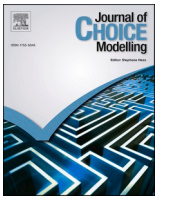
Journal of Choice Modelling 42 (2022) 100342

Contents lists available at ScienceDirect

Journal of Choice Modelling

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



Characterizing the impact of discrete indicators to correct for endogeneity in discrete choice models

Thomas E. Guerrero [a,b,c,*], C. Angelo Guevara [d,f], Elisabetta Cherchi [c], Juan de Dios Ortúzar [b,e,f]

[a] Department of Civil Engineering, Universidad Francisco de Paula Santander Ocaña, Sede el Algodonal Ocaña, Colombia
[b] Department of Transport Engineering and Logistics, Pontificia Universidad Católica de Chile, Santiago, Chile
[c] School of Engineering, Future Mobility Group, Newcastle University, Newcastle upon Tyne, UK
[d] Department of Civil Engineering, Universidad de Chile, Santiago, Chile
[e] BRT+ Centre of Excellence, Santiago, Chile
[f] Instituto Sistemas Complejos de Ingeniería (ISCI), Santiago, Chile

ARTICLE INFO

Keywords:
Endogeneity
Multiple indicator solutions method
Discrete and continuous indicators
Discrete choice models

ABSTRACT

Endogeneity is a common problem in econometric modelling that may lead to estimating inconsistent parameters. In the scientific literature, the Multiple Indicator Solutions (MIS) method is used to correct for endogeneity. This approach uses indicators that, in practice, tend to be collected as discrete using Likert scales; however, theoretically, the MIS method is derived considering continuous indicators. To close this research gap, this paper focuses on characterizing the impact of discrete indicators when correcting for endogeneity using the MIS method in the case of discrete choice models (DCM). Our findings show that (*i*) under some conditions, using discrete indicators instead of continuous ones seems not to be a problem, however, (*ii*) there is also evidence that indicates that the correction could fail under not unusual circumstances.

1. Introduction and motivation

Endogeneity arises when one or more of the explanatory variables of an econometric model are correlated with the model's error term. This may be caused by measurement/specification errors, omitted attributes and self-selection among other reasons (Guevara, 2015). The main problem of models affected by endogeneity is that their parameters may be inconsistent, leading to faulty policy analysis, forecasts, conclusions and/or behavioural assessments (Guevara and Ben-Akiva 2006).

The correction of endogeneity has been studied in depth in the case of linear models, where the problem has been considered from several viewpoints (Stock and Yogo, 2005; Ebbes et al., 2011). In the case of discrete choice models (DCM), which are highly non-linear, some research gaps have been closed in recent years (see e.g., Guevara and Polanco, 2016), and work by Guerrero et al. (2021), Wen and Chen (2017), Lurkin et al. (2017) and Mumbower et al. (2014), have successfully applied practical corrections for endogeneity in the transport field. Notwithstanding, several open questions still remain.

There are several methods available to correct for endogeneity in DCM, such as Berry-Levinsohn-Pakes - BLP (Berry et al., 1995), Latent Variables - LV (Ben-Akiva et al., 2002), Maximum Likelihood (Park and Gupta, 2009), Control Function - CF (Petrin and Train,

* Corresponding author. Department of Civil Engineering, Universidad Francisco de Paula Santander Ocaña, Sede el Algodonal Ocaña, Colombia.
E-mail addresses: teguerrerob@ufpso.edu.co (T.E. Guerrero), crguevar@ing.uchile.cl (C.A. Guevara), elisabetta.cherchi@ncl.ac.uk (E. Cherchi), jos@ing.puc.cl (J.D. Ortúzar).

https://doi.org/10.1016/j.jocm.2021.100342
Received 27 January 2021; Received in revised form 28 December 2021; Accepted 29 December 2021
Available online 5 January 2022

2010), Proxies (Guevara, 2015), and Multiple Indicator Solution - MIS (Guevara and Polanco, 2016). Louviere et al. (2005) provided an extensive compendium of advances in the treatment of endogeneity in DCM. The decision to use one or another method is based on what information is available, the assumptions the researcher is willing to make, and the associated computational costs (Guevara, 2015).

In this paper we consider the MIS method for correcting endogeneity in DCM. The MIS method is based on the use of indicators, which often come from surveys designed for knowing respondents' attitudes and/or perceptions about their decision making (Bahamonde-Birke et al., 2017). Indicators have been typically collected using Likert (1932) or verbal scales (Glerum et al., 2014), and they can be *attitudinal* because these represent the characteristic of the individuals toward life (Walker and Ben-Akiva, 2002; Daly et al., 2012; Bahamonde-Birke et al., 2017) or *perceptual* because they are exclusively related to the way certain alternatives are perceived, namely, they are intrinsically associated with an alternative (Bolduc and Alvarez-Daziano, 2009; Yáñez et al., 2010; Raveau et al., 2010).

Guevara (2015) shows that the main advantage of the MIS method in the correction of endogeneity is that it transforms the problem in a way that makes it possible to gather proper instrumental variables directly from the respondent. By this, the MIS method avoids having to collect or build instrumental variables from existing information, which is always hard and sometimes even controversial. The difficulty resides in that the instrumental variables need to satisfy two conflicting criteria: (*i*) to be exogenous (independent of the error term of the model) and (*ii*) to be relevant (strongly correlated with the endogenous variable). In practice, having proper instruments may be impossible and if the instruments are endogenous or *weak* (i.e., not sufficiently strongly correlated with the endogenous variable), their use can yield results as bad as when using an endogenous model (see, for example, Guevara, 2018; Guerrero et al., 2021). MIS offers a practical approach to address this problem. However, the relative easiness in the collection of indicators for the MIS comes with a caveat. In theory, to apply the MIS method to correct for endogeneity in DCM, the indicators must be continuous (Guevara, 2015), since this is a mathematical requirement for its derivation (Wooldridge, 2010), and this may be hard to fulfil.

The problem is that, in practice, the indicators tend to be discrete since they are typically obtained through Likert scales. Although there is some preliminary empirical evidence suggesting that discrete indicators can be as good as continuous ones for correcting endogeneity with the MIS method in DCM (Guevara and Polanco, 2016; Fernández-Antolín et al., 2016), it should be clear that this is only an approximation (Guevara, 2015). So, this part of the research will focus on characterizing the impact of using discrete indicators to correct for endogeneity with the MIS method in DCM. For this, we will correct DCM, with discrete and continuous indicators, using both real and simulated data, and compare the results to determine the possible impacts and differences of each approach. This will allow identifying some of the conditions under which discrete indicators could work adequately, providing support or refuting existing empirical evidence.

To do our tests we will use a purposely designed stated preference (SP) survey and Monte Carlo simulation. The SP survey was designed for the context of departure time choice, considering the main explanatory variables in this modelling context; that is, travel time, cost, additional travel time and schedule delay (Arellana et al., 2012), where the last variable followed the scheduling model of Small (1982). The Monte Carlo experiments were designed to test the effect of: (*i*) different criteria for the discretization of latent continuous indicators, (*ii*) the sample size and (*iii*) the distribution of the indicator.

The rest of the paper is organised as follows. Section 2 details the use of the MIS method in the case of DCM. The Monte Carlo experiment to assess the performance of discrete indicators using the MIS method is described in Section 3. In Section 4, we show the application of the MIS method for a departure time choice model estimated from SP data. In the final section we discuss the main findings, conclusions, and future research directions.

2. The MIS approach in discrete choice modelling

The MIS approach was initially introduced by Wooldridge (2010) for linear models, and later Guevara and Polanco (2016) extended it to DCM. To apply the MIS method at least two indicators are needed for each endogenous variable considered (Wooldridge 2010; Guevara and Polanco 2016). Therefore, given that we consider an *endogenous variable* (t_{in}) in our explanation then we must have available at least *two indicators* (I_{1in} and I_{2in}). This is an essential requirement of this approach. For explanatory purposes, let us consider as modelling context, a departure time choice decision represented by the DCM with utility function (U_{in}) as shown in (1):

$$U_{in} = ASC_i + \beta_t t_{in} + \beta_c c_{in} + \beta_x x_{in} + e_{in}, \tag{1}$$

where ASC_i is an alternative specific constant for alternative *i*, β_t, β_c, and β_x are parameters to be estimated. t_{in} and c_{in} (representing, for example, time and cost) are the model's explanatory variables. x_{in} can represent a qualitative attribute or the combination of several of them (e.g. schedule delay, additional travel time, safety, comfort or reliability), and e_{in} is an exogenous error term; the subscript *n* represents the individual. Although our SP experiment incorporated the variables sd_{in} (schedule delay) and v_{in} (additional travel time experienced in any trips during the week) in the choice task, they will be omitted in this section to simplify our explanation. For explanatory purposes, we will assume that x_{in} replaces the effects of both sd_{in} and v_{in}. This fact does not affect the mathematical derivation of the method. On the other hand, we will assume that t_{in} and c_{in} represent a set of known (measurable) attributes whereas the variable x_{in} is unknown to the modeller and correlated with t_{in}. Therefore, the modeller's specification would be as follows:

$$U_{in} = ASC_i + \beta_t t_{in} + \beta_c c_{in} + \varepsilon_{in}, \tag{2}$$

where the error term ε_{in} contains both $\beta_x x_{in}$ and e_{in}. Thus, as the error term ε_{in} is correlated with t_{in} in (2) through x_{in}, the DCM will

suffer from endogeneity as a result of omitting x_{in}; therefore, by definition, t_{in} is endogenous. Now, let us suppose that the variable x_{in} and the error terms ε_{in} can explain two indicators (I_{1in} and I_{2in}), as shown in (3) and (4):

$$I_{1in} = \alpha_1 + \alpha_{1x} x_{in} + \varepsilon_{1in} \tag{3}$$

$$I_{2in} = \alpha_2 + \alpha_{2x} x_{in} + \varepsilon_{2in} \tag{4}$$

For consistency, to apply the MIS method we will also assume that in this case the pairs of variables (x_{in}, ε_{1in}), (t_{in}, ε_{1in}), (x_{in}, ε_{2in}), (t_{in}, ε_{2in}) and (ε_{1in}, ε_{2in}) are mutually independent, and that the coefficients α_{1x} and α_{2x} are not null; α_1 and α_2 are intercepts to be estimated.

Although x_{in} in (2) is unknown to the researcher, for explanatory purposes we need to represent its effect in the model, and also to correct for the endogeneity resulting from its omission. In this case the correction is guaranteed because I_{1in} and I_{2in} can be expressed as a linear combination of x_{in} and other elements (α_1, α_2, ε_{1in} and ε_{2in}). For this, we can rewrite x_{in} as a function of I_{1in} using (3), that is, $x_{in} = \left(\frac{1}{\alpha_{1x}}\right)(I_{1in} - \alpha_1 - \varepsilon_{1in})$ and by defining $\theta_x = \frac{\beta_x}{\alpha_{1x}}$, the new expression for the utility function is given by (5):

$$U_{in} = ASC_i + \beta_t t_{in} + \beta_c c_{in} + \beta_x \left[\left(\frac{1}{\alpha_{1x}}\right)(I_{1in} - \alpha_1 - \varepsilon_{1in})\right] + e_{in} = ASC_i + \beta_t t_{in} + \beta_c c_{in} + \theta_x I_{1in} \underbrace{- \theta_x(\alpha_1 + \varepsilon_{1in}) + e_{in}}_{\omega_{in}} \tag{5}$$

The model in (5) tries to correct for the endogeneity caused by omitting x_{in} in (2); however, the modified model still suffers from a different source of endogeneity because the term I_{1in} is correlated with ω_{in} through ε_{1in}.

The variable t_{in} is not endogenous in (5) because x_{in} is no longer part of the error term; therefore, the only endogenous variable in (5) is I_{1in}, and to correct for endogeneity we need another instrumental variable. This may come from the second indicator I_{2in}, which by construction is correlated with I_{1in} only through x_{in}, but it is independent of the error term ω_{in} in (5). In linear models, this final correction for the MIS method is performed through *Two Stage Least Square* - TSLS (Wooldridge, 2010), but in DCM this is done using the CF method (Guevara and Polanco, 2016).

Thus, to apply the MIS method at least two indicators are needed for each endogenous variable considered. The MIS method uses one indicator to account for the omitted variable and the other as an instrument of the first one in the corrected model. Both effects are represented in (3) and (4). It can be noted that the omitted variable (x_{in}) correlates with both indicators. Then, in this way, I_{2in} can be used as an instrument of I_{1in} and vice versa.

In practice, for the endogenous DCM model in this example (2), the MIS method could be applied in two-stages as follows:

(i) Apply an ordinary least squares (OLS) regression to I_{1in} on I_{2in}, t_{in} and c_{in}, and obtain the residuals $\widehat{\delta}_{in}$. Both t_{in} and c_{in} must be included in this auxiliary regression.

(ii) Estimate the DCM considering $\widehat{\delta}_{in}$ and I_{1in} within the utility function.

Therefore, the DCM corrected for endogeneity using the MIS method (Guevara and Polanco, 2016) in two-stages would be as shown in (6) and (7):

$$I_{1in} = \alpha_t t_{in} + \alpha_c c_{in} + \alpha_{I_{2in}} I_{2in} + \delta_{in} \xrightarrow{OLS} \widehat{\delta}_{in} = I_{1in} - \widehat{I}_{1in} \tag{6}$$

$$U_{in} = \widehat{ASC}_i + \widehat{\beta}_t t_{in} + \widehat{\beta}_c c_{in} + \widehat{\beta}_{I_{1in}} I_{1in} + \widehat{\beta}_{\widehat{\delta}} \widehat{\delta}_{in} + \widetilde{\varepsilon}_{in} \tag{7}$$

It can be noted that in (7) the term $\widehat{\beta}_{\widehat{\delta}} \widehat{\delta}_{in} + \widetilde{\varepsilon}_{in}$ is an orthogonal decomposition of ω_{in} in (5), where the term $\widehat{\beta}_{\widehat{\delta}} \widehat{\delta}_{in}$ captures the endogenous effect of ω_{in} that was present in (5). In this way, no term in (7) is correlated with the error term $\widetilde{\varepsilon}_{in}$, and therefore, the endogeneity problem is accounted for.

3. A departure time SP experiment to assess the performance of discrete indicators using the MIS method

In this section, we describe the MIS method's application to a databank coming from a specially designed SP survey in the context of departure time choice. The idea is to illustrate how, in practical terms, the use of continuous versus discrete indicators affects the results when using real data. We only provide a general description of the databank, limited to the main issues relevant to applying the MIS method.

Our SP experiment considered a survey where respondents were asked to report, first, their socioeconomic characteristics and information associated with their weekly trips by car to work. Regarding the latter, respondents had to report the usual travel time (t), any additional travel time (v) experienced in any of the trips during the week (in other words, v can be interpreted as an additional time due to the congestion effects that only happens with some probability), and the usual departure time. This itinerary was considered as the *current* alternative. This information was also used as pivot to build two additional itineraries (one with an early departure time and another with a late departure time), that were eventually presented to respondents as alternatives in four hypothetical scenarios. To introduce a variable associated with the cost of travelling, we asked respondents to consider the existence of an urban toll, which depended on their trip departure times (Fig. 1a).

	Departure time to work	Travel time	Arrival time to work	Additional travel time experienced	Travel cost (COP)	Which itinerary would you choose for going to work?
Itinerary 1	6:00 am	45 min	6:45 am	55 min	$ 5200	☐
Itinerary 2	5:40 am	32 min	6:12 am	40 min	$ 3100	☐
Itinerary 3	6:30 am	37 min	7:07 am	46 min	$ 7700	☐

(a)

	Departure time to work	Travel time	Arrival time to work	Additional travel time experienced	How pleasant is this itinerary for you?	How convenient is this departure time for you?
Itinerary 1	6:00 am	45 min	6:45 am	55 min		
Itinerary 2	5:40 am	32 min	6:12 am	40 min		
Itinerary 3	6:30 am	37 min	7:07 am	46 min		

(b)

Fig. 1. Illustration SP choice experiment screens: (a) Example of choice task; (b) Grading to build indicators.

Departure time choice is typically modelled using the scheduling model formulated by Small (1982). Our questionnaire considered also asking for the preferred arrival time (PAT) to the respondents' jobs. As we knew the arrival time to the job (AT), because it is the departure time plus the travel time reported by the respondent, we were able to estimate the schedule delay early (*sde*) and the schedule delay late (*sdl*) attributes, as follows:

$$sde = Max\{(PAT) - (AT), 0\} \tag{8}$$

$$sdl = Max\{(AT) - (PAT), 0\} \tag{9}$$

Note that the values of (8) or (9) are either positive or zero. This implies that the individual will not experience disutility from rescheduling (Arellana et al., 2012; Thorhauge et al., 2016).

As our objective was to use the MIS method to correct for endogeneity, two indicators were collected for each set of itineraries shown. This part of the survey considered showing some (not all, to reduce the cognitive load) of the itineraries later presented in the choice tasks and asking the respondents to grade them using the two following indicators:

1) How pleasant is this itinerary for you?
2) How convenient is this departure time for you?

The graded itineraries did not include the cost variable, because the indicators were designed to capture the effect of schedule differences, travel time and additional travel time experienced on the itineraries (Fig. 1b). To complete the information for those itineraries that were not graded (to reduce the cognitive load), we used the multiple imputation approach proposed by Gopalakrishnan et al. (2020).

The hypothetical scenarios shown to respondents in the SP experiment, were generated using a D-efficient design (Rose and Bliemer, 2009), considering the principles of level balance and minimal overlap (Zwerina et al., 2005). We contacted a convenience sample, via the internet, and used the Qualtrics software to implement the survey. Each respondent was confronted with four hypothetical scenarios with three itineraries, where they had to choose the preferred one. With the data gathered, we estimated several models, some of which were corrected for endogeneity due to omitted attributes with the MIS method.

To study the impact of the discreteness of the indicators in the quality of the correction, we used three different scales for them, which were assigned at random to respondents. Thanks to the randomness, any difference in the quality of the correction attained in each case may be attributable to the degree of discreetness considered. The scale used were (*i*) from 0 to 5 (without decimals), with probability 50%; (*ii*) from 0 to 5 (including one decimal), with probability 25%; and (*iii*) from 0 to 100 (without decimals), with probability 25%. With scale (*i*) we could guarantee a purely discrete grade, whereas with scales (*ii*) and (*iii*) we could secure a more continuous grade. We did some preliminary runs to determine the performance of each grading scales for correcting endogeneity using the MIS method and found some numerical problems when using grading scale (*iii*). For this reason, we later scaled the data obtained from (*iii*), to convert it to the scale from 0 to 5 (including one decimal).

After data cleaning, the total number of valid choices recorded was 437 pseudo-observations graded with discrete indicators and 476 pseudo-observations graded with continuous indicators. The survey was applied in the five main cities of Colombia (Bogotá, Cali, Medellín, Bucaramanga and Barranquilla).

Our methodological contribution is related with the performance of the two types of indicators, discrete and continuous, to correct for endogeneity using the MIS method in this context. To the best of our knowledge, no research has reported indicators to correct for endogeneity in a time-of-day choice modelling context.

The estimated models were of Multinomial Logit (MNL) type with linear utilities. As future research we propose to correct for endogeneity using a more complex specification such as a Mixed Logit model with an error component to treat the pseudo panel nature

of the SP data. The estimated models were the following:

1) Benchmark model: Containing all the explanatory variables considered in the SP experiment, with representative utility given by (10):

$$V_{in} = ASC_i + \beta_t t_{in} + \beta_c c_{in} + \beta_{sd} sd_{in} + \beta_v v_{in} \tag{10}$$

Note that the parameter β_{sd} is specified as either β_{sde} or β_{sdl} for the early or late itineraries, respectively. The current itinerary is taken as reference (*ASC* fixed to zero) and does not have schedule delay. This model does not suffer from endogeneity since it considers all the attributes faced by the respondent and is therefore used as benchmark.

2) Endogenous model: As above but excluding sd_{in} and v_{in} in (10).
3) MIS-1 model[1]: This is a model corrected for endogeneity and estimated including the variable $\widehat{\delta}_{in}$ (coming from the first stage as explained above) and an indicator (I^1_{in}) within the utility function that can be discrete or continuous. The first and second stages of the MIS method follow expressions (11) and (12), respectively. Note that, as explained above, indicator 2 is used as an instrument for indicator 1 .

$$I^1_{in} = \alpha_1 + \alpha_t t_{in} + \alpha_c c_{in} + \alpha_{I^2_{in}} I^2_{in} + \delta_{in} \xrightarrow{OLS} \widehat{\delta}_{in} = I^1_{in} - \widehat{I}^1_{in} \tag{11}$$

$$V_{in} = \widehat{ASC}_i + \widehat{\beta}_t t_{in} + \widehat{\beta}_c c_{in} + \widehat{\beta}_{I^1_{in}} I^1_{in} + \widehat{\beta}_{\widehat{\delta}} \widehat{\delta}_{in} \tag{12}$$

4) MIS-2 model: Analogous to MIS-1, but in this case indicator 1 was used as an instrument of indicator 2 .

$$I^2_{in} = \alpha_1 + \alpha_t t_{in} + \alpha_c c_{in} + \alpha_{I^1_{in}} I^1_{in} + \delta_{in} \xrightarrow{OLS} \widehat{\delta}_{in} = I^2_{in} - \widehat{I}^2_{in} \tag{13}$$

$$V_{in} = \widehat{ASC}_i + \widehat{\beta}_t t_{in} + \widehat{\beta}_c c_{in} + \widehat{\beta}_{I^2_{in}} I^2_{in} + \widehat{\beta}_{\widehat{\delta}} \widehat{\delta}_{in} \tag{14}$$

Table 1 presents the loglikelihood $l(\theta)$ and subjective values of time, SVT $= \left(\frac{\widehat{\beta}_t}{\widehat{\beta}_c}\right)$, for the various models estimated. We show the results for SVT[2] because the correction of endogeneity implies a change of scale (Guevara and Ben-Akiva, 2012); therefore, we can only analyze the ratios of the estimated coefficients to assess the phenomena under study. Besides, SVT is an important measure in the planning and social evaluation of transport projects (Ortúzar and Willumsen, 2011).

Further, $l(\theta)$ and SVT for the MIS-1 and MIS-2 models correspond to the mean for 100 simulations. This is because – as mentioned above - to deal with the missing indicators we used a multiple imputation process formulated by Gopalakrishnan et al. (2020), and this process involved simulating the possible imputation values. These experiments were repeated with 100 random samples to avoid sampling bias.

Both models (benchmark and MIS) have the same number of parameters. The only difference between them is that the benchmark model uses the explanatory variables: ASC_i, t_{in}, c_{in}, sd_{in}, and v_{in} (equation (10)), whereas the MIS model uses I^1_{in} and $\widehat{\delta}_{in}$ instead of sd_{in}

Table 1

Summary of the $l(\theta)$ and SVT $\left(\frac{\widehat{\beta}_t}{\widehat{\beta}_c}\right)$ from the estimated models.

Databank	Sample Size	Model	$l(\boldsymbol{\theta})$	SVT (COP/min)
Discrete	437	Endogenous	−341.18	472.27
		Benchmark	−331.74	136.55
		MIS-1	−313.16	86.66
		MIS-2	−313.16	224.64
Continuous	476	Endogenous	−362.82	351.39
		Benchmark	−346.98	140.26
		MIS-1	−333.27	116.81
		MIS-2	−333.27	138.41

[1] The MIS-1 model includes both time and cost (t_{in} and c_{in}, respectively) as explanatory variables in the first-stage regression because the MIS approach must apply an ordinary least squares (OLS) regression to the first indicator on the second indicator and every explanatory variable (t_{in} and c_{in}) including the endogenous variable (t_{in}) and obtain the residuals $\widehat{\delta}_{in}$ (Wooldridge, 2010; Guevara and Polanco, 2016). This process is analogous for MIS-2 but the OLS regression to the second indicator on the first indicator and every explanatory variable (t_{in} and c_{in}), including the endogenous variable (t_{in}) must be applied.

[2] SVT is estimated in Colombian Pesos (COP)/minute. 1 US dollar is 3800 COP, approximately.

and v_{in} (equation (12)). A potentially counterintuitive result in Table 1 is that the benchmark model has a worse fit to the data than the MIS model. We think this finding may be due to the fact that the benchmark model is not capturing the whole effect due to the schedule differences. On the other hand, the indicators collected in the SP survey are apparently capturing relevant variance that the omitted variables used in the benchmark model cannot capture. To build the indicators, respondents were asked to grade the whole itinerary, i.e., not just the explanatory variables t_{in}, sd_{in}, and v_{in}. As modellers, we used those three variables linearly to represent the itinerary in the benchmark model. However, the respondent may be considering other aspects that we did not consider. When asking about the indicators, we are perhaps capturing a more significant part of the effect, and this could explain the better fit.

The SVT estimated from the Benchmark models are similar to the values reported in the literature (Deloitte Consulting SLU, 2017). We checked that the sign (negative) of the estimators for t_{in}, c_{in}, v_{in}, and sd_{in} for the Benchmark models were as expected according to microeconomic theory. They also fulfilled the condition $\widehat{\beta}_{sdl} < \widehat{\beta}_{sde} < 0$ as stated in the literature (Arellana et al., 2012; Koster and Verhoef, 2012; Thorhauge et al., 2016). This means that people care more about being late instead of early. Note that the SVT achieved from the Benchmark models with both databanks are similar (136.55 and 140.26 COP/min).

The results in Table 1 show that the Endogenous models have clearly the worst performance. They show bias up to 246% in the estimates of the SVT in comparison with the Benchmark models. Given that the Endogenous models are a restricted version of the Benchmark models (10), it is possible to apply the likelihood ratio (LR) test (Ortúzar and Willumsen 2011, page 281) to compare them.

LR is asymptotically distributed χ^2_r with r degrees of freedom, where r is the number of linear restrictions required to transform the more general model into the restricted version. In our case, $r = 2$ (because the restrictions are that both β_{sde} and β_{sdl} are zero), therefore, the LR test for both databanks are $LR = -2(-341.18 + 331.74) = 18.88$ and $LR = -2(-362.82 + 346.98) = 31.68$. These values must be compared with the critical value for two degrees of freedom at the 95% confidence level ($\chi^2_2 = 5.99$). As LR $> \chi^2_2$ (in both cases), the null hypothesis of model equality is confidently rejected, and we can conclude that the Benchmark models are indeed superior.

On the other hand, the models corrected using the MIS method cannot be compared with the Endogenous or Benchmark model in terms of the LR test, because they are not a restricted version of them. Notwithstanding, the results in Table 1 show that the SVT reached from models MIS-1 and MIS-2 (116.81 COP/min and 138.41 COP/min, respectively) for the case with continuous indicators are closer to that of the Benchmark model. In this case, the bias attained from these estimates is 16.7% and 1.3%, respectively.

On the contrary, the SVT estimated from the databank with discrete indicators for MIS-1 and MIS-2 (86.66 and 224.64 COP/min, respectively) imply bias of 36.5% and 64.5% (respectively) in comparison with the SVT from the Benchmark model (136.55 COP/min). Note, also, that the difference between the SVT estimated from the databank with continuous indicators is only 21.6 COP/min (138.41–116.81), while the difference in the case of the discrete indicators is unusually large: 138.2 COP/min). Therefore, and in line with our initial hypothesis about endogeneity correction with the MIS method, using continuous indicators indeed performs better than using discrete indicators.

We also checked the parameter estimates and the MIS-1 and MIS-2 models' performance for the 100 replications needed to apply the multiple imputation method. First, the sign of the estimators for t_{in} and c_{in} was always as expected (negative) according to microeconomic theory. Second, when the continuous version of I^1_{in} (*How pleasant is this itinerary for you?*) was included as an explanatory variable within the utility function (MIS-1 model), it always obtained a positive sign. This means that respondents tended to prefer itineraries that they graded better. Besides, the coefficient of $\widehat{\delta}_{in}$ was significantly different from zero for all imputations, implying that the model including the continuous version of I^1_{in} effectively suffered from endogeneity (Rivers and Vuong 1988).

On the other hand, when the continuous version of I^2_{in} (*How convenient is this departure time for you?*) was included as an explanatory variable within the utility function (MIS-2 model), it showed the same performance, in terms of signs, as I^1_{in}. As for this case the coefficient of $\widehat{\delta}_{in}$ was also significantly different from zero for all imputations, the presence of endogeneity is also evident in model MIS-2 (Rivers and Vuong 1988). Finally, the results in Table 1 show that the continuous versions of I^1_{in} and I^2_{in} have the same performance in terms of $l(\theta)$; however, the SVT are closer to the Benchmark model when I^1_{in} was used as an instrument of I^2_{in}. This may be due to the fact that the question stated in I^2_{in} captures better the effect of the additional travel time and schedule delay in our SP experiment than that in I^1_{in}.

4. A Monte Carlo experiment to assess the performance of discrete indicators using the MIS method

Our Monte Carlo experiment focused on three aspects: distribution of the indicator, sample size, and discretization process. This last aspect was considered relevant because, as we explain below, we assumed that the discrete indicators were obtained from continuous indicators; therefore, the discretization process should be explicit.

Our simulation experiment tried to emulate the same choice tasks of the departure time SP survey described in Section 3. For this, we generated data and estimated a DCM, following the scheduling model formulated by Small (1982), with three departure time choice alternatives: the current itinerary, an early itinerary and a late itinerary. The experiment was repeated 200[3] times to build a sampling distribution of the estimators (which becomes complicated in two stages methods) and to properly consider circumstantial effects.

We considered a DCM represented by the following utility function for alternative i and individual n:

[3] We are grateful to an anonymous reviewer for making us look deeper into this finding, using 200 repetitions, because we found significant differences compared to the results achieved for 100 repetitions; however no further differences appeared when using 300 repetitions.

$$U_{in} = ASC_i + \beta_t t_{in} + \beta_c c_{in} + \beta_{sd} sd_{in} + \beta_v v_{in} + e_{in} \quad (15)$$

where, c_{in}, t_{in}, v_{in} and sd_{in} emulate the same explanatory variables of the model used in the SP experiment explained in the previous section, e_{in} is an error term that distributes Extreme Value Type I, and ASC_i is the alternative specific constant of alternative i. The parameter β_{sd} in (15) can be specified as either β_{sde} or β_{sdl} for the early or late itineraries, respectively. Again, the current itinerary does not have a schedule delay. The explanatory variables c_{in}, t_{in}, v_{in} and sd_{in} were generated using three distributions: Normal, Uniform and Exponential.

For simulation purposes, and without loss of generality, we assumed that the values of the parameters in (15) were as follows: $\beta_t = -4$, $\beta_c = -2$, and $\beta_v = -1$. The *ASC* for the early and late itineraries had values of −0.5 and −1.0, respectively. Finally, following the findings of Thorhauge et al. (2016), we defined the schedule delay early parameter ($\beta_{sde} = -1$) with a smaller absolute value than the schedule delay late parameter ($\beta_{sdl} = -3$).

To replicate the correlation among the SP survey databank variables (c_{in}, t_{in}, v_{in} and sd_{in}), in the simulated experiment we used the copulas approach[4] and the covariance matrix from SP survey databank to simulate correlated random variables. On the other hand, two appropriate indicators were generated for each itinerary, using equations (16) and (17) below. As these indicators are computed from the variables sd_{in}, v_{in} and the error term ε_{in}, they have a continuous nature and will be labelled as *CI*, where the superscripts 1 and 2 identify the indicator. To ensure consistency in the distribution of *CI* in (16) and (17), we forced sd_{in}, v_{in} and ε_{in} to have the same distribution. For example, if sd_{in}, v_{in} and ε_{in} distribute Normal, then *CI* distributes Normal.

$$CI_i^1 = \alpha_i^1 + \alpha_{sd}^1 sd_{in} + \alpha_v^1 v_{in} + \varepsilon_{in}^1 \quad (16)$$

$$CI_i^2 = \alpha_i^2 + \alpha_{sd}^2 sd_{in} + \alpha_v^2 v_{in} + \varepsilon_{in}^2 \quad (17)$$

Here, α_i are intercepts of the linear equations; they take the value of 1 for the early and current itineraries, and −1 for the late itinerary. The selection of these values is consistent with the fact that the early itinerary tends to be graded better than the late itinerary. Finally, the remaining parameters in (16) and (17) had the same value: $\alpha_{sd} = \alpha_v = -2$. This was done because in the first stage regression of the MIS method with the real dataset, these parameters turned out to be similar.

On the basis of the continuous indicators above, discrete indicators (DI_i^1 and DI_i^2) for itinerary i were generated using two discretization criteria and four discretization algorithms; these were implemented trying to avoid changing the original data distribution (Gonzalez-Abril et al., 2009). The idea was to keep the high interdependence between the discrete indicators (DI_i^1 and DI_i^2) and the attributes sd_{in}, v_{in} and ε_{in}. The six discretization methods are described as follow:

1) The discrete indicator *DI* is equal to the continuous indicator's value rounded to the nearest integer.
2) The continuous indicator is assigned to a quintile interval (0–20, 20–40, …) and the closest integer belonging to the mean percentile of each interval (10, 30, …) is used as the discrete indicator *DI*.
3) The discrete indicator *DI* is estimated using the *minimum description length principle* (MDLP) algorithm (Fayyad and Irani, 1993).
4) The discrete indicator *DI* is estimated using the *class-attribute interdependence maximization* (CAIM) algorithm (Kurgan and Cios, 2004).
5) The discrete indicator *DI* is estimated using the *class-attribute contingency coefficient* (CACC) algorithm (Tsai et al., 2008), and
6) The discrete indicator *DI* is estimated using the *Ameva* algorithm (Gonzalez-Abril et al., 2009).

There were two objectives behind criteria (1) and (2). The first was to consider something that we observed during the data collection process: although we required respondents to grade the indicators on a continuous scale (i.e., using decimals), they tended to grade them without decimals. Therefore, we considered that a similar approach to this grading process was the criteria listed in (1). The second reason was based on the measure proposed by Guevara (2015), where the discrete indicators were constructed from the quantiles of an empirical distribution obtained from a Monte Carlo simulation. In our case, we proposed to obtain the continuous indicator as the closest integer belonging to the mean percentile of each quintile interval (0–20, 20–40, …) in our Monte Carlo simulation. Detailed descriptions of algorithms (3) to (6) are given in the Appendix. Discretization algorithms have played an essential role in data mining and computer science (Tsai et al., 2008), as these areas often involve using continuous attributes that need to be implemented using categorical or discrete versions. So, many discretization algorithms have been proposed such as MDLP, CAIM, CACC and Ameva. We used these because they have been proved appropriate in the computer science field. Besides, the discretization process using these algorithms guarantees two important facts. The first is to maximize the interdependence between the class variables (in our practical case *sd* and *v*) and the discretized indicator. The second is to minimize the number of discrete intervals (Kurgan and Cios, 2004).

MDLP, CAIM, CACC and Ameva have a common characteristic: they are *supervised* algorithms (Fayyad and Irani, 1993; Kurgan and Cios, 2004; Tsai et al., 2008; Gonzalez-Abril et al., 2009). This means that they discretize attributes by considering the interdependence between class labels and the attribute values, without needing to specify in advance some parameters of discretization such as the

[4] The copulas are multivariate functions that join or "*couple*" two or more univariate distribution functions to construct continuous multivariate distribution functions. The copula represents a convenient parametric way to model the dependency structure on joint distributions of random variables, particularly for a set of random variables (Nelsen, 1999).

maximum number of intervals or the significance level for a χ^2 test.[5] This is an advantage in attempting to avoid a wide margin for error (e.g., experts' absence to define parameters, measurement errors, Gonzalez-Abril et al., 2009). Many empirical studies in the scientific literature specialized in this topic show that supervised discretization methods are more effective than unsupervised discretization methods to maintain high predictive accuracy when training and validation accuracy are determined (Senavirathne and Torra, 2019).

The main difference between the algorithms used is that, whereas CAIM, CACC, and Ameva have their own discretization criterion[6] (*CAIM criterion*, *cacc value*, and *Ameva coefficient*, respectively), the MDLP algorithm bases its search for discrete intervals on the decision tree technique.[7] The search for discrete intervals based on discretization criteria or decision tree techniques allows finding the width of the discrete intervals given the range of values of a continuous attribute. In this way, each algorithm automatically selects a number of discrete intervals and, at the same time, finds the width of every interval based on the interdependency between class and attribute values.

Simulations were carried out for various sample sizes and for three distributions of the indicators, which depended on the distribution of the variables and error terms in (16) and (17). The aim was to analyze the effect of both conditions when correcting for endogeneity using the MIS method. Regarding sample size, we simulated three different situations: 5000, 2000 and 500 individuals. Finally, the variation of the distribution of the indicators was built by giving the variables sd_{in}, v_{in}, and the error term ε_{in} Normal, Uniform and Exponential distributions. In this way we could guarantee the following cases:

1) If sd_{in}, v_{in} and ε_{in} distribute Normal, the indicator distributes Normal.
2) If sd_{in}, v_{in} and ε_{in} distribute Uniform, the indicator distributes Triangular (if two terms are added) or Irving-Hall (if three terms are added).
3) If sd_{in}, v_{in} and ε_{in} distribute Exponential, the indicator distributes Gamma.

The observed variables in all cases were generated by means of the copulas approach (Nelsen, 1999), using the respective marginal distributions and considering a degree of correlation among them equal to the one observed in the SP experiment.

Finally, using the above data, choices were simulated and the databank from the Monte Carlo experiment was built. With this data, we estimated the following four models:

1) A *true* model, that considered all the explanatory variables described in (15). This model was used as benchmark.
2) An *endogenous* model estimated excluding the variables sd_{in} and v_{in} in (15) to cause endogeneity.
3) A *MIS DI* model that corrects the *endogenous* model including the variable $\widehat{\delta}_{in}$ (coming from the first stage) and the discrete indicator DI_{in} within the utility function. The utility function of this model is shown in (18). Again, the indicator DI^2_{in} was used as an instrument of DI^1_{in}.

$$U_{in} = \widehat{ASC}_i + \widehat{\beta}_t t_{in} + \widehat{\beta}_c c_{in} + \widehat{\beta}_{DI^1_{in}} DI^1_{in} + \widehat{\beta}_{\widehat{\delta}} \widehat{\delta}_{in} + \widetilde{\varepsilon}_{in} \tag{18}$$

4) A *MIS CI* model corrected for endogeneity, estimated including $\widehat{\delta}_{in}$ (coming from the first stage) and the continuous indicator CI_{in} within the utility function. The utility function of this model is shown in (19). This time, CI^2_{in} was used as an instrument of CI^1_{in}.

$$U_{in} = \widehat{ASC}_i + \widehat{\beta}_t t_{in} + \widehat{\beta}_c c_{in} + \widehat{\beta}_{CI^1_{in}} CI^1_{in} + \widehat{\beta}_{\widehat{\delta}} \widehat{\delta}_{in} + \widetilde{\varepsilon}_{in} \tag{19}$$

The four models described above were estimated for the different sample sizes, distribution of the indicators and discretization processes. As correcting for endogeneity in DCM implies a change of scale in the estimators (Guevara and Ben-Akiva, 2012), again we checked parameter ratios (SVT $= \frac{\widehat{\beta}_t}{\widehat{\beta}_c}$) in the comparisons.

Table 2 shows the Monte Carlo experiment results for the various sample sizes, when the indicators distribute Normal, and we use the first discretization criterium. All experiment runs were generated using the open-source software R (R Development Core Team, 2008). The ratios among parameters ($\frac{\widehat{\beta}_t}{\widehat{\beta}_c}$) correspond to the mean of 200 replications. The p-value is for the null hypothesis (H_0) that the parameter ratio is equal to its population value. When H_0 cannot be rejected, the method is said to have properly corrected for endogeneity, and when H_0 is rejected, the method is said to have failed. For testing purposes, we took as critical a p-value of 0.05.

[5] χ^2 is a statistical measure that conducts a significance test on the relationship between the values of a feature and the class. Kerber (1992) argues that in an accurate discretization, the relative class frequencies should be fairly consistent within an interval (otherwise the interval should be split to express this difference) but two adjacent intervals should not have similar relative class frequency (in that case the adjacent intervals should be merged into one). The χ^2 statistic determines the similarity of adjacent intervals based on some significance level. It tests the hypothesis that two adjacent intervals of a feature are independent of the class. If they are independent, they should be merged; otherwise, they should remain separate (Liu et al., 2002).

[6] For more details regarding these criteria see the Appendix Section.

[7] Decision tree learning is a supervised machine learning technique for inducing a decision tree from training data. A decision tree (also referred to as a classification tree or a reduction tree) is a predictive model which is a mapping from observations about an item to conclusions about its target value (Tan, 2015).

Table 2
Monte Carlo statistics for Normal indicators and discretization criterion 1 (rounded to the nearest integer).

Sample size	Model	$\frac{\widehat{\beta}_t}{\widehat{\beta}_c}$	Actual $\frac{\beta_t}{\beta_c}$	% Bias $\frac{\widehat{\beta}_t}{\widehat{\beta}_c}$	p-value
5000	True	2.007	2.0	0.4	0.11
	Endogenous	1.865	2.0	6.8	<0.01
	MIS CI	2.008	2.0	0.4	0.14
	MIS DI	1.993	2.0	0.4	0.22
2000	True	2.001	2.0	0.0	0.39
	Endogenous	1.867	2.0	6.7	<0.01
	MIS CI	1.998	2.0	0.1	0.39
	MIS DI	1.997	2.0	0.1	0.38
500	True	2.028	2.0	1.4	0.08
	Endogenous	1.880	2.0	6.0	<0.01
	MIS CI	2.042	2.0	2.1	0.09
	MIS DI	2.060	2.0	3.0	0.01

As can be seen, the true model always recovers the actual parameter ratio since its p-value is always far above the 5% threshold. On the contrary, the bias for the endogenous model is always significant. Under this simulation setting, where the indicators distribute Normal, and the discretization criterium is simply rounding to the nearest integer, correcting for endogeneity using the MIS method works with both discrete and continuous indicators. There is a case where the correction does not work (discrete indicators and 500 individuals). We think that this fact is due to an adverse effect yielded by small sample size.

Table 3 shows the p-values corresponding to the ratio $\frac{\widehat{\beta}_t}{\widehat{\beta}_c}$ for the MIS DI model, for the three sample sizes, when varying the discretization criteria and the distribution of the indicators. Although not shown in the table, for the sake of space, it was always possible to recover the ratio between parameters for the true model and for the MIS method using continuous indicators. On the other hand, the ratio among parameters for the endogenous model failed to meet the target in several cases.

The p-values in green highlight the cases where the ratio $\frac{\widehat{\beta}_t}{\widehat{\beta}_c}$ for the MIS DI model works. On the contrary, the values in red show the

Table 3
P-values corresponding to $\frac{\widehat{\beta}_t}{\widehat{\beta}_c}$ for the MIS DI model.

Discretization criterion	Indicator distribution	Sample size		
		5000	2000	500
1. Rounded to nearest integer	Normal	0.22	0.38	0.01
	Triangular	0.35	0.39	0.02
	Gamma	0.40	0.26	0.01
2. Percentiles	Normal	0.38	0.32	0.03
	Triangular	0.39	0.35	0.03
	Gamma	0.10	0.40	0.02
3. MDLP	Normal	<0.01	<0.01	<0.01
	Triangular	<0.01	<0.01	<0.01
	Gamma	<0.01	<0.01	0.03
4. CAIM	Normal	<0.01	<0.01	<0.01
	Triangular	<0.01	<0.01	0.12
	Gamma	<0.01	<0.01	0.38
5. CACC	Normal	<0.01	<0.01	0.21
	Triangular	<0.01	<0.01	0.20
	Gamma	<0.01	<0.01	0.15
6. Ameva	Normal	<0.01	<0.01	<0.01
	Triangular	<0.01	<0.01	0.19
	Gamma	<0.01	<0.01	<0.01

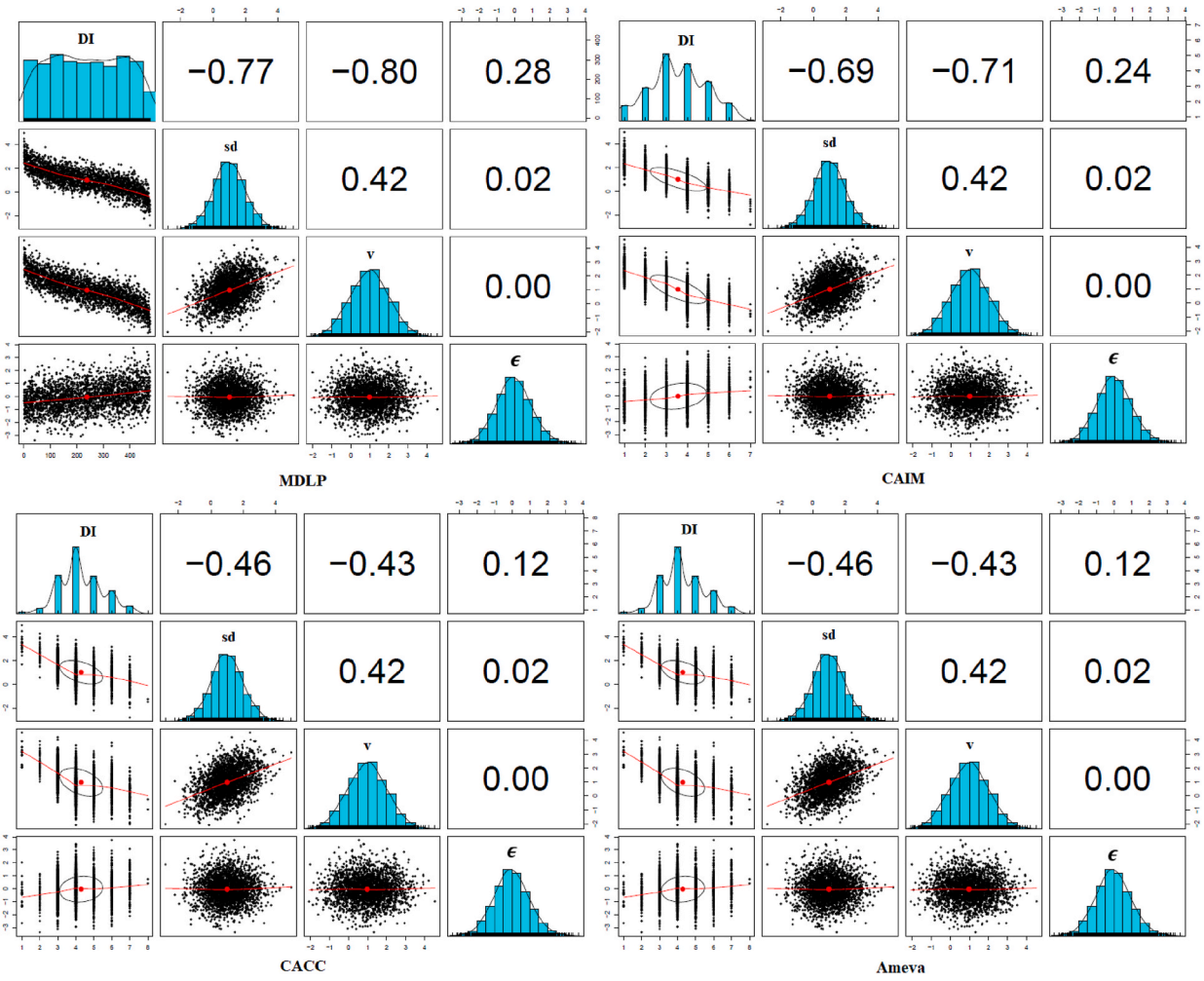


Fig. 2. Scatter plot for discrete indicator according to the algorithm.

cases where the correction failed under the 5% threshold. Finally, the values in black correspond to cases where the root-mean-squared error (RMSE) reached by the simulation was too large, yielding meaningless estimates of the p-value. This seems to be an effect of sample size because it only occurred for the smallest sample size (500 simulated individuals).

An interesting finding corresponds to the failure to correct for endogeneity when the MDLP, CACC, CAIM and Ameva algorithms are used. These algorithms are widely used in machine learning and data mining, among others (Tsai et al., 2008; Gonzalez-Abril et al., 2009). Although their use guarantees correlation between the generated discrete indicator and the class attribute (i.e., the variables sd_{in} and v_{in}), in addition to preserving the distribution of the discrete indicator, this does not guarantee endogeneity correction.

We noted that formally the MDLP algorithm does not preserve the continuous indicator distribution, but its discretization process involved discrete indicators with more granular values than other algorithms. We had hypothesised that this last fact could be due to the MDLP working better given that the estimation of the discrete indicators was made over a wide range of values. This would imply that, in practice, the approach tended to emulate a continuous distribution of values. However, our hypothesis was incorrect because, despite this fact, the correction was unsuccessful.

To illustrate the impact of the different discretizing algorithms in the empirical distribution of the indicator, Fig. 2 shows a scatter plot of matrices, with bivariate scatter plots below the diagonal, histograms on the diagonal, and the Pearson correlation index above the diagonal. The first column corresponds to the discrete indicator, the second and third represent the variables sd_{in} and v_{in}, respectively, and the last column represents the error term ε_{in} in (16) and (17). The results of MDLP, CAIM, CACC and Ameva are shown from left to right and top to bottom. The values above the main diagonal correspond to the correlation between the variables that are part of the scatter plot of matrices.

Fig. 2 shows that, as designed by the copula and discretization algorithm, the correlation between variables and indicators is different. For example, when the MDLP algorithm is used, the correlation between the indicator and sd_{in}, v_{in}, and ε_{in} achieved the highest values (−0.77, −0.80 and 0.28), in comparison with the other algorithms. Further, although all methods discretized the continuous indicator, the MDLP results appear substantially more granular (taking values between 0 and 500), while the other

methods take values between 1 and 8. On the other hand, the MDLP results appear almost uniform, completely losing the continuous indicator's bell distributional shape, while the other methods do a better job at retaining it. Considering these findings, the granularity in the discrete indicators coming from the MDLP algorithm does not improve the correction. Besides, the correlation between the indicator and the class variables seems essential during the discretization process. Although these results are valuable, we consider that more research is needed to reach a firm conclusion in this sense.

5. Conclusions and future research directions

Endogeneity is an anomaly that may yield inconsistent model parameters. We designed a Monte Carlo experiment and applied a SP survey, in the context of departure time choice, to examine the impact of using discrete indicators to correct for endogeneity with the MIS method in DCM. The reason was that, in practice, indicators tend to be discrete, but they should be continuous to be consistent with the mathematical derivation of the MIS method (Wooldridge, 2010).

From the Monte Carlo experiments, it was possible to show, first, that small sample sizes can lead to erroneous conclusions. As the mean square error increased, it erroneously led to conclude that the correction with discrete indicators worked. Second, the most straightforward criteria to produce discrete indicators, that is, *rounding the continuous value to the nearest integer* and *assign it to percentiles* (described in Section 4), worked properly. This may be due to the fact that these criteria preserve the correlation between the indicator and the class variables (sd_{in} and v_{in}). Similarly, we used four algorithms reported in the literature to discretize continuous indicators. We used these algorithms because they are typically used in the computer science field. Our findings show that the algorithm used to produce discrete indicators (from continuous ones) affects the endogeneity correction. Third, none of the complex algorithms tested to perform the discretization process worked successfully. Despite the MDLP algorithm achieving substantially more granular discrete indicators (similar to a continuous distribution) than those of the other three methods, the correction failed. However, more research on this is needed to reach a firm conclusion.

On the other hand, using real data we were able to show that the correction with continuous indicators worked better than the correction with discrete indicators. In particular, we checked the parameter ratios (to bypass the scaling problem), finding that the subjective value of time (SVT) for the model corrected with continuous indicators was closer to that of the Benchmark model than those computed from the model corrected with discrete indicators. Notwithstanding, the SVT achieved with the endogenous model was much worse, having a bias of approximately 250% with respect to the benchmark model.

An interesting finding from this part of the research were the two indicators themselves, as they were able to capture the effect of the explanatory variables related to the schedule delay (sd) and any additional travel time (v). This is a methodological contribution, as to the best of our knowledge, no previous research has suggested appropriate indicators to correct endogeneity under this modelling context.

Our findings also allow us to recommend that in future data collection, of this type, we should ask for indicators on a continuous scale, leaving aside the commonly used Likert scales. Besides, if discrete grades are used, respondents should be required to use wider ranges to achieve ratings throughout the entire scale. Our conclusions are valid for the conditions shown in our SP survey and the Monte Carlo experiment.

Finally, the findings and analyses performed in this paper have several limitations that future research should address. For example, the impact of different types of discrete choice models and functional forms should be considered, as we only modelled the simplest case of a Logit model with fixed parameters. We also recommend exploring a mathematical derivation contemplating discrete distributions for the indicators. Finally, we think an interesting extension would be to consider that the discrete indicators arise from an ordered logit/probit in the data generating process of the Monte Carlo experiment.

CRediT author contribution statement

Thomas E. Guerrero: Study conception, design, analysis and interpretation of results, draft manuscript preparation. **C. Angelo Guevara:** Study conception, analysis and interpretation of results, draft manuscript preparation. **Elisabetta Cherchi and Juan de Dios Ortúzar:** Analysis and interpretation of results, draft manuscript preparation.

Declaration of competing interest

The authors declare that they have no conflict of interest.

Acknowledgments

This research was partially funded by Agencia Nacional de Investigación y Desarrollo (ANID), FONDECYT 1191104 and by the Instituto Sistemas Complejos de Ingeniería (ISCI), through the grant ANID PIA/BASAL AFB180003. We are also grateful for the support received from the BRT+ Centre of Excellence (www.brt.cl), financed by the Volvo Research and Educational Foundations.

Appendix

A detailed description of the four algorithms used in this research is reported below. The pseudocodes, description and notation used were extracted directly from the papers proposing the algorithms.

Class-attribute interdependence maximization (CAIM) algorithm

The CAIM algorithm was developed by Kurgan and Cios (2004). The CAIM criterion measures the dependency between a class variable C and a discretization variable D for attribute F, for a given quanta matrix[8] (see Table 4), and is defined as:

$$CAIM(C,D|F)=\frac{\sum_{r=1}^{n}\frac{max_r^2}{M_{+r}}}{n} \tag{20}$$

Table 4
2D quanta matrix for attribute F and discretization scheme D

Class	Interval					Class Total
	$[d_0,d_1]$	…	$(d_{r-1},d_r]$	…	$(d_{n-1},d_n]$	
C_1	q_{11}	…	q_{1r}	…	q_{1n}	M_{1+}
:	:	…	:	…	:	:
C_i	q_{i1}	…	q_{ir}	…	q_{in}	M_{i+}
:	:	…	:	…	:	:
C_s	q_{S1}	…	q_{Sr}	…	q_{Sn}	M_{S+}
Interval Total	M_{+1}	…	M_{+r}	…	M_{+n}	

where n is the number of intervals, r iterates through all intervals (i.e., $r=1,2,\ldots,n$), $\max_r$ is the maximum value among all q_{ir} values (maximum value within the rth column of the quanta matrix), $i=1,2,\ldots,S$, M_{+r} is the total number of continuous values of attribute F that are within the interval $(d_{r-1},d_r]$.

Kurgan and Cios (2004) describe the pseudocode of the CAIM algorithm. Given data consisting of M examples, S classes, and continuous attributes F_i. For every F_i do the following steps:

Step 1.
1.1 Find maximum (d_n) and minimum (d_0) values of F_i.
1.2 Form a set of all distinct values of F_i in ascending order, and initialize all possible interval boundaries B with minimum, maximum and all the midpoints of all the adjacent pairs in the set.
1.3 Set the initial discretization scheme as $D:\{[d_0,d_n]\}$ set GlobalCAIM = 0.
Step 2.
2.1 Initialize $k=1$.
2.2 Tentatively add an inner boundary, which is not already in D, from B, and calculate the corresponding CAIM value.
2.3 After all the tentative additions have been tried, accept the one with the highest value of CAIM.
2.4 If (CAIM > GlobalCAIM or k < S) then update D with the accepted in Step 2.3 boundary and set GlobalCAIM = CAIM, else terminate.
2.5 Set $k=k+1$ and go to 2.2.

Output: Discretization scheme D.

Class-attribute contingency coefficient (CACC) algorithm

Tsai et al. (2008) developed this algorithm. Given a dataset with i continuous attributes, M examples, and S target classes, for each attribute A_i, CACC first finds the maximum d_n and minimum d_0 of A_i and then forms a set of all distinct values of A_i in the ascending order. As a result, all possible interval boundaries B with the minimum and the maximum, and all the midpoints of all the adjacent boundaries in the set are obtained.

Then, CACC would iteratively partition the attribute A_i. In the *kth* loop, CACC would compute for all possible cutting points to find the one with the maximum *cacc* value and then partition this attribute accordingly into $k+1$ intervals. The *cacc* value is defined as follows:

$$cacc=\sqrt{\frac{y'}{y'+M'}} \tag{21}$$

where $y'=M\left[\left(\sum_{i=1}^{s}\sum_{r=1}^{n}\frac{q_{ir}^2}{M_{i+}M_{+r}}\right)-1\right]\Big/\log(n)$. To reduce the computational cost of the discretization, CACC also uses a greedy method (as in CAIM) to generate the sub-optimal discretization scheme. In other words, for every loop, CACC not only finds the best division

[8] A quanta matrix is a two-dimensional frequency matrix where the class variable and the discretization variable of attribute F are treated as two random variables.

point but also records a *Globalcacc* value.

If the generated *cacc* value in loop $k+1$ is less than the *Globalcacc* obtained in loop k, CACC would terminate and output the discretization scheme. Besides, to generate a rational discrete result, such a greedy mechanism is ignored if the number of generated intervals is less than the number of target classes. Since the main framework of CACC is similar to that of CAIM, the complexity of CACC for discretizing a single attribute is still $O(m\log(m))$, where m is the number of distinct values of the discretized attribute.

Ameva algorithm

Gonzalez-Abril et al. (2009) developed this algorithm. Let $X=\{X_1,\ldots,X_N\}$ be a training data set of a continuous attribute X of mixed-mode data such that each example x_i belongs to only one of the l classes of the class variable denoted by $\mathscr{I}=\{C_1,\ldots,C_l\}$. A continuous attribute discretization is a function $D: X\rightarrow\mathscr{I}$ which assigns a class $C_i\in\mathscr{I}$ to each value $x\in X$.

Let us consider a discretization D which discretizes X into k discrete intervals $\{L_1,\ldots,L_k\}$ where $L_1=[(d_0,d_1]$ and $L_j=(d_{j-1},d_j]$ for $j=2,\ldots,k$ such that d_k is the maximal value and d_0 is the minimal value of attribute X, and the values d_i are arranged in ascending order. Thus, a discretization variable is defined as $\mathscr{L}(k;X,\mathscr{I})=\{L_1,\ldots,L_k\}$ which verifies that, for all $x_i\in X$, a unique L_j exists such that $x_i\in L_j$ for $i=1,\ldots,N$ and $j=1,\ldots,k$. The discretization variable $\mathscr{L}(k;X,\mathscr{I})$ (denoted as $\mathscr{L}(k)$ for the sake of simplicity) of attribute X and the class variable $\mathscr{I}$ are treated from a descriptive point of view and Table 5 is drawn up where n_{ij} denotes the total number of continuous values belonging to the ith class that are within the jth interval, n_i is the total number of instances belonging to the ith class, and n_j is the total number of instances belonging to the jth interval for $i=1,\ldots,l$ and $j=1,\ldots,k$.

Table 5
Contingency table for attribute X and discretization variable $\mathscr{L}(k)$

$C_i \mid L_j$	L_1	…	L_j	…	L_k	n_j
C_1	n_{11}	…	n_{1j}	…	n_{1k}	n_1
…	…		…		…	…
C_i	n_{i1}	…	n_{ij}	…	n_{ik}	n_i
…	…		…		…	…
C_l	n_{l1}	…	n_{lj}	…	n_{lk}	n_l
n_j	n_1		n_j		n_k	N

Given discrete attributes $\mathscr{I}$ and $\mathscr{L}(k)$, the contingency coefficient, denoted by $\chi^2(k)\overset{def}{=}\chi^2(\mathscr{L}(k),\mathscr{I}|\mathrm{X})$, defined as:

$$\chi^2(k)=N\left(-1+\sum_{i=1}^{l}\sum_{j=1}^{k}\frac{n_{ij}^2}{n_i n_j}\right) \tag{22}$$

is considered. It is straightforward to prove that:

$$\max_{X,\mathscr{L}(k),\mathscr{I}}\chi^2(k)=N(min\{l,k\}-1) \tag{23}$$

Hence, the Ameva coefficient, $Ameva(k)\overset{def}{=}Ameva(\mathscr{L}(k),\mathscr{I}|\mathrm{X})$ is defined as follows:

$$Ameva(k)=\frac{\chi^2(k)}{k(l-1)} \tag{24}$$

for $k,l\geq 2$. The Ameva criterion has the following properties:

- The minimum value of $Ameva(k)$ is 0 and when this value is achieved then both discrete attributes $\mathscr{I}$ and $\mathscr{L}(k)$ are statistically independent and vice versa.
- The maximum value of $Ameva(k)$ indicates the best correlation between the class labels and the discrete intervals. If $k\geq l$ then, for all $x\in C_i$ a unique j_0 exists such that x $x\in L_{j_0}$ (the remaining intervals $(k-l)$ have no elements); and if $k<l$ then, for all $x\in L_j$, a unique i_0 exists such that $x\in C_{i_0}$ (the remaining classes have no elements), that is, the highest value of the Ameva coefficient is achieved when all values within a particular interval belong to the same associated class for each interval.
- The aggregated value is divided by the number of intervals k, hence the criterion favours discretization schemes with the lowest number of intervals.
- To make a comparison with the CAIM coefficient, the aggregated value is divided by $l-1$.
- From (23), it follows that $Ameva_{max}(k)\overset{def}{=}max_{X,\mathscr{L}(k),\mathscr{I}}\,Ameva(k)=\frac{N(k-1)}{k(l-1)}$ if $k<l$ and $\frac{N}{k}$ otherwise. Hence, $Ameva_{max}(k)$ is an increasing function of k if $k\leq l$, and a decreasing function of k if $k>l$. Therefore, $max_{k\geq 2}Ameva_{max}(k)=Ameva_{max}(l)$, that is, the maximum of the Ameva coefficient is achieved in the optimal situation (all values of C_i are in a unique interval L_j and vice versa).

Minimum description length principle (MDLP) algorithm

Fayyad and Irani (1993) developed this algorithm. The MDLP criterion for the partition induced by a cut point T for a set S of N examples and for the continuous-valued attribute A is accepted if:

$$Gain(A,T;S) > \frac{\log_2(N-1)}{N} + \frac{\Delta(A,T;S)}{N} \tag{25}$$

and it is rejected otherwise. Note that the quantities required to evaluate this criterion, namely the information entropy of S, S_1 and S_2 are computed by the cut point selection algorithm as part of the cut point evaluation.

References

Arellana, J., Daly, A., Hess, S., Ortúzar, J. de D., Rizzi, L.I., 2012. Development of surveys for study of departure time choice: two-stage approach to efficient design. Transport. Res. Rec. 2303, 9–18.

Bahamonde-Birke, F.J., Kunert, U., Link, H., Ortúzar, J. de D., 2017. About attitudes and perceptions: finding the proper way to consider latent variables in discrete choice models. Transportation 44, 475–493.

Ben-Akiva, M.E., Walker, J.L., Bernardino, A.T., Gopinath, D.A., Morikawa, T., Polydoropoulou, A., 2002. Integration of choice and latent variable models. In: Mahmassani, En H.S. (Ed.), In Perpetual Motion: Travel Behaviour Research Opportunities and Challenges. Pergamon, Amsterdam.

Berry, S., Levinsohn, J., Pakes, A., 1995. Automobile prices in market equilibrium. Econometrica 63, 841–890.

Bolduc, D., Alvarez-Daziano, R., 2009. On estimation of hybrid choice models. In: Hess, En S., Daly, A. (Eds.), Choice Modelling: the State-Of-The-Art and the State-Of-Practice. Emerald Group Publishing Limited, Bingley.

Daly, A., Hess, S., Patruni, B., Potoglou, D., Rohr, C., 2012. Using ordered attitudinal indicators in a latent variable choice model: a study of the impact of security on rail travel behaviour. Transportation 39, 267–297.

Deloitte Consulting SLU, 2017. Evaluación Socioeconómica para la Primera Línea de metro de Bogotá. Deloitte Consulting SLU, Bogotá.

Ebbes, P., Papies, D., Van Heerde, H.J., 2011. The sense and non-sense of holdout sample validation in the presence of endogeneity. Market. Sci. 30, 1115–1122.

Fayyad, U., Irani, K.B., 1993. Multi-interval discretization of continuous valued attributes for classification learning. In: Proceedings of 13th International Joint Conference on Artificial Intelligence, pp. 1022–1027, 1993.

Fernández-Antolín, A., Guevara, C.A., De Lapparent, M., Bierlaire, M., 2016. Correcting for endogeneity due to omitted attitudes: empirical assessment of a modified MIS method using RP mode choice data. J. Choice Modelling 20, 1–15.

Glerum, A., Atasoy, B., Bierlaire, M., 2014. Using semi-open questions to integrate perceptions in choice models. J. Choice Modelling 10, 11–33.

Gonzalez-Abril, L., Cuberos, F.J., Velasco, F., Ortega, J.A., 2009. Ameva: an autonomous discretization algorithm. Expert Syst. Appl. 36, 5327–5332.

Gopalakrishnan, R., Guevara, C.A., Ben-Akiva, M., 2020. Combining multiple imputation and control function methods to deal with missing data and endogeneity in discrete-choice models. Transp. Res. Part B Methodol. 142, 45–57.

Guerrero, T.E., Guevara, C.A., Cherchi, E., Ortúzar, J. de D., 2021. Addressing endogeneity in strategic urban mode choice models. Transportation 48, 2081–2102.

Guevara, C.A., 2015. Critical assessment of five methods to correct for endogeneity in discrete-choice models. Transport. Res. Pol. Pract. 82, 240–254.

Guevara, C.A., 2018. Overidentification tests for the exogeneity of instruments in discrete choice models. Transp. Res. Part B Methodol. 114, 241–253.

Guevara, C.A., Ben-Akiva, M.E., 2006. Endogeneity in residential location choice models. Transport. Res. Rec. 60–66, 1977.

Guevara, C.A., Ben-Akiva, M.E., 2012. Change of scale and forecasting with the control-function method in logit models. Transport. Sci. 46, 425–437.

Guevara, C.A., Polanco, D., 2016. Correcting for endogeneity due to omitted attributes in discrete-choice models: the multiple indicator solution. Transportmetrica A: Transport. Sci. 12, 458–478.

Kerber, R., 1992. Chimerge: discretization of numeric attributes. In: Proc. AAAI-92, Ninth National Conference Artificial Intelligence. AAAI Press/The MIT Press, pp. 123–128.

Koster, P., Verhoef, E.T., 2012. A rank-dependent scheduling model. J. Transport Econ. Pol. 46, 123–138.

Kurgan, L.A., Cios, K.J., 2004. CAIM discretization algorithm. IEEE Trans. Knowl. Data Eng. 16, 145–153.

Likert, R., 1932. A technique for the measurement of attitudes. Arch. Psychol. 140, 1–55.

Liu, H., Hussain, F., Tan, C.L., Dash, M., 2002. Discretization: an enabling technique. Data Min. Knowl. Discov. 6 (4), 393–423.

Louviere, J.J., Train, K., Ben-Akiva, M., Bhat, C., Brownstone, D., Cameron, T.A., Waldman, D., 2005. Recent progress on endogeneity in choice modelling. Market. Lett. 16, 255–265.

Lurkin, V., Garrow, L.A., Higgins, M.J., Newman, J.P., Schyns, M., 2017. Accounting for price endogeneity in airline itinerary choice models: an application to Continental U.S. markets. Transport. Res. Pol. Pract. 100, 228–246.

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. Transport. Res. Pol. Pract. 66, 196–212.

Nelsen, R.B., 1999. An Introduction to Copulas. Springer, New York.

Ortúzar, J. de D., Willumsen, L.G., 2011. Modelling Transport, fourth ed. John Wiley & Sons, Chichester.

Park, S., Gupta, S., 2009. Simulated maximum likelihood estimator for the random coefficient logit model using aggregate data. J. Market. Res. 46, 531–542.

Petrin, A., Train, K., 2010. A control function approach to endogeneity in consumer choice models. J. Market. Res. 47, 3–13.

Raveau, S., Alvarez-Daziano, R., Yañez, M.F., Bolduc, D., Ortúzar, J. de D., 2010. Sequential and simultaneous estimation of hybrid discrete choice models: some new findings. Transport. Res. Rec. 2156, 131–139.

R Development Core Team, 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. http://www.Rproject.org.

Rivers, D., Vuong, Q., 1988. Limited information estimators and exogeneity tests for simultaneous probit models. J. Econom. 39, 347–366.

Rose, J.M., Bliemer, M.C.J., 2009. Constructing efficient stated choice experimental designs. Transport Rev. 29, 587–617.

Senavirathne, N., Torra, V., 2019. Rounding based continuous data discretization for statistical disclosure control. J. Ambient Intell. Hum. Comput. 1–19.

Small, K.A., 1982. The scheduling of consumer activities: work trips. Am. Econ. Rev. 72, 467–479.

Stock, J.H., Yogo, M., 2005. Testing for weak instruments in linear IV regression. In: Andrews, En D.W., Stock, J.H. (Eds.), Identification and Inference for Econometric Models, Essays in Honour of Thomas Rothenberg. Cambridge University Press, Cambridge.

Tan, L., 2015. Code comment analysis for improving software quality. In: The Art and Science of Analyzing Software Data. Morgan Kaufmann, pp. 493–517.

Thorhauge, M., Cherchi, E., Rich, J., 2016. How flexible is flexible? Accounting for the effect of rescheduling possibilities in choice of departure time for work trips. Transport. Res. Pol. Pract. 86, 177–193.

Tsai, C.J., Lee, C.I., Yang, W.P., 2008. A discretization algorithm based on class-attribute contingency coefficient. Inf. Sci. 178, 714–731.

Walker, J.L., Ben-Akiva, M.E., 2002. Generalized random utility model. Math. Soc. Sci. 43, 303–343.

Wen, C.H., Chen, P.H., 2017. Passenger booking timing for low-cost airlines: a continuous logit approach. J. Air Transport. Manag. 64, 91–99.

Wooldridge, J., 2010. Econometric Analysis of Cross-Section and Panel Data, second ed. The MIT Press, Cambridge, Mass.
Yáñez, M.F., Raveau, S., Ortúzar, J. de D., 2010. Inclusion of latent variables in mixed logit models: modelling and forecasting. Transport. Res. Pol. Pract. 44, 744–753.
Zwerina, K., Huber, J., Kuhfeld, W., 2005. A General Method for Constructing Efficient Choice Designs. SAS Institute, Ludwigshafen.