correlated (e.g., consider unobserved time and distance attributes). The covariance matrix is the obvious candidate to obtain priors for multiple imputed unobserved attributes. It is quite likely, however, to over-identify the model system by applying too many constraints on covariances, resulting in a BI-MNL model that does not converge.

To avoid model identifiability problems, an alternative approach to specifying a prior on the full covariance matrix is to develop regression equations between variables. Using this approach, a regression equation is developed between times and distances, which are correlated in the calibration sample. Using a single regression equation reduces the risk of specifying an un-identifiable model system but fails to replicate all covariances in the calibration sample (e.g., imputed travel time and household income). In the regression approach, traveler ns unobserved value of attribute R (e.g., travel time) is a linear (quadratic, nonlinear, etc.) function of the unobserved value of attribute S (e.g., travel cost), such that

$$R_n = \alpha + \delta S_n + \varepsilon_n \tag{13}$$

In Equation (13), α and δ are estimated empirically using the calibration sample, and ε_n is assumed to be normally distributed across choosers. Applying this equation to influence the relationship between imputed attributes requires priors on α, δ, and ε such that correlation in the calibration sample is preserved.

DOI: 10.1002/atr

Key Concept #4: The priors described in this section collectively constrain and influence the BI-MNL model such that imputed unobserved attribute values follow distributions with parameters similar to those of the calibration sample, corrected for their unobserved status and for correlation present in the observed calibration data.

The BI-MNL model developed and described in this research is implemented using MCMC sampling based estimation and the Gibbs sampler. Sampling-based estimation differs from maximum likelihood estimation in the following way. Monte Carlo methods use computational algorithms that generate repeated random (or pseudo-random) samples from known distributions, such as the multivariate normal or binomial. Statistics such as parameter estimates are computed based on the samples generated rather than solving a function, such as least squares or maximum likelihood. The task for MCMC-based estimation, then, is to construct a statistical model such that the appropriate sampling distributions are specified, can be sampled, and then statistics from them obtained.

The MCMC estimation yields multiple imputations of all parameters obtained from the BI-MNL model; however, only the means of parameters are reported here. MCMC methods also require assessment of model convergence, model fit, and model identifiability as described in the study conducted by Congdon [12,17,13]. Further details of these issues and the MCMC approach are omitted for brevity; however, these issues were addressed as is done in standard practice. Readers wishing additional details on Bayesian approaches and estimation considerations should consult Gelman *et al*. [18], those wishing additional details on MCMC methods should consult Gilks *et al*. [19] and Brooks and Roberts [20], and those interested on additional details on the Gibbs sampler should consult Robert and Casella [21].

Finally, it should be noted that other model forms could be used to structure prediction of choice outcomes, such as the nested and mixed logit models; however, the MNL model is known for performing well when errors are identically and independently distributed. The alternative models are more complex and left for potential future research.

4. DATA DESCRIPTION AND IDENTIFICATION OF BAYESIAN PRIORS

The Maricopa Regional Household Travel Survey (MRHTS) data from 2001 are used to develop and illustrate the BI-MNL model described previously [22]. The MRHTS is a revealed choice survey of households in Maricopa County, Arizona, and consists of 4018 households and 10 030 household members within Maricopa County and a small portion of Pinal County that contains the city of Apache Junction. To facilitate this analysis, a subsample of data was used on 1587 household members observed during their work commute with four mode choices: single occupant vehicle (SOV), high occupant vehicle (HOV), transit, and non-motorized (NM) modes.

Key Concept #5: The BI-MNL model is demonstrated using the MRHTS data set. It is a data set of 1587 household member commute trips, with unobserved mode choice attributes obtained from network skim values (see Introduction). The validation exercise implemented here employs a random calibration sample of $M = 517$ (about a third) to obtain priors and then applies the BI-MNL model to predict non-chosen modes for the remaining sample of 1070 household members (where the unobserved attribute values are *assumed to be missing*). Because the unobserved mode attributes of these 1070 household members are actually known, an assessment of how well the BI-MNL model performs at imputed attribute values is conducted. In practice, however, the analyst would apply the BI-MNL model on a data set where unobserved attribute values are not known with intent to forecast current or future values of unobserved mode attributes. The focus here is not on the data themselves but the methodology developed to impute non-chosen mode attributes.

Priors needed for the BI-MNL model are those that influence the relationship between travel choices and choice trade-offs of the population, as described previously. To obtain these priors, a classical MNL model is estimated using the calibration sample described previously. Descriptive statistics of the MRHTS calibration sample are shown in Table I. Travel times in minutes range between a

DOI: 10.1002/atr

Table I. Descriptive statistics of 517 working household members from Maricopa Regional Household Travel Survey.

Variable name	Minimum	Maximum	Mean	Standard deviation
Age of the household member	16.000	99.000	41.220	12.879
Number of vehicles in household	0.000	7.000	1.895	0.938
Number of bicycles in household	0.000	8.000	1.119	1.393
Household size	1.000	8.000	2.823	1.442
Number of workers in household	1.000	5.000	1.793	0.744
Single occupant vehicle travel time (minutes)	2.509	99.420	18.946	10.378
High occupant vehicle travel time (minutes)	2.509	98.810	18.118	9.468
Transit travel time (minutes)	1.500	155.700	33.295	22.713
Non-motorized travel time (minutes)	0.821	176.905	25.564	34.636
Commute distance (miles)	0.110	81.780	10.617	8.358

minimum of less than a minute (NM) to more than 175 minutes (transit), whereas commute distances ranged from 1/10 of a mile to over 81 miles.

The MNL model parameter estimates based on the calibration sample are shown in Table II. The objective of this MNL model is to yield parameters that will serve as priors in the BI-MNL model. Because travel times and distances will be imputed for all modes and for a variety of travelers, an effort is made to include variables in the MNL model that reflect all modes and a wide range of traveler characteristics. Moreover, type I risk threshold was set to 15% ($\alpha = 0.15, t = 1.42$) because of the intended use of this model—to predict unobserved values. In other words, policy decisions will not be based on these parameters, but instead, they are used to assist in the prediction of unobserved behavior—thus arguing for a lenient alpha. As shown in the table, all coefficient signs are plausible and make reasonable sense in the travel choice context. As it turns out, most *t*-ratios are 2 or greater.

The bolded variables shown in Table II were applied as priors in the BI-MNL model. Not all estimated coefficients are used as priors; however, many are used so that imputed attribute values are consistent with the choice preferences reflected in the calibration sample. All priors on MNL coefficients and marginal rates of substitution were assumed to be normally distributed. Asymptotic theory suggests that coefficients of MNL models are asymptotically normal, whereas the ratio of two normally distributed variables is more unpredictable—sometimes being approximately normal

Table II. Multinomial logit model results for 517 household member calibration sample.

Variable description	Parameter estimate	Standard error
Alternative specific constant for single occupant vehicle	0.4431	0.7916
Alternative specific constant for high occupant vehicle	−1.5040	1.1240
Alternative specific constant for Transit	−0.0476	1.5840
Number of vehicles in household, specific to single occupant vehicle	2.1872	0.4120
Number of vehicles in household, specific to high occupant vehicle	1.0770	0.4650
Single occupant vehicle travel time (minutes)	−0.0778	0.0409
High occupant vehicle travel time (minutes)	−0.0855	0.0530
Transit travel time (minutes)	−0.1340	0.0590
Number of workers in household, specific to high occupant vehicle	0.9100	0.3180
Number of workers in household, specific to Transit	1.0590	0.4780
White ethnicity (indicator), specific to high occupant vehicle	−0.8760	0.4850
Low income status (indicator), specific to Transit	1.1200	0.6680
Commute distance, specific to non-motorized	−0.2120	0.0890
Number of bikes in household, specific to non-motorized	0.4540	0.2290
***Single occupant vehicle travel time/high occupant vehicle travel time**	0.909	|0.909|/20
***Transit travel time/non-motorized commute distance**	0.632	|0.632|/20
***Number of vehicles (high occupant vehicle)/number of workers (Transit)**	1.010	|1.010|/20
***Number of bikes (non-motorized modes)/number of workers (high occupant vehicle)**	0.498	|0.498|/20
***ASC Alternative specific constant single occupant vehicle/ASC Transit**	−9.036	|-9.036|/20

*These parameters are the sampled ratios of two model parameters.

DOI: 10.1002/atr

distributed [14], whereas other times being ill defined and without finite moments [14,15]. As described later, steps are taken to prevent ill-behaved distributes of marginal rates of substitution.

Because the objective of the BI-MNL model is to predict realistic Xs consistent with observed choice behavior and inherent variation in these Xs, calibration of model error is necessary so that unexplained choice behavior in the sample is preserved. As such, the analyst may not have control over what variables need to be imputed, and therefore must develop a BI-MNL model that exploits the existing data and choice relationship within the data—resulting in constraints on both the deterministic relationship between variables and the stochastic nature of decision making in the sample.

The most straightforward calibration metric to preserve choice certainty for a fixed data set and model specification is the log likelihood. To assist in model calibration, the standard errors of the priors used in the BI-MNL are adjusted as needed to produce a MNL model with equivalent log likelihood as observed in the calibration sample. In other words, the imputed choice data should reflect the same level of certainty as the calibration data. Relatively smaller standard errors of parameters suggest greater certainty in choice, whereas relatively larger standard errors suggest less certainty in choice (with respect to a specific variable). In the specification here, the log-likelihood value for the MNL model with imputed data was −453.56, whereas the log-likelihood of the MNL model with skim data was −468.86. The slight improvement of the BI- MNL model is justified because imputed *time* and *distance* vary within a TAZ, whereas *time* and *distance* determined via network skim values are constant within a TAZ (e.g., everyone living within a TAZ traveling to another TAZ is assumed to have the same travel time if taking a bus) thus increasing error and reducing fit.

When imputed *times* and *distances* are more consistent with observed choices in the calibration sample, a smaller log-likelihood is obtained. Conversely, if imputed *times* and *distances* are less consistent with observed choices (than observed in the calibration sample) then standard errors are too large and must be reduced to properly calibrate the BI-MNL log likelihood. This model calibration process is iterative much like all statistical modeling, where the "best fit" model is obtained through trial and error.

Additional priors needed in the BI-MNL model are those that influence the distributions of imputed values—as it is undesirable to predict travel times and distances outside the ranges observed in the exogenous sample. These Bayesian priors are extracted from the descriptive statistics for the chosen and the non-chosen alternatives of household members in the calibration sample.

Descriptive statistics of the calibration sample of 517 household members are shown for chosen and non-chosen alternatives in Table III. The tables show that chosen and non-chosen attribute values

Table III. Descriptive statistics of chosen and non-chosen modes.

Chosen mode attributes				
Variable name	Minimum	Maximum	μ_c	Standard deviation
Non-motorized travel time	0.823	177.278	20.645	35.668
Single occupant vehicle travel time	2.635	59.920	15.870	9.531
High occupant vehicle travel time	3.214	38.150	17.653	10.482
Transit travel time	14.344	37.000	23.319	6.035
Non-motorized commute distance	0.130	28.010	3.262	5.636
Single occupant vehicle commute distance	0.255	53.280	10.620	7.786
High occupant vehicle commute distance	0.405	29.590	10.800	9.708
Transit commute distance	0.540	50.480	13.028	12.516
	Minimum	Maximum	μ_{nc}	Standard deviation
Non-motorized travel time	1.614	337.215	67.810	51.244
Single occupant vehicle travel time	2.612	55.890	18.840	12.209
High occupant vehicle travel time	2.612	56.110	17.715	9.053
Transit travel time	11.900	118.000	29.418	9.571
Non-motorized commute distance	0.255	53.280	10.714	8.097
Single occupant vehicle commute distance	0.130	50.480	8.570	10.039
High occupant vehicle commute distance	0.130	53.280	10.313	8.064
Transit commute distance	0.130	53.280	10.248	7.965

c indicates chosen; nc, non-chosen.

DOI: 10.1002/atr

differ, as expected. Note that these are not paired comparisons—those who took transit in the sample had longer average transit distances (13.28 miles) than those who rejected this mode (10.28 miles).

In practice, the analyst will not know the absolute values of chosen and non-chosen parameter values *a priori*, and of course, MNL models are postulated on utility differences. As a result, the values in Table III are not ideal for informing priors in the BI-MNL model. Instead, Table IV lists the ratio of chosen to non-chosen attribute values from the calibration sample. These ratios are more transferable to a future sample of choosers with unknown absolute values of attributes. For example, it is easier to defend a 1.187 prior on the ratio of imputed chosen to non-chosen SOV travel times than to place a 15.87-minute prior on the mean of chosen SOV travel times. Ratios are also more consistent with utility theory, as it is the relative attractiveness of alternatives that is important in choice making. As a result, the phi values in Table IV are used to inform priors in the BI-MNL model. These prior values suggest that the observed distributions of *times* and *distances* for chosen alternatives are multiplied by a constant (with some variation) to obtain the distributions of non-chosen *times* and *distances*.

The final priors needed to calibrate the BI-MNL model are used to influence the relationship between imputed *time* and *distance*. This is necessary—especially in this specific case—because *time* and *distance* are correlated in the true population, and ignoring this correlation will result in unrealistic imputed values. In fact, when this is ignored in the imputation model, *time* and *distance* are nearly orthogonal. Recall that an analyst may not have influence over what attributes are being collected in a travel survey, and so correlated attributes are certainly possible, even though they are less than ideal and can create challenges in analysis and modeling.

The empirical relationships between observed *time* and *distance* across choosers in the sample for the travel NM, SOV, HOV, and transit modes are shown in Figure 2. In estimating this BI-MNL model, (travel) *time* is modeled as a function of (travel) *distance* (see Equation (14)). The four empirically derived equations corresponding to priors assigned to NM, SOV, HOV, and transit mode *times* are given by:

$$\text{time}_{\text{NM}} = 6.329(\text{dist}_{\text{NM}}) \tag{14}$$

$$\text{time}_{\text{SOV}} = 0.0008(\text{dist}_{\text{SOV}})^3 - 0.067(\text{dist}_{\text{SOV}})^2 + 2.567(\text{dist}_{\text{SOV}}) + 0.144 \tag{15}$$

$$\text{time}_{\text{HOV}} = 0.0009(\text{dist}_{\text{HOV}})^3 - 0.0757(\text{dist}_{\text{HOV}})^2 + 2.577(\text{dist}_{\text{HOV}}) + 0.1497 \tag{16}$$

$$\text{time}_{\text{Transit}} = 0.008(\text{dist}_{\text{Transit}})^3 - 0.425(\text{dist}_{\text{Transit}})^2 + 6.795(\text{dist}_{\text{Transit}}) + 0.3053 \tag{17}$$

These polynomial expressions are used to assign priors on imputed *time* across modes. These empirically derived forms were chosen over perhaps more theoretically motivated relationships for

Table IV. Ratio priors (phi) used to inform the Bayesian Imputation Multinomial Logit.

Variable name	μ_c	μ_{nc}	phi = μ_{nc}/μ_c
Non-motorized travel time	20.645	67.810	3.285
Single occupant vehicle travel time	15.870	18.840	1.187
High occupant vehicle travel time	17.653	17.715	1.004
Transit travel time	23.319	29.418	1.262
Non-motorized commute distance	3.262	10.714	3.284
Single occupant vehicle commute distance	10.620	8.570	0.807
High occupant vehicle commute distance	10.800	10.313	0.955
Transit commute distance	13.028	10.248	0.787

c indicates chosen; nc, non-chosen.

DOI: 10.1002/atr

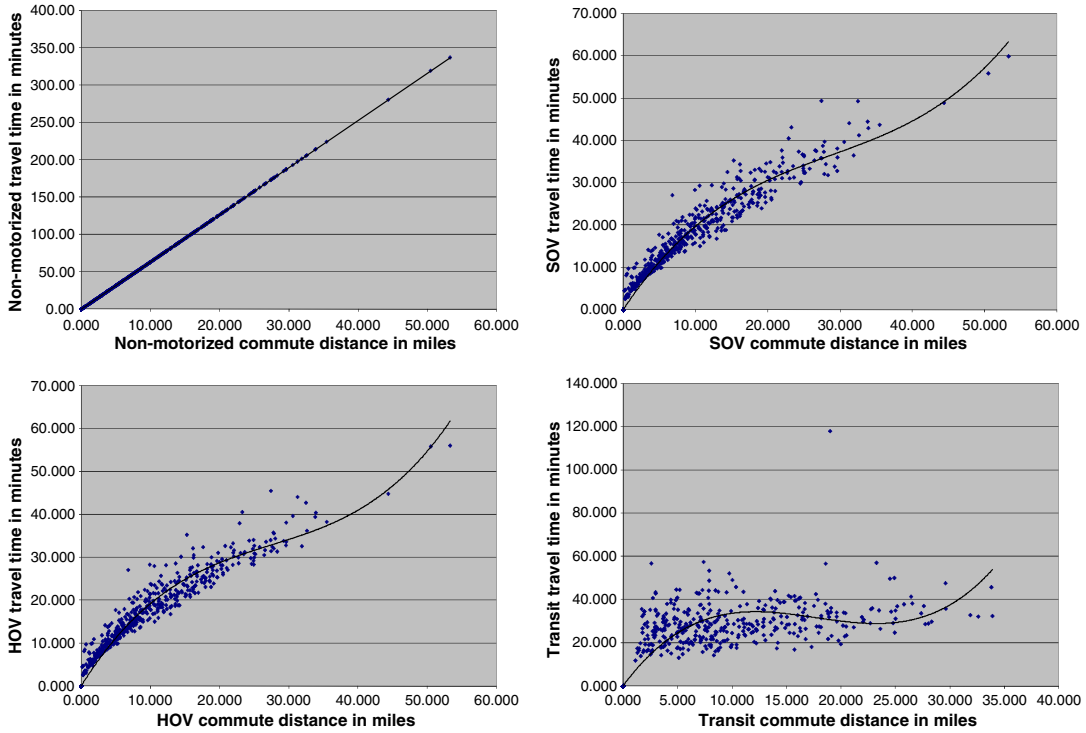

Figure 2. Travel time versus commute distance in calibration sample: non-motorized (upper left), single occupant vehicle (SOV; upper right), high occupant vehicle (HOV; lower left), and transit (lower right).

several reasons. First, nonlinearities, although slight in some cases, were observed for *SOV*, *HOV*, and *Transit* modes—and thus, the cubic specifications seem more empirically defensible than linear ones over the range of observed values. Moreover, the focus is not on making inferences about these relationships but instead on predicting time given distance. Finally, *SOV*, *HOV*, and *Transit* modes within a region will not serve travelers "door to door" within a region (e.g., transit may cover 70% of a trip distance), and thus, distance and time are likely to be nonlinear for these modes in practice. Applying these empirical relationships within the BI-MNL, once a non-chosen *distance* is imputed (using a prior on the ratio of chosen *distance* described previously), *time* is imputed as a function of distance. When this approach is taken, the priors on ratios for *time* are not needed in the BI-MNL model. Alternatively, Equations (15)–(17) could have been estimated to relate *distance* as a function of *time*, and the priors on ratios for *distance* would not be needed in the BI-MNL model. In Equations (15)–(17), priors are placed on the equation coefficients and their standard errors. As was needed in model calibration previously, standard errors influence the observed variation in *time* values as a function of *distance* and are adjusted so that imputed variation matches the variation in the calibration sample.

The calibrated BI-MNL model specification results in the ability to impute values based on random draws from the conditional posterior distribution of missing variables given the observed variables and model parameters with informative priors as described previously. The mean values obtained from this multiple imputation Bayesian MNL model are assessed in the following section.

5. BAYESIAN IMPUTATION MULTINOMIAL LOGIT MODELING RESULTS

It is important to understand how the model is assessed. First, the BI-MNL model is calibrated using an exogenous calibration sample, as described previously. Calibration is achieved when the BI-MNL model log-likelihood with imputed data is similar to the log-likelihood obtained using the calibration sample (using the same sample size and model specification). The remaining sample of 1070 households—the validation sample used to demonstrate this methodology—includes *time* and *distance* values obtained from network model skims. Network skim values were used because they were

DOI: 10.1002/atr

available. This is a known limitation of this approach; however, the methodology is the main contribution, and this limitation is discussed later.

Key Concept #6: The validation sample was created by deleting all non-chosen time and distance values (for which skims are actually known). The calibrated BI-MNL model is then used to impute time and distance for this validation sample. In the assessment, the two data sets are compared: one containing time and distance obtained from network skims and a data set containing time and distance imputed from the BI-MNL model. Recall that the BI-MNL model used to impute time and distance is calibrated entirely on the smaller random calibration sample of households.

The remainder of this section describes the results of the variety of different tests to assess the goodness of fit of imputed versus skim values of *time* and *distance*. To facilitate the discussion, a snapshot of the imputed data output from the BI-MNL model is first presented. Table V shows the imputed results for three respondents (chooser ID 11, 12, and 13). Alternatives are listed (1 through 4 corresponding to NM, SOV, HOV, and transit modes, respectively), as is the indicator for choice; thus, respondent 11 chose the SOV mode when surveyed. The BI-MNL model is used to impute *time* and *distance* for modes 1, 3, and 4. For example, the skim *time* for NM is 70 minutes, whereas the BI-MNL model imputed *time* is 48.66 minutes. For respondent 13 (who also chose SOV), the NM skim *time* is 56.32 minutes, whereas the imputed *time* is 63.63 minutes. Note that skim distances are the same across respondents because the skims use the network distance between origin and destination TAZ zone centroids.

The imputed *times* and *distances* from the BI-MNL model across subjects are evaluated in the remainder of this section. They are compared with skim *times* and *distances*, which are considered as "truth" throughout the comparisons. In discussions that follow, the "truth" data set refers to the complete set of data including skim derived *times* and *distances*, whereas the "imputed" data set refers to the same data set with imputed *times* and *distances*.

5.1. Correlation matrices

The choice data sets (truth and imputed) reveal correlations among variables. The correlation coefficient between imputed *time* and *distance* is 0.511, compared with the true correlation of 0.519—thus, correlation between these variables has been substantively preserved in the imputed data. The correlation of *time* and *distance* with other model variables is also replicated quite well, with all non-zero correlations reflected in the imputed data. Many of the observed correlations are near zero, as were imputed correlations, and thus, the model in no case produced a significant non-zero correlation in error.

5.2. Predicted travel time and commute distance

The average predicted travel *time* and commute *distance* for imputed and skim data sets are shown in Table VI. The *time* is measured in minutes, and *distance* is measured in miles. As demonstrated previously, the observed correlation between *time* and *distance* is preserved in the imputed data set.

Table V. Snapshot of skim and imputed *time* and *distance* values.

Respondent ID	Alternative	Choice	Mean imputed time	Mean imputed distance	Skim time	Skim distance
11	1	0	48.66	7.689	70	11.06
11	2	1	20.13	11.06	20.13	11.06
11	3	0	20.99	11.71	18.06	11.06
11	4	0	31.98	8.76	51.6	11.06
12	1	0	27.99	4.42	52.53	8.3
12	2	1	15.56	8.3	15.56	8.3
12	3	0	11.82	5.59	15.56	8.3
12	4	0	33.33	13.06	50.4	8.3
13	1	0	63.63	10.05	56.32	8.9
13	2	1	17.52	8.9	17.52	8.9
13	3	0	27.18	18.07	17.52	8.9
13	4	0	22.03	4.49	50.05	8.9

DOI: 10.1002/atr

Table VI. Average predicted travel time and commute distance.

Average travel time and distance	
Variable	Mean
Imputed data	
Average travel time (minutes)	35.935
Average distance (miles)	11.433
Skim data	
Average travel time (minutes)	32.953
Average distance (miles)	10.454

As seen in the table, the average imputed *time* is 35.9 minutes for the imputed data and 32.9 minutes for skim data, whereas average imputed *distance* is 11.4 miles compared with 10.4 miles.

5.3. Average and absolute difference in travel time and commute distance

The average difference for travel *time* and commute *distance* and the average absolute difference reveal information about the variability in imputed compared with skim values. The former shows the average difference where positive differences cancel with negatives, whereas the latter shows the absolute value of differences. Table VII shows these average and average absolute differences. The average differences are 2.9816 minutes and 0.9883 miles for *time* and *distance*, respectively, whereas the average absolute differences are 16.077 minutes and 5.547 miles. The suspected reasons for these differences are discussed in the next section.

The MNL models estimated on the skim and imputed data are used to reveal meaningful information about the viability and reasonableness of the imputed values produced by the BI-MNL model and provide the ability of imputed data to replicate observed choice behavior. Table VIII shows the MNL results for the skim and imputed data sets.

The MNL model results are quite similar for both data sets. Coefficient signs and magnitudes are similar across the models, and the models possess signs of coefficients as expected in practice. The travel time coefficient for HOV is larger than that for SOV, reflecting the greater disutility of travel time in a ride share situation. The significance of variables is similar across the models. The goodness of fit statistics are quite similar; however, the imputed data set reveals slightly improved fit compared with the skim data set, as reflected by the log-likelihood values, the rho-squared values, and the chi-squared values.

6. AVERAGE PREDICTED PROBABILITIES

Although examination of the MNL model estimates on imputed and skim data is insightful, additional and perhaps greater insight is derived through examination of the predicted choice probabilities of these models. The average predicted probabilities across the sample (i.e., $\bar{P}(C_i) = \sum_{n=1}^{M} P(C_i)_n / M$) for the imputed and skim data sets are tabulated and compared in Table IX. In this table P(SOV), P(HOV), P(Transit), and P(NM) are the probabilities of choosing the alternatives SOV, HOV, transit, and NM modes averaged over the sample of choosers, respectively. Inspection of the table reveals that

Table VII. Average and absolute difference of predicted time and distance.

Average difference in travel time and distance for imputed and skim data sets*	
Variable	Mean
Average time difference (minutes)	2.9816
Average distance difference (miles)	0.9883
Average absolute difference in travel time and distance for imputed and skim data sets*	
Variable	Mean
Average absolute time difference (minutes)	16.0772
Average absolute distance difference (miles)	5.5476

*Imputed *times* and *distances* vary within zones, whereas skim *times* and *distances* do not.

DOI: 10.1002/atr

Table VIII. Multinomial logit estimation results for skim and imputed data sets.

	Parameter estimates for skim data set					Parameter estimates for imputed data set			
Variable name	Variable	Coefficient	Standard error	b/Standard error	P[−Z| > z]	Coefficient	Standard error	b/Standard error	P[−Z| > z]
Single occupant vehicle specific variables									
Alternative specific constant	CONS1	0.654	0.448	1.457	0.1450	−0.277	0.554	−0.5	0.6160
Travel time	B1	−0.022	0.02	−1.074	0.2820	−0.012	0.012	−1.036	0.3000
Number of vehicles in household	B2	2.16	0.225	9.58	0.0000	2.15	0.222	9.716	0.0000
High occupant vehicle specific variables									
Alternative specific constant	CONS2	−0.475	0.61	−0.78	0.4350	−0.853	0.673	−1.267	0.2051
Travel time	B3	−0.028	0.027	−1.04	0.2980	−0.051	0.017	−2.963	0.0030
Number of workers in household	B4	0.901	0.191	4.714	0.0000	0.94	0.193	4.863	0.0000
Number of vehicles in household	B5	0.799	0.254	3.143	0.0017	0.764	0.249	3.064	0.0022
Ethnicity being white	B6	−1.068	0.266	−4.017	0.0001	−1.048	0.266	−3.929	0.0001
Transit specific variables									
Alternative specific constant	CONS3	−0.822	0.8	−1.018	0.3080	−0.769	0.837	−0.919	0.3581
Travel time	B7	−0.052	0.024	−2.156	0.0311	−0.081	0.021	−3.825	0.0001
Number of workers in household	B8	1.428	0.21	6.77	0.0000	1.413	0.021	6.659	0.0000
Households in low income category	B9	0.813	0.335	2.421	0.0155	0.79	0.34	2.319	0.0204
Non-motorized specific variables									
Non-motorized commute distance	B10	−0.083	0.045	−1.841	0.0656	−0.153	0.047	−3.263	0.0011
Number of bikes in an household	B11	0.412	0.124	3.319	0.0009	0.367	0.126	2.9	0.0037
		Goodness of fit statistics				Goodness of fit statistics			
		Number of observations			1070	Number of observations			1070
		Skipped observations			0	Skipped observations			0
		Log likelihood function			−468.86	Log likelihood function			−453.56
		ρ^2			0.683	ρ^2			0.694
		Adjusted ρ^2			0.682	Adjusted ρ^2			0.692
		ρ^2 constants only			0.215	ρ^2 constants only			0.241
		Adjusted ρ^2 constants only			0.212	Adjusted ρ^2 constants only			0.238
		χ^2			258.12	χ^2			288.73
		Prob[χ^2 > value]			0.000	Prob[χ^2 > value]			0.000

DOI: 10.1002/atr

Table IX. Average predicted probabilities of choosing an alternative for imputed and skim data sets.

Alternative	(A) Average probability of selection: validation data	(B) Average probability of selection: imputed data	Average difference in probabilities (A − B)
Single occupant vehicle	0.8560	0.8596	−0.0003
High occupant vehicle	0.0650	0.0648	0.0002
Transit	0.0531	0.0532	−0.0001
Non-motorized	0.0220	0.0224	−0.0004

average predicted probabilities are arbitrarily close, suggesting that the imputed values produce on average the same choice probabilities as the skim derived *time* and *distance* variables.

Although the average predicted probabilities reveal that average choice behavior is duplicated on average with the imputed data, it is possible that variation across choosers is relatively large. To measure the variation across choosers, the mean difference between the predicted probabilities for imputed and skim data sets is computed, such that $\bar{P}_{\Delta}(C_i) = \sum_{n=1}^{M} \left(P(C_{\text{Observed}}) - P\left(C_{\text{Imputed},i}\right)\right)/M$. Table IX shows the average differences and reveals that they are arbitrarily small. The largest difference, for example, is 0.03% for $i = \text{SOV}$.

6.1. Correlation between predicted probabilities

If imputed data are consistent with the choice process revealed in the validation data, then the predicted choice probabilities between the skim and imputed MNL models should be highly correlated. The predicted probabilities are consistently highly correlated, with the highest correlation being 0.971 for SOV and the lowest being 0.78 for NM travel. Not surprisingly, these modes reflect the most and least often chosen modes, respectively.

7. DISCUSSION AND CONCLUSIONS

The primary objective of this research was to develop a multivariate model for imputing unobserved attribute values in the travel mode context. The theoretical development and validation of a BI-MNL model for obtaining multiple imputations of non-chosen attribute values was described first. Imputed values are based on random draws from the conditional posterior distribution of missing variables, given the observed variables and model parameters with informative priors. Comparison of BI-MNL model imputed and network skim derived *time* and *distance* values reveals close agreement across a variety of statistical comparisons. The observed correlation between *time* and *distance* (0.519) is well replicated between the imputed *time* and *distance* (0.511). An area of further potential BI-MNL model refinement is revealed by the difference between skim and imputed *time* and *distance* values, which differs on average by about 10%.

An evaluation of the BI-MNL model imputed *time* and *distance* values is afforded by comparing MNL model estimation results using the skim and imputed data sets. These comparisons offer compelling evidence that the choice behavior revealed using network skim derived *time* and *distance* values is closely replicated with BI-MNL model imputed *time* and *distance* values. For example, the average predicted probability difference (across the 1070 individuals in the validation sample) between skim and imputed data MNL models for choosing SOV is 0.03% (SOV reflects the largest difference). The correlation between predicted probabilities is 0.971 for choosing SOV and 0.784 for choosing NM modes.

Although the results of the BI-MNL model suggest that imputing non-chosen attributes is feasible, further model refinement and improvement is possible and even desirable. Because of data availability, network skim values served as the calibration sample. However, network skim values are not ideal and suffer from a host of network calibration problems and because network *distance* values are constant across modes within an O-D pair and reflect zone centroid distances. Skim values are inaccurate because the distance from zone *i* to zone *j* by car is likely to differ across individuals. Network derived *times* also do not account for the *distance* differences that might occur within a TAZ. For these reasons, a superior validation sample is needed to demonstrate the true potential of this approach. Alternatives

DOI: 10.1002/atr

to the relationship between imputed *times* and *distances* also could be explored, using linear forms with access time intercepts as perhaps more intuitively appealing specifications.

The use of network skim values causes a negative ripple effect on the calibration of the BI-MNL model as described in this paper. First, because the log-likelihood is preserved in the BI-MNL model calibration process (i.e., log-likelihood of BI-MNL model $\cong$ log-likelihood of MNL model on skim data), imputed *times* and *distances* will vary more than skim values to explain the revealed choice behavior. In other words, whereas the skim *distances* (and to a lesser extent *times*) are constant across mode within an individual, the imputed *distances* vary across modes for an individual (again see Table V), consistent with actual travel situations. This limitation of network skim derived *times* and *distances* in the calibration of the BI-MNL model is likely to explain the observed differences in the values of *times* and *distances* between imputed and skim values. Specifically, using network skim data for the calibration sample produces no within-zone variation in time and distance, yet the imputed data do contain within-zone variation. As a result, average differences are small (all travel times within a zone perform well), but absolute average differences within a zone are large and contribute to this poor goodness of fit statistic. This ripple effect of using network skims to calibrate the BI-MNL model can be overcome with a robust implementation plan including but not limited to the use of improved calibration data, as described in the next section.

8. IMPLEMENTATION CONSIDERATIONS

Two practice-oriented outcomes would motivate the use of this approach for obtaining unobserved choice attributes. The first is the desire to examine the impact of travel choice policies, prices, or projects on travelers' mode choices within zones. For example, the marginal impact of expanded rural bus transit on households within large rural TAZs would be well served with intrazonal variation in attributes as compared with average zonal values typically applied using alternative approaches. An analyst's desire to capture travel preferences within zones is well served by this approach.

A second motivation might arise from a desire to replace zonal skims and their known associated inaccuracies by collecting detailed survey information on a subset of households and then imputing unobserved attributes for the complete sample. Although this approach requires costly and perhaps more onerous respondent load for the subsample needed to acquire attribute information across all mode choices, the resulting complete data set would provide richer and more complete information than obtained using more traditional methods, particular in regard to within zone attributes.

A practical implementation of this approach remains as further research. It should intend to demonstrate the advantages of imputing within-zone variation of unobserved mode attributes for all travelers. One surely can argue that within-zone variation in travel exists to a significant degree as evidenced by heterogeneity in travel choices within zones. This variation could inform transportation planning activities with additional and important insights as to the impacts of travel policies, pricing, and programs and their market penetration within transportation analysis zones.

REFERENCES

1. Hensher DA, Rose JM, Greene WH. *Applied Choice Analysis: A Primer*. Cambridge University Press: Cambridge, 2005.
2. Jou YJ, Cho HJ, Lin PW, Wang CY, American Society of Civil Engineers. Incomplete information analysis for the Origin-Destination survey table. *Journal of Urban Planning and Development* 2006; **132**(4):193–200.
3. Brownstone D. Multiple imputation methodology for missing data, non-random response, and panel attrition. *Theoretical Foundations of Travel Choice Modeling*. University of California Transportation Center: Berkeley, 1998.
4. Rubin D. Inference and missing data. *Biometrika* 1976; Biometrika Trust. **63**(3):581–592.
5. Schafer JL. *Analysis of Incomplete Multivariate Data*. CRC Press/Chapman Hall: Boca Raton, FL, 1997.
6. Laird NM. Missing data in longitudinal studies. *Statistics in Medicine* 1988 Jan-Feb; **7**(1–2):305–15.
7. Datla SK, Zhaobin L, Sharma S. A nearest neighbourhood approach for estimation of missing traffic data during holiday periods. Proceedings of the Transportation Research Board 87th Annual Meeting. National Academies, Washington DC, 2008.
8. Wang J, Zou N, Chang G. Empirical analysis of missing data issues for ATIS applications: travel time prediction. Proceedings of the Transportation Research Board 87th Annual Meeting. National Academies, Washington DC, 2008.

DOI: 10.1002/atr

9. Zhao F, Chung S. A study of alternative land use forecasting models. Florida Department of Transportation Report BD015-10: FDOT 99700-3596-119, 2006.
10. Steimetz SSC, Browstone D. Estimating commuters "value of time" with noisy data: a multiple imputation approach. *Transportation Research Part B: Methodological* 2005; **39**(10):865–889.
11. Brownstone D, Golob TF. The effectiveness of ridesharing incentives: discrete-choice models of commuting in Southern California. *Regional Science and Urban Economics* 1992; **22**:5–24.
12. Congdon P. *Bayesian Statistical Modelling*. Wiley: New York, NY, 2001.
13. Congdon P. Bayesian predictive model comparison via parallel sampling. *Computational Statistics and Data Analysis*. Wiley: New York, NY, 2004.
14. Washington S, Karlaftis M, Mannering F. *Statistical and Econometric Methods for Transportation Data Analysis*. Chapman Hall/CRC Press: Boca Raton, FL, 2003.
15. Marsaglia G. Ratios of normal variables. *Journal of Statistical Software* 2006; (16):4.
16. Daly A, Hess S, Train K. 2009. Assuring finite moments for willingness to pay in random coefficient models. Presented at the European Transport Conference, Leeuwenhorst, October 2009.
17. Congdon P. *Applied Bayesian Statistical Models*. Wiley: New York, NY, 2003.
18. Gelman A, Carlin J, Stern H, Rubin DB. *Bayesian Data Analysis*. Chapman and Hall: New York, NY, 1995.
19. Gilks W, Richardson S, Spiegelhalter D (eds.) *Markov Chain Monte Carlo in Practice*. Chapman and Hall: New York, NY, 1996.
20. Brooks S, Roberts GO. Assessing convergence on Markov chain Monte Carlo algorithms. *Statistics and Computing* 1998; **8**:319–335.
21. Robert C, Casella G. *Monte Carlo Statistical Methods*. Springer. New York, NY, 1997.
22. NuStats. Maricopa regional household travel survey, Final Report. Submitted to the Maricopa Association of Governments, Phoenix Arizona, 2002.

DOI: 10.1002/atr