Transportation Research Part B 142 (2020) 45–57

Contents lists available at ScienceDirect

Transportation Research Part B

journal homepage: www.elsevier.com/locate/trb

Combining multiple imputation and control function methods to deal with missing data and endogeneity in discrete-choice models

Raja Gopalakrishnan[a,b,*,*], C. Angelo Guevara[c,d], Moshe Ben-Akiva[e]

[a] Singapore University of Technology and Design, 8 Somapah Road, 487372, Singapore
[b] Singapore-MIT Alliance for Research and Technology, 09-02 CREATE Tower, 138602, Singapore
[c] Universidad de Chile, Departamento de Ingeniería Civil, Blanco Encalada 2002, Santiago, Chile
[d] Instituto Sistemas Complejos de Ingeniería, Av. República 701, Santiago, Chile
[e] Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

ARTICLE INFO

Article history:
Received 13 April 2020
Revised 27 September 2020
Accepted 4 October 2020
Available online 15 October 2020

Keywords:
Imputation
Missing data
Endogeneity
Discrete choice
Control function
Monte-Carlo simulation
Missing at random
Limited information maximum likelihood
Urban freight
Commercial vehicle parking

ABSTRACT

While collecting data for estimating discrete-choice models, researchers often encounter missing information in observations. In addition, endogeneity can occur whenever the error term is not independent of the observed variables. Both problems result in inconsistent estimators of the model parameters. The problems of missing information and endogeneity may occur in the same variable in the data, if, e.g., partially missing price information is correlated with another omitted variable. Extant approaches to correct for endogeneity in discrete choice models, such as the control function method, do not address the problem when the error term is correlated with a variable having missing information. Likewise, approaches to address missing information, such as the multiple imputation method, cannot handle endogeneity problems. To address this challenge, we propose a novel hybrid algorithm by combining the methods of multiple imputation and the control function. We validate the algorithm in a Monte-Carlo experiment and apply it to real data of heavy commercial vehicle parking from Singapore. In this case study, we were able to substantially correct for price endogeneity in the presence of missing price information.

1. Introduction

In a multi-stakeholder environment, collecting complete information on all the variables of interest for developing discrete-choice models is an onerous task. The researcher might encounter data with missing information, which could either be partial or total, that may also be prone to endogeneity problems from different sources.

Customarily, the data is presented with rows representing an individual observation and columns indicating the attributes. The data suffers from missing information when a column value is available for some observations and missing for

* Corresponding author at: Centre for Railway Information Systems, Chanakyapuri, New Delhi 110 021, India.
E-mail addresses: r.gopalakrishnan@cris.org.in (R. Gopalakrishnan), crguevar@ing.uchile.cl (C.A. Guevara), mba@mit.edu (M. Ben-Akiva).

https://doi.org/10.1016/j.trb.2020.10.002

others. This problem may arise from different causes. For example, in a survey, respondents may either not know or may refuse to answer a question. In both instances, the data collected will have missing information on the response variable. Similarly, longitudinal studies may be prone to attrition problems when subjects are observed in a series of experiments conducted in different epochs; those who participated in the initial experiments may no longer be available for the follow-on experiments. Even while collating data from secondary sources, some fields might have partial missing values. In merging data from multiple secondary sources, the extant of missing information increases with the addition of each data source.

Endogeneity may also occur in the model when the data has limitations arising from one or more of the following causes - omitted variables correlated with an observed attribute, errors in measurement, simultaneity or self-selection (Guevara and Ben-Akiva, 2012). A textbook definition of endogeneity highlights violation of assumption of independence of the error term with observed attributes in the data (Greene, 2011). Ignoring endogeneity may result in inconsistent parameter estimates of the developed model. Guevara and Ben-Akiva (2012) present an example of a dwelling choice model, where ignoring the quality attribute that is positively correlated with rent might spuriously reduce the magnitude (negative value) of the estimated price coefficient. In fact, the estimated price coefficient could even be positive rendering the model unsuitable for forecasting (Guevara and Ben-Akiva (2006); Bhat and Guo (2004); Sermons and Koppelman (2001)).

In real-world applications, both data deficiencies of missingness and endogeneity may occur in the same attribute. For example, prices, which are almost inevitably endogenous, were found to be partially missing in a study on airline consumer search and purchase behavior (Hotle et al., 2015) and on-line shopping (Cavallo, 2018). Similarly, in transportation studies, partially missing values of travel time are encountered by Steimetz and Brownstone (2005) and Guo et al. (2018) while endogeneity due to an omitted attribute correlated with travel time is also reported in literature (Wardman and Whelan, 2011; Tirachini et al., 2013).

The methods that have been applied so far consider the data deficiencies, of missing information or endogeneity, separately. A method to overcome missing information is to ignore observations with missing values in any column - termed the complete-case method (Little and Rubin, 2002). Adopting this method might decrease the sample size considerably, thereby reducing the confidence in the parameter estimates. Further, the model may no longer represent the target population. Imputation methods (Little and Rubin, 2002), both heuristics and model-based, are used to estimate the missing column values in the data. Rubin (1987) developed the multiple imputation method which avoids the pitfalls from using adhoc-methods of handling missing values. Under assumption of Missing at Random (MAR) (Rubin, 1987), the missing values are imputed from a conditional distribution whose parameters are estimated from a model developed from observations with complete information. The key assumption of MAR is that the conditional probability of missing information in a variable, given other observed variables, is independent of its value. In the multiple imputation method, random draws from the conditional distribution are used to generate multiple values of the missing variable. Each draw of the missing variable constitutes an imputed but complete data set. Each complete data set is then treated in a manner as if there was complete information. Examples of applications of the multiple imputation methods for discrete choice, include bid-rent (Li, 2014), and value of travel-time (Steimetz and Brownstone, 2005) models.

Endogeneity in discrete-choice models is primarily corrected by applying the Berry-Levinson-Pakes (BLP, Berry et al. (1995)) or the control function (Rivers and Vuong, 1988) method. Guevara (2015) presents an overview of the endogeneity correction methods in discrete choice literature. BLP method can be applied when the source of endogeneity is an alternative specific omitted attribute that does not vary across agents and data from different markets are observed. A few applications which use the BLP method to correct for endogeneity include those in the automobile and satellite television markets (Berry et al., 1995; Goolsbee and Petrin, 2004). When the omitted attribute varies across agents or the data from a single market is given, control function method (Rivers and Vuong, 1988) may be applied. Examples of applications of the control function method include residential location choice (Ferreira, 2010; Guevara and Ben-Akiva, 2012) and consumer choice models (Petrin and Train, 2010).

The control function method, which is shown to correct for endogeneity in discrete-choice models, may be considered as a single imputation of the omitted attribute under complete information of the endogenous variable. However, when the endogenous variable suffers from partially missing information, there is a gap in literature on the methods applicable. This research proposes, illustrates and assesses a novel hybrid algorithm, which combines the methods of multiple-imputation and control function, to address both model mis-specifications - missing information and endogeneity. Our algorithm also corrects for a limitation in previous literature (Steimetz and Brownstone, 2005) in the multiple imputation method when applied to develop discrete-choice models. We demonstrate, in a Monte-Carlo experiment, the efficacy of the proposed algorithm when the endogenous variable is MAR and an associated instrumental variable can be constructed. It is followed by an application of the algorithm to model overnight parking location choice behavior using data collected in Singapore.

The article is organized as follows - Section 2 presents a short introduction to the methods of multiple imputation and control function. Section 3 describes our proposed novel algorithm that corrects for both missing information and endogeneity in the same variable. The algorithm is validated in a Monte-Carlo experiment and the results are presented in Section 4. Section 5 illustrates a real-world application of the algorithm in modeling the overnight parking subscription behavior of heavy commercial vehicles in cities. Finally, Section 6 concludes by summarizing the main findings, discussing the implications for policy and practice, and outlining the directions for future research.

2. Methodology review : multiple imputation and control function

2.1. Theory of multiple imputation

We briefly describe the theory of multiple imputation (see e.g., Little and Rubin, 2002). Let $\boldsymbol{\beta}$ be a vector of parameters that we wish to estimate from the sample data, $\boldsymbol{X}$. Let the prior distribution of $\boldsymbol{\beta}$ be $f(\boldsymbol{\beta})$. Using Bayes rule, the posterior probability distribution, $f(\boldsymbol{\beta}|\boldsymbol{X})$, can be written as:

$$f(\boldsymbol{\beta}|\boldsymbol{X}) \propto \boldsymbol{L}(\boldsymbol{X}|\boldsymbol{\beta})f(\boldsymbol{\beta}) \tag{1}$$

where $\boldsymbol{L}$ is the likelihood of observing the sample data $\boldsymbol{X}$.

Let $\boldsymbol{X_p}$: $\boldsymbol{X}\backslash\boldsymbol{p}$ denote the data $\boldsymbol{X}$ without the variable $\boldsymbol{p}$ that has missing information. Let $\boldsymbol{p^{obs}}$ and $\boldsymbol{p^{mis}}$ denote the observed and missing elements in the variable $\boldsymbol{p}$. Further, the missingness mechanism in variable $\boldsymbol{p}$ is assumed to be **MAR**.

The objective is to estimate the vector of parameters, $\boldsymbol{\beta}$, from the data with missing information. The expected value of $\boldsymbol{\beta}|\boldsymbol{X_p}, \boldsymbol{p^{obs}}$ with missing data can be written as:

$$\begin{aligned} \mathrm{E}(\boldsymbol{\beta}|\boldsymbol{X_p}, \boldsymbol{p^{obs}}) &= \int_{\boldsymbol{p^{mis}}} \mathrm{E}(\boldsymbol{\beta}|\boldsymbol{X_p}, \boldsymbol{p^{obs}}, \boldsymbol{p^{mis}})\ dP(\boldsymbol{p^{mis}}|\boldsymbol{X_p}, \boldsymbol{p^{obs}}) \\ &= \int_{\boldsymbol{p^{mis}}} \mathrm{E}(\boldsymbol{\beta}|\boldsymbol{X_d})\ dP(\boldsymbol{p^{mis}}|\boldsymbol{X_p}, \boldsymbol{p^{obs}}) \end{aligned} \tag{2}$$

The idea behind the multiple imputation method (Rubin, 1987) is equivalent to evaluating $\mathrm{E}(\boldsymbol{\beta}|\boldsymbol{X_p}, \boldsymbol{p^{obs}})$ in Eq. (2) by Monte-Carlo integration. Each random draw of the missing value, $\mathbf{p_d^{mis}}\ \forall\ d = 1, \ldots, D$, from its distribution, $\boldsymbol{p^{mis}}|\boldsymbol{X_p}, \boldsymbol{p^{obs}}$, generates a vector, $\boldsymbol{p_d} \equiv [\boldsymbol{p^{obs}}\ \boldsymbol{p_d^{mis}}]^T$, with imputed values. $\boldsymbol{X_d} \equiv [\boldsymbol{X_p}\ \boldsymbol{p_d}]$ denotes a complete dataset with imputed values.

The Bernstein-von Mises theorem (see e.g., Train, 2009) relates the Bayesian estimators to the classical Maximum Likelihood Estimator (MLE). The following results summarize the relationship:

$$\mathrm{E}(\boldsymbol{\beta}|\boldsymbol{X_d}) \xrightarrow{p} \hat{\boldsymbol{\beta}}_{\boldsymbol{d}} \tag{3}$$

$$(\boldsymbol{\beta} - \mathrm{E}(\boldsymbol{\beta}|\boldsymbol{X_d})) \xrightarrow{d} \mathrm{Normal}(0, -\frac{\boldsymbol{H_d}^{-1}}{N}) \tag{4}$$

where N is the sample size, $\hat{\boldsymbol{\beta}}_{\boldsymbol{d}}$ is the MLE for the d^{th} complete dataset with imputed values, and $-\boldsymbol{H_d}$ is the associated Fischer-information matrix. Substituting the maximum likelihood estimator computed in each imputation in Eq. (3) into Eq. (2), we obtain

$$E(\boldsymbol{\beta}|\boldsymbol{X_p}, \boldsymbol{p^{obs}}) \approx \bar{\boldsymbol{\beta}} = \frac{1}{D}\sum_{d=1}^{D} \hat{\boldsymbol{\beta}}_d \tag{5}$$

where $\bar{\boldsymbol{\beta}}$ denotes the average of the values $\hat{\boldsymbol{\beta}}_d, d = 1, \ldots, D$.

Using the law of total variance, the variance of the random variable $\boldsymbol{\beta}|\boldsymbol{X_p}, \boldsymbol{p^{obs}}$ can be computed as under:

$$\begin{aligned} \mathrm{Var}(\boldsymbol{\beta}|\boldsymbol{X_p}, \boldsymbol{p^{obs}}) &= \mathrm{E}(\mathrm{Var}[(\boldsymbol{\beta}|\boldsymbol{X_p}, \boldsymbol{p^{obs}})|\boldsymbol{p^{mis}}]) + \mathrm{Var}(\mathrm{E}[(\boldsymbol{\beta}|\boldsymbol{X_p}, \boldsymbol{p^{obs}})|\boldsymbol{p^{mis}}]) \\ &= \mathrm{E}(\mathrm{Var}(\boldsymbol{\beta}|\boldsymbol{X_d})) + \mathrm{Var}(\hat{\boldsymbol{\beta}}_d) \end{aligned} \tag{6}$$

Conventionally, the first term of Eq. (6) is termed as within-sample variance with the latter denoting the between-sample variance. Under D random imputation draws, an estimate of the variance, $\boldsymbol{\beta}|\boldsymbol{X_p}, \boldsymbol{p^{obs}}$ can be obtained from the expression below (Little and Rubin, 2002):

$$\mathrm{Var}(\boldsymbol{\beta}|\boldsymbol{X_p}, \boldsymbol{p^{obs}}) \approx \frac{1}{D}\sum_{d=1}^{D} \mathrm{Var}(\boldsymbol{\beta}|\boldsymbol{X_d}) + \frac{1}{D-1}\sum_{d=1}^{D} (\hat{\boldsymbol{\beta}}_d - \bar{\boldsymbol{\beta}}) \cdot (\hat{\boldsymbol{\beta}}_d - \bar{\boldsymbol{\beta}})^T \tag{7}$$

Rubin and Schenker (1986) report that interval estimates with three imputation runs are close to the nominal values even with 60% missing information when the data is exactly normal with small sample size or any data for large samples. Other studies show that 2–10 imputation runs are sufficient to recover the point estimates (Bodner, 2008; von Hippel, 2018). However, the number of imputation runs required to recover the standard errors may be much higher and is shown to depend quadratically on the fraction of missing information (von Hippel, 2018).

For discrete-choice models, the Cramer-Rao theorem can be used to compute the variance of the maximum likelihood estimator. Eq. (7) can be used to calculate the final variance for a given sample.

Unconditional estimates of mean and variance of the population parameters can be inferred by drawing multiple random samples from a population. The mean and variance - covariance matrix is computed as below:

$$\mathrm{E}(\boldsymbol{\beta}) = \mathrm{E}_{\boldsymbol{X}}(E(\boldsymbol{\beta}|\boldsymbol{X})) \tag{8}$$

$$\mathrm{Var}(\boldsymbol{\beta}) = \mathrm{Var}_{\boldsymbol{X}}(\mathrm{E}(\boldsymbol{\beta}|\boldsymbol{X})) + \mathrm{E}_{\boldsymbol{X}}(\mathrm{Var}(\boldsymbol{\beta}|\boldsymbol{X})) \tag{9}$$

where the estimates for $E(\boldsymbol{\beta}|\boldsymbol{X})$ and $\mathrm{Var}(\boldsymbol{\beta}|\boldsymbol{X})$ are obtained from Eqs. (5) and (7) respectively.

2.2. Theory of control function

The control function method is applicable when the endogenous variable can be assumed to be a function of a variable (an instrument) that is exogenous to the model, i.e., that does not belong to the structural equation of the utility, in the case of a discrete choice model based on the Random Utility Maximization principle. The method has its roots in the simultaneous equation models in linear regression when one of the dependent variables is latent (Heckman, 1978). Rivers and Vuong (1988) demonstrated applicability of the method to obtain consistent parameter estimates in binary probit models. The method relies on the availability of an instrument, an exogenous variable that is correlated with the endogenous variable. We describe the method (see, e.g., Train, 2009) in the following paragraphs.

We explain the control function method by referring to Eq. (10), where every agent $n \in N$ associates a utility U_{in} to each alternative i in the choice-set C_n. The symbols in bold represent vectors.

$$\begin{aligned} U_{in} &= \beta_y y_{in} + \beta_x x_{in} + \beta_p p_{in} + \overbrace{\xi_{in} + \epsilon_{in}}^{\eta_{in}}, \quad i \in C_n, \ \forall\ n = 1, \ldots, N \\ p_{in} &= \gamma_z z_{in} + \overbrace{\gamma_\xi \xi_{in} + \varepsilon_{in}}^{\delta_{in}} \end{aligned} \tag{10}$$

The utility is a linear function of an alternative specific variable $\boldsymbol{y}$, an exogeneous variable $\boldsymbol{x}$, and an endogeneous variable $\boldsymbol{p}$. $\boldsymbol{\eta}$, the error term, is a convolution of an omitted variable $\boldsymbol{\xi}$ and a random error $\boldsymbol{\epsilon}$. Eq. (10) also relates $\boldsymbol{p}$ to the instrument variable $\boldsymbol{z}$, omitted attribute $\boldsymbol{\xi}$, and an i.i.d. random error term $\boldsymbol{\varepsilon}$. Further, when $\boldsymbol{z}$ is exogenous, it is a valid instrument variable as it is correlated with $\boldsymbol{p}$ and independent of $\boldsymbol{\xi}$, $\boldsymbol{\varepsilon}$ and $\boldsymbol{\epsilon}$. When $\boldsymbol{\xi}$ is omitted, the error term in the utility function is correlated with $\boldsymbol{p}$, thereby causing endogeneity. In such a case, assumption of the error being independent of the data, which is needed for consistency in estimation, is no longer valid.

The distribution of η_{in}, conditional on observing δ_{in}, can be decomposed into sum of the conditional mean and an error term, $\tilde{\eta}_{in}$, independent of δ_{in}:

$$\eta_{in} = E(\eta_{in}|\delta_{in}) + \tilde{\eta}_{in} \tag{11}$$

Substituting Eq. (11) in the utility Eq. (10), we obtain:

$$U_{in} = \beta_y y_{in} + \beta_x x_{in} + \beta_p p_{in} + E(\eta_{in}|\delta_{in}) + \tilde{\eta}_{in} \tag{12}$$

The conditional expectation, $E(\eta_{in}|\delta_{in})$, is defined as the control function. Eq. (12) is cured of endogeneity as $\tilde{\eta}_{in}$, the error term in the utility function, is no longer correlated with p_{in}. Consistent estimates of the parameters of the choice model can be obtained by applying standard methods such as maximizing the likelihood function.

Often, the control function cannot be expressed in a closed form in δ_{in}. In addition, the distribution of $\tilde{\eta}_{in}$ is also unknown. However, when the choice model is binary probit, and the linear regression for the endogenous variable has a normal error term, the joint distribution of (η_{in}, δ_{in}) follows a normal distribution. The conditional expectation, $E(\eta_{in}|\delta_{in})$, when the variables (η_{in}, δ_{in}) are jointly normal, has the following closed form :

$$E(\eta_{in}|\delta_{in}) = \lambda \delta_{in} \tag{13}$$

A consistent estimate of δ_{in} is obtained by calculating the residuals, $\hat{\delta}_{in} = p_{in} - \hat{\gamma}_z z_{in}$, in the linear regression Eq. (10). $\hat{\delta}_{in}$ enters the utility function as an additional attribute and the parameters are estimated as if there is no endogeneity in the model. However, scale of the model changes as the variance of $\tilde{\eta}_{in}$ is no longer equal to the variance of η_{in} (Guevara and Ben-Akiva, 2012).

Under certain assumptions, Ruud (1983) showed that maximum likelihood estimators are consistent even when the distribution is misspecified. With this approximation, the control function method can also be extended to logit models even though it was initially developed for probit models.

3. Proposed hybrid algorithm

This section describes the proposed algorithm to address the problems of both endogeneity and missing information in discrete-choice models. We again refer to Eq. (10), where the utility function and the structural equation of the endogenous variable, $\boldsymbol{p}$, were described. In addition to endogeneity, variable $\boldsymbol{p}$ suffers from partially missing information. The researcher observes the variables $\boldsymbol{x}$, $\boldsymbol{z}$, the choice of every agent n and partial information on $\boldsymbol{p}$.

The proposed hybrid algorithm considers the following 5 assumptions:

A.1 The data generation process corresponds to that of a discrete-choice model.
A.2 Information on the endogenous variable, $\boldsymbol{p}$, is missing at random (MAR).
A.3 Missing values of the endogenous variable, $\boldsymbol{p}$, are imputed from an underlying linear regression model.
A.4 There exists an instrumental variable, $\boldsymbol{z}$, that is correlated with the endogenous variable, $\boldsymbol{p}$, and independent of the error terms.
A.5 Error term in the linear regression model, $\boldsymbol{\varepsilon}$, and the omitted variable, $\boldsymbol{\xi}$, in the utility specification, are normally distributed.

Assumption A.1 implies that the underlying model could be a probit, but also a logit, mixed logit or, e.g., a member of the MEV (GEV) family. MAR assumption implies that the missing and observed values of $\boldsymbol{p}$ are drawn from an underlying conditional distribution. Further, as assumption A.2 holds across alternatives, no alternative specific categorical variable is included in the specification of the imputation model. However, we may include an alternative specific constant (e.g., a labeled alternative) in the utility specification of the discrete-choice model. Assumption A.3 is not restrictive as it permits transformation of the regressands as long as the coefficients are linear. A normally distributed error term in assumption A.5 subsumes all unknown factors which influence the endogenous variable. The omitted attribute $\boldsymbol{\xi}$ can be assumed to be a smaller subset of the unknown factors.

We extend, with the addition of a control function step, the multiple imputation algorithm applied by Steimetz and Brownstone (2005) in estimating the parameters in a lane-choice model. Our algorithm is divided into three parts - linear regression, multiple imputation and control function steps.

Linear regression step :

L.1 The observed values of the endogenous variable, $\boldsymbol{p^{obs}}$, are regressed with the exogenous variables and instruments.

$$\boldsymbol{p^{obs}} = \boldsymbol{Z\theta} + \tilde{\boldsymbol{\varepsilon}}, \qquad \tilde{\boldsymbol{\varepsilon}} \sim N(0, \sigma^2),$$

where $\boldsymbol{Z} \equiv [\boldsymbol{z}\ \boldsymbol{X_p}]$.

L.2 Estimate parameters $\hat{\boldsymbol{\theta}} = (\boldsymbol{Z}^T\boldsymbol{Z}^{-1})\boldsymbol{Z}^T\boldsymbol{p^{obs}}$, where $\hat{\boldsymbol{\theta}} \in \Re^K$, and K is the number of parameters in the linear regression model.

Multiple imputation step :

The sum of squared residuals, $\frac{\sum \boldsymbol{e}^T\boldsymbol{e}}{\sigma^2} \sim \chi^2_{N-K}$[1] distribution (see e.g, Schenker and Welsh, 1988), where $\boldsymbol{e}$ is the residual vector obtained in the linear regression step. For each iteration, $d = 1, \ldots, D$, of the imputation:

MI.1 Draw a random number, χ_d, from χ^2_{N-K} distribution.

MI.2 Impute the variance of the error term, which is computed as $\hat{\sigma}^2_d = \frac{\sum \boldsymbol{e}^T\boldsymbol{e}}{\chi_d}$.

MI.3 Draw random vector, $\hat{\varepsilon}_d$, from $N(0, \hat{\sigma}^2_d)$.

MI.4 Draw $\hat{\boldsymbol{\theta}}_d$ from a multivariate normal distribution with mean, $\hat{\boldsymbol{\theta}}$, and variance-covariance matrix, $\hat{\boldsymbol{\Theta}}_d = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\hat{\sigma}^2_d$.

MI.5 Impute the vector of missing values, $\boldsymbol{p^{mis}_d}$, using the equation,

$$\boldsymbol{p^{mis}_d} = \boldsymbol{Z}\hat{\boldsymbol{\theta}}_d + \hat{\boldsymbol{\varepsilon}}_d$$

At the end of the multiple imputation step, we obtain D complete data sets.

Control function step :

For the proposed hybrid method that combines multiple imputation and control-function methods, we apply the latter on each completed dataset, $d = 1, \ldots, D$, separately.

Alternative methods such as Full-information maximum likelihood (Guevara and Ben-Akiva, 2012; Villas-Boas and Winer, 1999; Park and Gupta, 2009) specify the joint distribution of $(\boldsymbol{\eta}, \boldsymbol{\delta})$ and estimate, in a single-step, the parameters of both the utility function and the structural equation. Even though it is more efficient, specifying the conditional distribution in the control function method is more general than the joint distribution (Train, 2009) and is computationally easier (Guevara and Ben-Akiva, 2012).

Under this setting, we apply the control function method that can be viewed as a single imputation of the omitted attribute when complete information on the endogenous variable is available.

CF.1 Regress the vector of endogenous variables, $\boldsymbol{p_d} = [\boldsymbol{p^{obs}}\ \boldsymbol{p^{mis}_d}]^T$ with the instrument $\boldsymbol{z}$ and exogenous variables in the utility function.

$$\boldsymbol{p_d} = \boldsymbol{Z}\ \boldsymbol{\gamma} + \boldsymbol{\delta}_d$$

CF.2 Include the estimated residual $\hat{\boldsymbol{\delta}}_d$ as an additional variable in the utility function.

CF.3 Estimate the parameters, $\hat{\boldsymbol{\beta}}_d$, of the discrete-choice model.

The instrument $\boldsymbol{z}$ is correlated with the true values of the endogenous variable, $\boldsymbol{p}$. However, we use the imputed values $\boldsymbol{p}_d$ while applying the control function. The following lemma shows that $\boldsymbol{z}$ remains a valid instrument for the imputed endogenous variable, $\boldsymbol{p}_d, d = 1, \ldots, D$.

[1] We correct the model proposed by Steimetz and Brownstone (2005) where the authors assume $\frac{\sum \boldsymbol{e}^T\boldsymbol{e}}{\sigma^2}$ to be χ^2_K distributed.

Lemma 1. *Let* $\boldsymbol{z}$ *be a valid instrument* that *is correlated with the true values of the endogenous variable* $\boldsymbol{p}$, *i.e.,* $\text{Cov}(\boldsymbol{z}, \boldsymbol{p}) = \rho \neq 0$. *Let* $\boldsymbol{p}_d$ *be the vector of imputed values from an underlying model* $\boldsymbol{p_d} = \boldsymbol{Z\theta} + \tilde{\boldsymbol{\epsilon}}$, *where* $\tilde{\boldsymbol{\epsilon}}$ *is an i.i.d. error term.*

Under these conditions, $\boldsymbol{z}$ *is a valid instrument for each complete data set i.e.,*

(i) $\text{Cov}(\mathbf{z}, \mathbf{p}_d) \neq 0 \ \forall \ d = 1, \ldots, D.$
(ii) $\boldsymbol{z}$ *is independent of the error terms.*

Proof. The missing values $\boldsymbol{p}_d$ are imputed from an underlying model with an i.i.d. error term. Therefore, the relationship between $\boldsymbol{p}_d$ and $\boldsymbol{p}$ can be written as

$$\boldsymbol{p}_d = \boldsymbol{p} + \boldsymbol{\omega}_d$$

where $\boldsymbol{\omega}_d$ is an i.i.d random variable. Therefore,

$$\begin{aligned}\text{Cov}(\boldsymbol{z}, \boldsymbol{p}_d) &= \text{Cov}(\boldsymbol{z}, \boldsymbol{p} + \boldsymbol{\omega}_d) \\ &= \text{Cov}(\boldsymbol{z}, \boldsymbol{p}) + \text{Cov}(\boldsymbol{z}, \boldsymbol{\omega}_d)\end{aligned}$$

Now $\text{Cov}(\boldsymbol{z}, \boldsymbol{\omega}_d) = 0$ by construction. Therefore, $\text{Cov}(\boldsymbol{z}, \boldsymbol{p}_d) = \rho \neq 0$. Besides, $\boldsymbol{z}$ is independent from the error terms as it was a valid instrument for the true $\boldsymbol{p}$. □

The final parameter estimates of the discrete-choice model and the associated standard errors are computed from the results obtained in each imputed data set using Eqs. (5) and (7).

4. Monte-Carlo simulation experiment

The hybrid algorithm is tested in a Monte-Carlo experiment where the data is generated from an underlying binary choice model. In the following paragraphs, we describe the data generation process and the experiments.

The data generation process underlying the Monte-Carlo experiments considers that the utility function, $\boldsymbol{U}$, and structural equation for the endogenous variable, $\boldsymbol{p}$, be defined as under:

$$\begin{aligned} U_{in} &= 0.5y_{in} + 2x_{in} + p_{in} + 4\xi_{in} + \epsilon_{in}, \ \ i = \{1, 2\}, \ \ \forall \ n = 1, \ldots, N \\ y_{in} &= \begin{cases} 1 & \text{if i =1, } \ \forall \ n = 1, \ldots, N \\ 0 & \text{otherwise.} \end{cases} \\ p_{in} &= 2z_{in} + 0.5\xi_{in} + \varepsilon_{in} \end{aligned} \tag{14}$$

where $\boldsymbol{x}$, $\boldsymbol{\xi}$, $\boldsymbol{z}$ and $\boldsymbol{\epsilon}$ are respectively the exogenous variable, omitted attribute, instrumental variable and the random error term in the utility function $\boldsymbol{U}$. Without loss of generality, y_{in}, the alternative specific constant is included in the utility specification of the first alternative. The endogenous variable, $\boldsymbol{p}$, is a function of the instrument, $\boldsymbol{z}$, and the omitted variable, $\boldsymbol{\xi}$.

The exogenous variable, $\boldsymbol{x}$, and the instrument, $\boldsymbol{z}$, are i.i.d random draws from a Normal (0,1) and Uniform (0,1) distribution respectively. The error term in the utility function, $\boldsymbol{\epsilon}$, is type-I extreme value (Gumbel) distribution with mode, 0, and scale parameter, $\boldsymbol{\mu} = 1$. The errors in the endogenous variable regression equation, $\boldsymbol{\varepsilon}$, and the omitted attribute, $\boldsymbol{\xi}$, are independent draws from a standard normal distribution. When $\boldsymbol{\xi}$ is omitted, the error term in the utility function is correlated with $\boldsymbol{p}$ thereby causing endogeneity. Further, $\boldsymbol{z}$ is a valid instrument variable as it is correlated with $\boldsymbol{p}$ and independent of $\boldsymbol{\xi}$ and $\boldsymbol{\epsilon}$. A population of 100,000 observations was constructed. To avoid any sampling bias, the experiments were repeated with 100 random samples of 8000 observations each drawn from the population. The computations were performed using the open-source statistical package, *R* (R Core Team, 2013).

Two sets of experiments are performed. In the first set of experiments, complete information on the endogenous variable, $\boldsymbol{p}$, is assumed and models are estimated as described below:

M-1 Full information on the variables $\boldsymbol{p}$, $\boldsymbol{\xi}$.
M-2 Full information on the variable $\boldsymbol{p}$ but omitting the attribute, $\boldsymbol{\xi}$. It introduces endogeneity in the model. Two submodels are estimated:

(a) Without correcting for endogeneity.
(b) Correcting for endogeneity by applying the control function method.

Models, M-1 and M-2, respectively validate the data generating process and the applicability of the control function method to correct for endogeneity under this setting.

In the second set of experiments, missingness in the endogenous variable is introduced by removing all the values of $\boldsymbol{p}$ in alternative 2. Further, attribute $\boldsymbol{\xi}$ is completely omitted in both the alternatives. Multiple imputation method is applied to impute the missing values of p_{2n}. The values of p_{2n} are imputed from a linear regression of p_{1n} with the instrument z_{1n}. Two models are estimated by:

M-3 Applying multiple imputation algorithm to impute missing values p_{2n} without correcting for endogeneity.

Table 1
Models with complete information on endogenous variable.

	M-1 : True model		M-2 (a) : ξ omitted		M- 2(b) : ξ omitted, 2SCF	
Par. (true value)	Value (std.error)	t-value*	Value (std.error)	t-value*	Value (std.error)	t-value*
$\hat{\beta}_y$ $(\beta_y = 0.5)$	0.500 (0.0706)	0.00394	0.209 (0.0546)	−5.319	0.479 (0.148)	−0.145
$\hat{\beta}_x$ $(\beta_x = 2)$	1.99 (0.121)	−0.0946	0.883 (0.0855)	−13.1	1.98 (0.356)	−0.0650
$\hat{\beta}_\xi$ $(\beta_\xi = 4)$	3.95 (0.186)	−0.260				
$\hat{\mu}$ $(\mu = 1)$	1.02 (0.0569)	0.298	0.692 (0.0267)	−11.6	0.321 (0.0482)	−14.1
$\hat{\beta}_\delta$ $(\beta_\delta = 0)$					1.57 (0.397)	3.95
Sample size	8000		8000		8000	
No. of simulation runs	100		100		100	
Init. log-likelihood	−5545.18		−5545.18		−5545.18	
Final log-likelihood	−1473.36		−4422.22		−4328.72	

(*) t-tests are against the respective population values.

M-4 Applying the hybrid algorithm.

We conducted the experiments, M-3 and M-4, with 20 imputations for p_{2n} as literature suggests that while 2–10 imputations are sufficient for consistent point estimates, larger imputations may be required for consistent standard errors (von Hippel, 2018).

Variance of the error term in the utility equation increases when an attribute is omitted. As the scale changes across the different models, the absolute values of the coefficients are no longer comparable (Guevara and Ben-Akiva, 2012). Only the ratio of the coefficients, $\frac{\hat{\beta}_x}{\hat{\beta}_p}$, are comparable. When the sample sizes are large, the coefficients obtained using maximum likelihood estimation are distributed asymptotically normal. However, the ratio distribution of two standard normally distributed random variables follows a Cauchy distribution (Kamerud et al., 1978), which does not have well defined first moment. Therefore, the mean and variance of the ratio coefficients across samples is influenced by extreme values often encountered during estimation.

When the variable $\boldsymbol{p}$ denotes price and β_p is fixed to -1, the utilities are money-metric. The coefficients of the non-price attributes in the utility function are interpretable as willingness-to-pay (WTP) for the attribute. This method is termed estimation in WTP space (Train and Weeks (2005), Scarpa et al. (2008)) in contrast to estimation in the parameter space where the scale of the error term in the utility equation is fixed. Train and Weeks (2005) report that the models estimated in WTP space have better fit and are less prone to extreme values in the WTP coefficients. Following a similar approach, we estimate in the WTP space, by fixing the coefficient $\beta_p = 1$, for the endogenous variable $\boldsymbol{p}$ across the different models. Accordingly, the estimated scale of the utility function is reported in the results.

Table 1 presents the results from the first set of experiments that validate the data generation process. Model $M-1$ is estimated with complete information where all the parameters including the alternative specific constant, β_y, are recovered as evidenced by the low t-values. In the second sub-experiment $M-2(a)$, the variable $\boldsymbol{\xi}$ is omitted, thereby causing endogeneity. Ignoring endogeneity during estimation introduces bias in the results. $\hat{\beta}_x$ is significantly different from its true value (t-value $= -13.1$). The sub-experiment, $M-2(b)$ applies the 2-stage control function method (2SCF) (Rivers and Vuong, 1988). We are able to recover the parameters β_x and β_y, even though the data suffers from endogeneity as is seen in the high t-value of 3.95 in the coefficient for the control function residual term, β_δ (Rivers and Vuong, 1988). In addition, the scale of the error term is not recovered after applying control function method (Guevara and Ben-Akiva, 2012), as expected.

The second set of experiments, models $M-3$ and $M-4$ are estimated with the twin data deficiencies, missing $\boldsymbol{p}$ and omitted variable $\boldsymbol{\xi}$. As there is an underlying linear regression model for variable $\boldsymbol{p}$ for both choice alternatives, p_{2n} is imputed multiple times ($D = 20$) during estimation. Table 2 summarizes the estimation results. When the multiple imputation algorithm alone is applied in model $M-3$, we encounter bias in the estimates in both β_x and β_y. In fact, the bias in β_x is statistically significant with a t-value of -2.07 forcing us to reject the null hypothesis, $\beta_x = 2$. However, by applying the hybrid algorithm in model $M-4$, we are able to recover the true values of β_x (1.96) and the alternative specific constant β_y (0.480) with a statistically insignificant bias. The experiment demonstrates that the hybrid algorithm is able to simultaneously correct for endogeneity and missing information. However, the estimated coefficient, $\hat{\beta}_x$ has a high standard error which indicates a loss in efficiency due to the missing information. The box-plot of $\hat{\beta}_x$ estimated for different models is presented in Fig. 1. Further, in $M-4$,the t-value of the coefficient of the residual variable, $\hat{\beta}_\delta$, was not found to be statistically different from zero, possibly due to a large percentage (50%) of missing values.

To assess the impact of the proportion of missing values on the model parameters, we estimated model $M-4$ by varying the extent of missing values of the endogenous variable $\boldsymbol{p}$. Missing values in the information of the variable $\boldsymbol{p}$ for both alternatives were randomly inserted at three levels −10%, 30% and 50%. The results are shown in Table 3. In all three instances, we were able to recover the true parameters for the exogenous variables, $\boldsymbol{x}$ and $\boldsymbol{y}$, with insignificant statistical bias. However, as the proportion of missing values in the endogenous variable $\boldsymbol{p}$ increases, the confidence in the estimated parameter

Table 2
Models with missing information on endogenous variable.

	M-3 : Multiple imputation		M-4 : Hybrid algorithm	
Par. (true value)	Value (std.error)	t-value*	Value (std.error)	t-value*
$\hat{\beta}_y$ $(\beta_y = 0.5)$	0.394 (0.114)	−0.923	0.480 (0.162)	−0.126
$\hat{\beta}_x$ $(\beta_x = 2)$	1.62 (0.185)	−2.07	1.96 (0.385)	−0.104
$\hat{\mu}$ $(\mu = 1)$	0.314 (0.0204)	−33.6	0.262 (0.0430)	−17.2
$\hat{\beta}_\delta$ $(\beta_\delta = 0)$			0.270 (0.225)	1.20
Sample size	8000		8000	
No. of imputations	20		20	
No. of simulation runs	100		100	
Init. log-likelihood	−5545.18		−5545.18	
Final log-likelihood	−5150.61		−5147.99	

(*) t-tests are against the respective population values.

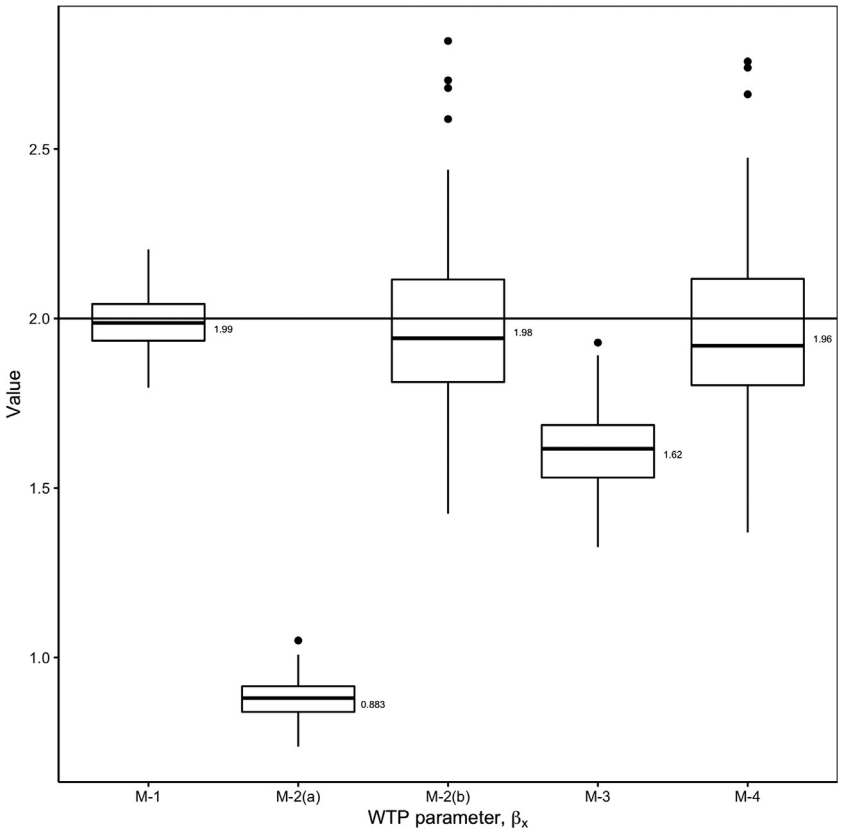

Fig. 1. Willingness-to-pay coefficients β_x for different models.

Table 3
Impact of proportion of missing values on model $M-4$.

	10% missing values		30% missing values		50% missing values	
Par. (true value)	Value (std.error)	t-value*	Value (std.error)	t-value*	Value (std.error)	t-value*
$\hat{\beta}_y$ $(\beta_y = 0.5)$	0.479 (0.152)	−0.137	0.478 (0.156)	−0.139	0.480 (0.161)	−0.121
$\hat{\beta}_x$ $(\beta_x = 2)$	1.97 (0.372)	−0.0728	1.97 (0.383)	−0.0659	1.97 (0.398)	−0.0589
$\hat{\mu}$ $(\mu = 1)$	0.303 (0.047)	−14.8	0.276 (0.0441)	−16.4	0.260 (0.436)	−17.0
$\hat{\beta}_\delta$ $(\beta_\delta = 0)$	1.31 (0.373)	3.51	0.792 (0.301)	2.63	0.278 (0.235)	1.18
Sample size	8000		8000		8000	
No. of imputations	20		20		20	
No. of simulation runs	100		100		100	
Init. log-likelihood	−5545.18		−5545.18		−5545.18	
Final log-likelihood	−4551.57		− 4904.08		−5147.60	

(*) t-tests are against the respective population values

for the residual term, $\hat{\beta}_{\delta}$ reduces. The residual term, $\hat{\delta}$, may no longer be a consistent estimator of the omitted attribute, thereby increasing the standard errors.

Finally, as a preliminary analysis of the impact that sample size and number of imputations may have in the proposed method, the experiments, $M-3$ and $M-4$, were repeated by reducing the sample size and the number of imputations to 4000 and $D=10$ respectively. Our conclusions remain unchanged.

5. Case study : modeling of overnight parking subscription of heavy commercial vehicles

5.1. Description

In this section, we apply the hybrid algorithm to model the overnight parking subscription of heavy commercial vehicles (HCV) in an urban area. The case study is set in Singapore, an island city-state in south-east Asia. Singapore has approximately 32,000 HCVs which are parked in nearly 42,000 overnight parking spaces distributed in 526 Traffic Analysis Zones (TAZ) of a total 1169 TAZs in the city. Availability of parking demand data for every HCV is facilitated by Singapore's garage law (Enoch, 2004). It mandates the vehicle owner to purchase a monthly subscription to overnight parking and report it to the transport authority. The parking supply is owned by both public and private agencies. The two public parking operators in Singapore are termed public-1 and public-2, while, the private operators are categorized as private-open and private-closed, based on the availability of parking price information. The parking prices of private-closed alternatives are missing, in contrast to private-open parking alternatives, whose prices are listed in a website operated by the transport authorities (OneMotoring, 2017). Gopalakrishnan (2019) provides a detailed description of the problem.

5.2. Model specification

The problem is modeled as a multinomial logit model in which the alternatives correspond to 539 overnight parking locations, which are aggregated at TAZ and park-types (Public-1, Public-2, Private-open and Private-closed) levels. While various potential variations of the underlying model may be feasible, we decided to maintain its core as simple as possible, for illustrative purposes.

Logit models are based on the assumption of Independence of Irrelevant Alternatives (IIA) property. Nevertheless, in this case, the aggregation of continuous areas into discrete alternatives such as TAZ may lead to violation of the IIA property, compromising efficiency and forecasting power, but not consistency. We acknowledge that spatial models other than logit have been estimated in literature (Bhat and Guo, 2004), but decided to maintain the IIA assumption to concentrate this illustrative example on the endogeneity and missing data problem that motivates this research.

The utility derived from choosing a specific parking alternative is modeled as a function of vehicle, vehicle-owner characteristics and attributes of the parking location. The distance of the parking location from the vehicle-owner premises, age of vehicle, weight, fleet size, industry group of the vehicle-owner, price of parking, and park-type are included in the utility function. Alternative specific dummy (or constant ASC) variables are included for the different park-types with public-1 as the base value. Vehicle and vehicle-owner characteristics are included in the utility function for specific park-types - older vehicles preferring public-1 parking locations while industry category is an influencing variable in private-closed parking. Further, as land is scarce in Singapore, we hypothesize that vehicle owners with a large fleet prefer public or private-open parking while heavier (and larger) vehicles, such as cranes, and construction vehicles, are inclined to park in private-closed parking.

The proposed model suffers from both endogeneity and missing data problems. Price endogeneity arises from the omission of quality attributes that are correlated with price and/or the simultaneous determination of the supply and demand for overnight parking. As it will be shown, the price endogeneity problem is severe enough to reflect in that uncorrected models have estimates of the price coefficient that are statistically equal to zero. While it cannot be discarded that other sources of endogeneity may exist for other variables, there is no a-priory reason to think that that might be relevant, and the results of the uncorrected model seem to support that assertion. The missing data problem arises from the fact that prices of private-closed alternatives are unavailable.

Endogeneity and missing data can be addressed in this case study through the control-function and the multiple imputation approaches, respectively. The mix of both problems can be addressed by the hybrid method proposed in this article. In Section 5.3 we discuss the selection of instrumental variables needed for the correction of endogeneity and in Section 5.4 we present the price imputation model needed for the missing data problem. In Section 5.5 we present the results of the proposed combined approach.

It worth noting that, thanks to the IIA assumption behind the logit model of the case study, and to that price is missing only for private-closed alternatives that have an ASC, the missing data problem may be partially addressed in this case by ignoring all the observations and alternatives with missing values from the sample. However, despite consistency may be attained in such a case, excluding observations and alternatives with missing prices in a partial model would prevent estimating parameters that are specific to private-closed alternatives, as well as preclude predicting the demand for that alternative and the total demand for other alternatives. Besides, if observations and alternatives with missing values are ignored, efficiency will be affected, and it would also be impossible to identify potential non-IIA structures that may involve the missing alternative. The hybrid approach we propose in this article can circumvent all these limitations.

Table 4
Estimated coefficients of price imputation equation.

Coefficients	Estimate	Std.Error	*t*-value
Intercept	231.0	115.0	2.00
Capacity of parking alternative (log)	10.1	8.70	1.16
Occupancy of parking alternative (%)	7.42	37.9	0.196
Occupancy of parking alternatives in adjacent TAZ (%)	−49.8	50.1	−0.995
Driver accessibility log-sum	22.8	5.33	4.28
Industrial area accessibility log-sum	−4.86	3.30	−1.47
Floor area of manufacturing industries in TAZ ('000 sq.m)	−0.0885	0.0779	−1.14
Distance from CBD (km)	12.78	3.74	3.42
Region dummy variables			
Central region (excl. central areas)	−129.0	95.6	−1.35
East	−170.8	108.0	−1.58
North	−298.0	124.0	−2.41
North-east	−165.0	110.0	−1.51
West	−185.0	111.0	−1.67

R^2 : 0.332 Adjusted R^2 : 0.236 Sample size : 98.

5.3. Instrumental variable selection

The correction of endogeneity requires finding proper instrumental variables (Guevara and Ben-Akiva, 2012), which can be very challenging in practice (Bound et al., 1995; Mumbower et al., 2014) and becomes the critical stage in this regard. The instrument must be correlated with the endogenous variable and independent of the error term. Mumbower et al. (2014) delineate the instrumental variables into four categories, viz., cost-shifting, Stern-type, BLP-type and Hausman-type instruments. For the problem under study, we propose instrumental variables of the latter type with a shift.

Hausman-type price instruments relate the price of the alternative to prices of other alternatives that may share marginal costs but have different demand shifts (Guevara and Ben-Akiva, 2006; Guevara-Cue, 2010; Nevo, 2000; Petrin and Train, 2010). The average of prices in different alternatives or markets could then be used as an instrument for this case study. However, since parking price information is missing in some cases, to have complete coverage, we need to gather this relationship from an alternative variable. We select parking lot occupancy in adjacent TAZ as it is negatively correlated with parking price ($\rho = -0.240$) in the sample, fulfilling one of the two conditions for a valid instrument variable. The second condition, independence of the instrument with the error terms, can be justified under assumption that other parking lots share marginal costs but, since they are located in different TAZ, they do not share demand shocks. A formal test for this hypothesis can be formulated only when the number of available instruments exceeds the number of endogenous variables (Guevara, 2018; Lung-Fei, 1992), something that was not feasible for the available dataset.

5.4. Parking price imputation model

The application of the multiple imputation method to address the missing data problem relies in assuming that the data is MAR. However, it can be argued that this may not strictly be the case in this application because the price data is missing for a whole type of parking alternative. Nevertheless, for this application we propose using the private-open data to build an imputation function for the private-close missing data. We postulate that this approach is appropriate because private-closed and private-open parking lots are often adjacent to each other. The underlying assumption in our application is that the parking prices are related to the attributes of the location and we use those attributes to built the conditional distribution from which we draw the imputations. A similar flavor may be found in the rents for tenant-occupied or owner-occupied dwelling units; the rents for tenant-occupied home are easily available in property search sites while the equivalent prices for owner-occupied units may have to be imputed.

The private-open parking supply with known parking price information was distributed in over 98 TAZs. We assume the existence of a linear price equation relating the price of private-open parking alternatives to the exogenous variables and instruments. The estimated coefficients of the price imputation equation are presented in Table 4.

The model in Table 4 show that higher parking prices are associated with larger parking lots, which reflect the producer's incentive to increase parking supply when prices are high. Further, occupancy of the parking alternative is positively associated with parking price. When the occupancy in parking lots in adjacent TAZ is high, the parking prices are low reflecting the competition from adjacent parking lots. Parking prices are related positively to driver accessibility of a parking lot. The accessibility measures were developed as a log-sum variable (Sivakumar and Bhat, 2002; Gopalakrishnan, 2019) and estimated in a logit model with complete cases. In addition, the floor area occupied by manufacturing industry in a TAZ is negatively correlated with parking prices. This reflects the extent of additional parking supply possible in a TAZ. The distance from Central Business District(CBD) and the region depict the region-wise differences in parking prices. The imputation model has a low R^2, which indicates a poor goodness of fit reflecting the inherent difficulty in imputing prices. This also does not preclude the possibility of large errors when imputing prices for private-closed alternatives.

Table 5
Results with proposed hybrid method.

	Multiple imputation			Hybrid algorithm		
Parameter[a]	Value	Std.error	*t*-value	Value	Std.error	*t*-value
Distance from owner to parking location, ln(km), (1,2,3,4)	−1.36	0.0272	−50.3	−1.36	0.0272	−50
Parking price, *ln(SGD)*, (1,2,3,4)	−0.00582	0.0923	−0.063	−2.37	0.459	−5.15
Private-closed (dummy var.), (4)	1.99	0.211	9.47	2.5	0.232	10.8
Private-open (dummy var.), (3)	−0.126	0.147	−0.858	0.306	0.161	1.91
Public-2 (dummy var.), (2)	−0.37	0.14	−2.64	0.455	0.211	2.15
Vehicle age as of 1-Jan-2017, years, (1)	0.0265	0.00947	2.8	0.0268	0.00949	2.82
Transportation industry (dummy var.) (4)	−0.597	0.105	−5.66	−0.599	0.106	−5.67
Manufacturing industry (dummy var.), (4)	0.365	0.153	2.38	0.364	0.153	2.38
Vehicle weight, *ln(Tonnes)*, (4)	−0.545	0.0603	−9.03	−0.546	0.0604	−9.03
Control function residual, (3,4)				2.49	0.485	5.13
Fleet size, *ln*, (3,4)	−0.0453	0.0254	−1.78	−0.0468	0.0255	−1.83
Sample size	3830			3830		
Excluded observations	170			170		
Initial log likelihood	−22324.17			−22324.17		
Final log likelihood	−18256.81			−18225.03		
Rho-square for the init. model	0.182			0.184		

[a] Number in parenthesis denotes the park-type of the alternative where the variable is included. 1.Public-1, 2.Public-2, 3. Private-open, 4. Private-closed. The transformation of the variable is also mentioned in italics. The alternative specific constant for park-type, Public-1, is the base value. Since alternatives are aggregations of all parking of types in a TAZ, a log size variable correction (with fixed coefficient equal to 1) was included in both models.

5.5. Proposed hybrid method

The estimation of the proposed hybrid method to address both endogeneity and missing data is applied in two steps. Step 1, corresponds to the imputation model detailed in Section 5.4. In Step 2, the imputed prices are used to estimate two models - a logit model with multiple imputation but without correcting for endogeneity and the other by applying the hybrid algorithm developed in Section 3.

The prices of private -closed alternatives were imputed using the price imputation model in Table 4. The results of the two logit models - estimated with multiple imputation alone (without endogeneity correction) and the hybrid algorithm, with mean coefficient values and standard errors computed using Eq. (7), are presented in Table 5.

Applying the multiple imputation algorithm without control function corrections, at the left of Table 5, results in price coefficients which are close to zero (with a low *t*-value). This indicates presence of endogeneity in the model. Applying multiple imputation algorithm alone is insufficient to obtain consistent parameter estimates. After application of the hybrid algorithm, at the right of Table 5, price sensitivity coefficient increases (in magnitude) from −0.00582 to −2.37 (*t*-value of −5.15), a change in an order of magnitude. The coefficient of the control function residual term has a high *t*-value of 5.13 affirming endogeneity in the data (Rivers and Vuong, 1988).

The model with real data was estimated in the parameter space using pandas Biogeme (Bierlaire, 2020), instead of the WTP space. The magnitude of coefficients of distance-sensitivity, age of vehicle, industry-category changed marginally across both the methods - multiple imputation and hybrid algorithms - indicating minimal change in scale (Guevara and Ben-Akiva, 2012).

The estimated coefficients, from the application hybrid algorithm, provide policy insights into the trade-off between price and distance of the parking location from the owner premises. When the distance from the location of the vehicle-owner to the parking lot increases by 50%, the price that the owner is willing to pay decreases to 42% ($= e^{-\frac{1.36}{2.37}*1.5}$) of the current price of the parking lot. In monetary terms, an increase in distance of the parking location (from the owner premises) from 2 km to 3 km would reduce the attractiveness (price)of the parking location from say, SGD100 per month to SGD42 per month.

6. Conclusion

In this paper, a hybrid algorithm has been developed to correct for the data deficiencies of missing information and endogeneity occurring simultaneously. We correct and extend the state-of-the-art method (Steimetz and Brownstone, 2005) for multiple imputation in discrete choice modeling literature by embedding the control function method. The algorithm, which combines the features of multiple imputation and control function methods is shown to correct for both missing values and endogeneity in the same variable in a Monte-Carlo set-up. The algorithm is robust even when the alternatives are labeled.

Further, the algorithm, when applied to parking data, corrects both the signs and magnitude of price sensitivity coefficients. The results from real data also demonstrate that application of the multiple imputation method alone is insufficient to overcome price endogeneity.

We believe the methodology can easily be extended to solve similar problems in other domains. For example, in airline industry studies, fares of competing airlines may be missing (Hotle et al., 2015), while other attributes like schedule, type of aircraft and connections are observable. Thus, even under situations of an additional cause of endogeneity in missing attribute correlated with the air-fare, WTP parameters can be consistently estimated with the proposed method. Similarly in many problems in social science research, determining individual income elasticity is an important objective. However, it is well-known that survey respondents often either refuse or misreport income leading to missing values in the data (Bhat, 1994; Schenker et al., 2006; Kim et al., 2007; Sanko et al., 2014). Further, when the agent utility specification includes an omitted attribute, correlated with income, endogeneity is encountered which can be corrected with the hybrid algorithm. With the proliferation of sensors in data collection, the problem of missing information (e.g, Kong et al., 2013; Pan et al., 2010) and endogeneity (Fan et al., 2014; Gandomi and Haider, 2015; Guevara et al., 2017) occurring together may be even more frequent.

The algorithm is limited by the assumption of normality of the omitted attribute. In real-world data, the distribution of an omitted variable is unknown and might influence the model results in finite samples. Further, the assumption of the existence of a price equation might not hold in practical applications. Moreover, finding a valid instrument to correct for endogeneity can be a challenge when using real-world data. The results reiterate the importance of data collection as reflected in loss of efficiency when using imputation methods.

Declaration of Competing Interest

None.

CRediT authorship contribution statement

Raja Gopalakrishnan: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization. **C. Angelo Guevara:** Conceptualization, Methodology, Investigation, Writing - review & editing, Supervision. **Moshe Ben-Akiva:** Methodology, Resources, Supervision, Project administration.

Acknowledgements

This research was funded in part by the Singapore Ministry of National Development and the National Research Foundation, Prime Minister's Office under the Land and Liveability National Innovation Challenge Research (L2NIC) Programme, grant number L2 NIC Award No. L2 NICTDF1-2016-1. The authors acknowledge the support of the Urban Redevelopment Authority of Singapore, Land Transport Authority of Singapore, and Housing and Development Board of Singapore. The first author acknowledges the financial assistance received from the Ministry of Railways, Government of India. Second author gratefully acknowledges financial support from FONDECYT-Chile, grant 1191104 and ANID PIA/ BASAL AFB180003. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) only and do not reflect the views of the Singapore Ministry of National Development, National Research Foundation, Prime Minister's Office, Urban Redevelopment Authority of Singapore, Land Transport Authority of Singapore, and Housing and Development Board of Singapore.

References

Berry, S., Levinsohn, J., Pakes, A., 1995. Automobile prices in market equilibrium. Econometrica 63 (4), 841–890.
Bhat, C.R., 1994. Imputing a continuous income variable from grouped and missing income observations. Econ. Lett. 46 (4), 311–319.
Bhat, C.R., Guo, J., 2004. A mixed spatially correlated logit model: formulation and application to residential choice modeling. Transp. Res. Part B 38 (2), 147–168.
Bierlaire, M., 2020. A short introduction to PandasBiogeme. Technical report TRANSP-OR 200605.
Bodner, T.E., 2008. What improves with increased missing data imputations? Struct. Equ. Model. 15 (4), 651–675.
Bound, J., Jaeger, D.A., Baker, R.M., 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. J. Am. Stat. Assoc. 90 (430), 443–450.
Cavallo, A., 2018. More amazon effects: online competition and pricing behaviors. Technical Report. National Bureau of Economic Research.
Enoch, M., 2004. Managing transportation demand in Singapore. Traffic Eng. Control 45 (3), 100–102.
Fan, J., Han, F., Liu, H., 2014. Challenges of big data analysis. Natl. Sci. Rev. 1 (2), 293–314.
Ferreira, F., 2010. You can take it with you: proposition 13 tax benefits, residential mobility, and willingness to pay for housing amenities. J. Public Econ. 94 (9–10), 661–673.
Gandomi, A., Haider, M., 2015. Beyond the hype: big data concepts, methods, and analytics. Int. J. Inf. Manag. 35 (2), 137–144.
Goolsbee, A., Petrin, A., 2004. The consumer gains from direct broadcast satellites and the competition with cable TV. Econometrica 72 (2), 351–381.
Gopalakrishnan, R., 2019. Overnight parking of commercial vehicles in cities : data, models, and policy analysis. Singapore University of Technology and Design Ph.D. Thesis.
Greene, W.H., 2011. Econometric Analysis ([7. sup. th] Edition). Upper Saddle River, NJ: Prentice Hall.
Guevara, C., Ben-Akiva, M., 2006. Endogeneity in residential location choice models. Transp. Res. Rec. 1977 (1977), 60–66.
Guevara, C., Tang, Y., Gao, S., 2017. The initial condition problem with complete history dependency in learning models for travel choices. Transp. Res. Procedia 23, 758–771.
Guevara, C.A., 2015. Critical assessment of five methods to correct for endogeneity in discrete-choice models. Transp. Res. Part A 82, 240–254.
Guevara, C.A., 2018. Overidentification tests for the exogeneity of instruments in discrete choice models. Transp. Res. Part B 114, 241–253.
Guevara, C.A., Ben-Akiva, M.E., 2012. Change of scale and forecasting with the control-function method in logit models. Transp. Sci. 46 (3), 425–437.

Guevara-Cue, C.A., 2010. Endogeneity and sampling of alternatives in spatial choice models. Massachusetts Institute of Technology Ph.D. Thesis.
Guo, C., Gu, X., Li, Q., Qu, J., Zhang, L., 2018. Traffic time prediction based on imputation algorithms for missing values. In: 2018 International Conference on Network Infrastructure and Digital Content (IC-NIDC), pp. 223–228.
Heckman, J.J., 1978. Dummy endogenous variables in a simultaneous equation system. Econometrica 46 (4), 931–959.
von Hippel, P.T., 2018. How many imputations do you need? A two-stage calculation using a quadratic rule. Sociol. Methods Res. 49 (3), 699–718.
Hotle, S.L., Castillo, M., Garrow, L.A., Higgins, M.J., 2015. The impact of advance purchase deadlines on airline consumers' search and purchase behaviors. Transp. Res. Part A 82, 1–16.
Kamerud, D.B., Deaton, L., Bosch, A., Driscoll, M., Young, D., Gbur, E., Goodsell, C., Hertz, E., Rogers, G., Skalsky, M., et al., 1978. Random variable x/y, x, y normal. Am. Math. Mon. 85 (3), 206–208.
Kim, S., Egerter, S., Cubbin, C., Takahashi, E.R., Braveman, P., 2007. Potential implications of missing income data in population-based surveys: an example from a postpartum survey in california. Public Health Rep. 122 (6), 753–763.
Kong, L., Xia, M., Liu, X.-Y., Wu, M.-Y., Liu, X., 2013. Data loss and reconstruction in sensor networks. In: 2013 Proceedings IEEE INFOCOM. IEEE, pp. 1654–1662.
Li, W., 2014. Modeling Household Residential Choice Using Multiple Imputation. Massachusetts Institute of Techology Masters Thesis.
Little, R.J.A., Rubin, D.B., 2002. Statistical Analysis with Missing Data: Second Edition. John Wiley and Sons.
Lung-Fei, L., 1992. Amemiya's generalized least squares and tests of overidentification in simultaneous equation models with qualitative or limited dependent variables. Econom. Rev. 11 (3), 319–328.
Mumbower, S., Garrow, L.A., Higgins, M.J., 2014. Estimating flight-level price elasticities using online airline data: a first step toward integrating pricing, demand, and revenue optimization. Transp. Res. Part A 66 (1), 196–212.
Nevo, A., 2000. Mergers with differentiated products: the case of the ready-to-eat cereal industry. RAND J. Econ. 31 (3), 395–421.
OneMotoring, 2017. Heavy Vehicle Parking. https://www.onemotoring.com. Last accessed 30-Jun-2017.
Pan, L., Li, J., et al., 2010. K-nearest neighbor based missing data estimation algorithm in wireless sensor networks. Wirel. Sensor Netw. 2 (02), 115.
Park, S., Gupta, S., 2009. Simulated maximum likelihood estimator for the random coefficient logit model using aggregate data. J. Mark. Res. 46 (4), 531–542.
Petrin, A., Train, K., 2010. A control function approach to endogeneity in consumer choice models. J. Mark. Res. 47 (1), 3–13.
R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Rivers, D., Vuong, Q.H., 1988. Limited information estimators and exogeneity tests for simultaneous probit models. J. Econom. 39 (3), 347–366.
Rubin, D.B., 1987. The calculation of posterior distributions by data augmentation: comment: a noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: the SIR algorithm. J. Am. Stat. Assoc. 82 (398), 543–546.
Rubin, D.B., Schenker, N., 1986. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. J. Am. Stat. Assoc. 81 (394), 366–374.
Ruud, P.A., 1983. Sufficient conditions for the consistency of maximum likelihood estimation despite misspecification of distribution in multinomial discrete choice models. Econometrica 51 (1), 225–228.
Sanko, N., Hess, S., Dumont, J., Daly, A., 2014. Contrasting imputation with a latent variable approach to dealing with missing income in choice models. J. Choice Model. 12, 47–57.
Scarpa, R., Thiene, M., Train, K., 2008. Utility in willingness to pay space: a tool to address confounding random scale effects in destination choice to the Alps. Am. J. Agric. Econ. 90 (4), 994–1010.
Schenker, N., Raghunathan, T.E., Chiu, P.-L., Makuc, D.M., Zhang, G., Cohen, A.J., 2006. Multiple imputation of missing income data in the national health interview survey. J. Am. Stat. Assoc. 101 (475), 924–933.
Schenker, N., Welsh, A., 1988. Asymptotic results for multiple imputation. Ann. Stat. 16 (4), 1550–1566.
Sermons, M.W., Koppelman, F.S., 2001. Representing the differences between female and male commute behavior in residential location choice models. J. Transp. Geogr. 9 (2), 101–110.
Sivakumar, A., Bhat, C., 2002. Fractional split-distribution model for statewide commodity-flow analysis. Transp. Res. Rec. 1790 (02), 80–88.
Steimetz, S.S., Brownstone, D., 2005. Estimating commuters' "value of time" with noisy data: a multiple imputation approach. Transp. Res. Part B 39 (10), 865–889.
Tirachini, A., Hensher, D.A., Rose, J.M., 2013. Crowding in public transport systems: effects on users, operation and implications for the estimation of demand. Transp. Res. Part A 53, 36–52.
Train, K., Weeks, M., 2005. Discrete choice models in preference space and willingness-to-pay space. In: Applications of Simulation Methods in Environmental and Resource Economics. Springer, pp. 1–16.
Train, K.E., 2009. Discrete Choice Methods with Simulation. Cambridge University Press.
Villas-Boas, J.M., Winer, R.S., 1999. Endogeneity in brand choice models. Manag. Sci. 45 (10), 1324–1338.
Wardman, M., Whelan, G., 2011. Twenty years of rail crowding valuation studies: evidence and lessons from British experience. Transp. Rev. 31 (3), 379–398.