Information theoretic-based sampling of observations

van Cranenburgh, Sander; Bliemer, Michiel C.J.

DOI
10.1016/j.jocm.2018.02.003

Publication date
2018

Document Version
Final published version

Published in
Journal of Choice Modelling

Journal of Choice Modelling 31 (2019) 181–197

Contents lists available at ScienceDirect

Journal of Choice Modelling

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

Information theoretic-based sampling of observations

Sander van Cranenburgh [a,*], Michiel C.J. Bliemer [b]

[a] Delft University of Technology, Faculty of Technology, Policy & Management, Transport and Logistics Group, Jaffalaan 5, 2628 BX, Delft, The Netherlands

[b] The University of Sydney, Institute of Transport and Logistics Studies, Australia

A B S T R A C T

Due to the surge in the amount of data that are being collected, analysts are increasingly faced with very large data sets. Estimation of sophisticated discrete choice models (such as Mixed Logit models) based on these typically large data sets can be computationally burdensome, or even infeasible. Hitherto, analysts tried to overcome these computational burdens by reverting to less computationally demanding choice models or by taking advantage of the increase in computational resources. In this paper we take a different approach: we develop a new method called Sampling of Observations (SoO) which scales down the size of the choice data set, prior to the estimation. More specifically, based on information-theoretic principles this method extracts a subset of observations from the data which is much smaller in volume than the original data set, yet produces statistically nearly identical results. We show that this method can be used to estimate sophisticated discrete choice models based on data sets that were originally too large to conduct sophisticated choice analysis.

1. Introduction

In numerous fields, recent technological advances have led to a surge in the amount of data that are being collected. These emerging data sources are changing the data landscape as well as the methods by which data are analysed. For instance, in the field of transport mobile phone, GPS, WIFI, and public transport smartcard data (Iqbal et al., 2014; Jánošíková et al., 2014; Prato et al., 2014; Farooq et al., 2015) are nowadays complementing or fully replacing traditional travel survey methods (Rieser-Schüssler, 2012), and data-driven methods (such as e.g. machine learning), as opposed to theory-driven methods, are increasingly becoming part of the standard tool-box of transport analysts (Wong et al., 2017). Moreover, it is widely believed that the amount of data that are being collected will continue to increase rapidly in the decades to come (Witlox, 2015).

Although these emerging data sources are widely believed to assist in understanding and solving numerous societal problems, they pose all sorts of new challenges to analysts. For choice modellers one major challenge relates to the computational burden. In particular, the size of these new data renders estimation of sophisticated state-of-the-art discrete choice models, such as Mixed Logit models (Revelt and Train, 1998) computationally burdensome, or even technically infeasible (Vlahogianni et al., 2015). This, in turn, is limiting the use of these emerging data sources in numerous fields where choice models are used. Moreover, even if model estimation is technically feasible on the large data set, many different model specifications are often tested. Therefore, long estimation times (which for current data sets may already easily take several days) quickly become prohibitive.

To deal with increasingly large choice data sets two types of approaches are commonly taken by analysts. The first approach is to revert to less computationally demanding models, such as Multinomial Logit (McFadden, 1974) and Nested Logit (Daly and Zachery, 1978) models. However, despite being very effective in reducing the computational efforts, this approach severely limits the analyst's ability to adequately model complex types of choice behaviour. As such, this approach is far from desirable. The second commonly taken

* Corresponding author.
E-mail addresses: s.vancranenburgh@tudelft.nl (S. van Cranenburgh), michiel.bliemer@sydney.edu.au (M.C.J. Bliemer).

http://dx.doi.org/10.1016/j.jocm.2018.02.003
Received 9 June 2017; Received in revised form 2 February 2018; Accepted 20 February 2018
Available online 6 April 2018
1755-5345/

approach is to modify estimation code in order to increase computational power, e.g. by taking advantage of parallel computing or cluster computing facilities. Recent technological advances have made it easier to employ cloud computing facilities and high performance computation clusters. However, many analysts do not have access to such facilities and existing widely available estimation software is typically not ready for taking full advantage of these technologies.

To the best of the authors' knowledge, no efforts have been undertaken to scale down large choice data sets,[1] such that initially large data sets can be used with standard discrete choice estimation software packages, such as Biogeme, Nlogit, and Alogit. While removing valid observations is considered a sin by many analysts working with discrete choice models, in fields like machine-learning down-scaling of data sets is more common practice (e.g. Arnaiz-González et al., 2016; Loyola et al., 2016). As we will argue in this paper, using a carefully sampled subset of choice observations can give nearly identical estimation results as compared to using the complete dataset. Hence, we believe that this approach is worth exploring from a practical point of view.

This study proposes a new information theoretic-based method that lowers the computational burden to estimate sophisticated discrete choice models based on large data sets. The method – which we call Sampling of Observations (SoO) – is inspired by, and closely related to, efficient experimental design. It reduces the size of the data by combining practices from the field of experimental design in Stated Choice (SC) studies (see e.g. Rose and Bliemer, 2009), with established notions from information theory (Shannon and Weaver, 1949). SoO constructs a subset of the data that consists of a manageable number of observations that are *jointly* highly informative on the behaviour that is being studied. It does so by sampling observations from the full data set in such a way that the D-error statistic of the subset is minimised, which means that Fisher information is maximised. The D-error is evaluated using what we call the sampling model. This model is a 'simplified' version of the sophisticated choice model that the analyst ultimately wishes to estimate. By using a simple model in the sampling stage, SoO is computationally cheap and fast to conduct.

The remaining part of this paper is organised as follows. Section 2 presents the methodology on information theoretic-based sampling of observations. Section 3 explores the efficacy of the method using Monte Carlo analyses. Finally, Section 4 closes with conclusions and a discussion.

2. Methodology

2.1. Preliminary: the effect of sample size

For asymptotically consistent estimators the standard errors associated with the estimates decrease with increasing sample size. Specifically, in case the observations are randomly drawn from the target population, standard errors decrease at a rate of $1/\sqrt{N}$ (Fisher, 1925), where N denotes the sample size, see Fig. 1. This implies that the slope flattens out fast: at a rate of $1/N^{1.5}$. This reflects the fact that relatively speaking less and less *new* information is revealed on the data generating process to the model when more and more randomly drawn observations are included in the data set.

When a data set is too large to analyse using sophisticated discrete choice models, one way to down-scale it is by randomly sampling N^* observations from the large data set consisting of N observations (where N^* much smaller than N). The increase in the standard errors incurred by such a naïve random sampling method can be derived from Fig. 1. Specifically, the standard errors are $\sqrt{N/N^*}$ times larger as compared to when the full data set would have been used. For example, in case an analyst scales down a data set by a factor 50 using this method, the standard errors will become about 7 times larger as compared to when the full data set would have been used. This implies that the model parameter will be considerably less precisely recovered.

2.2. Information-theoretic based sampling of observations

In this paper we propose a more sophisticated way to scale-down choice data sets. This method – which we abbreviate as Sampling of Observations (SoO) – is based on information-theoretic principles. Colloquially speaking, when deciding on whether to sample a particular observation, or not, it takes into account the 'information value' of that observation. After all, not all observations are equally informative on the behaviour that is being studied. For instance, all else equal, it is more informative to observe 100 choice observations that contain trade-offs between attributes than 100 choice observations which all contain a strongly dominant alternative[2] (Bliemer et al., 2017); and it is more informative to observe 100 choice observations across choice sets having many alternatives, than to observe 100 choice observations across binary choice sets.

Fig. 2 outlines the SoO methodology. In essence, SoO aims to extract a subset of observations that allows for statistically efficient recovery of the model parameters. It involves taking five steps. Next, we describe each step in detail.

Steps:

1. Define and parameterise the sampling model

[1] A well-known method developed by McFadden (1978) to reduce the computational burden is Sampling of Alternatives (SoA). This method reduces the computational burden by taking out the need to compute the utilities across all alternatives. Therefore, SoA is effective in reducing the computational burden specifically in case the number of alternatives is large. However, as the computational burden of these new data is specifically caused by the large number of observations, this method does not help much to reduce computational issues with such data.

[2] An alternative A is said to dominate alternative B if alternative A performs at least as well as alternative B in terms of all attributes, and performs better in at least one attribute.

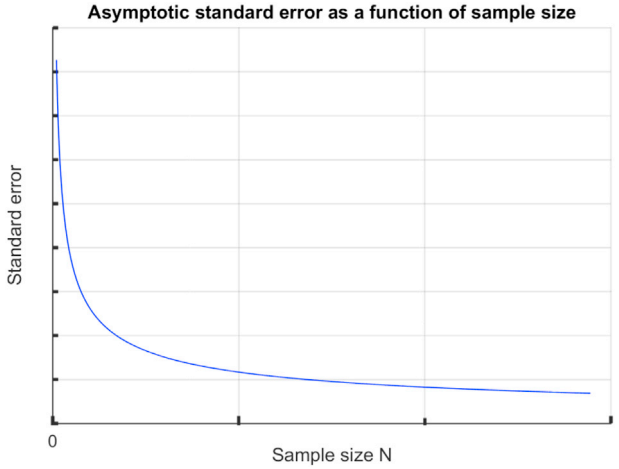

Fig. 1. Asymptotic standard error as a function of sample size.

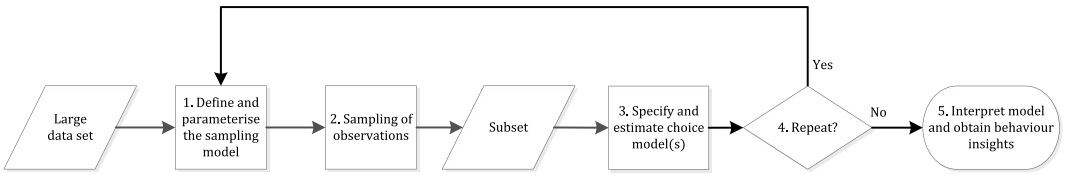

Fig. 2. Information-theoretic based sampling of observations.

The information value of a choice observation is evaluated using what we call the sampling model. This model resembles the sophisticated choice model that the analyst ultimately wishes to estimate, but is simpler. That is, the choice probabilities of the sampling model are fast to compute (hence, take a closed-form formula) and the second order derivatives of the Log-Likelihood function w.r.t. the models parameters can analytically be derived. These are needed to compute information measures, such as the D-error statistic, in the next step.

An important question is how to choose the sampling model? Naturally, the 'best' sampling model is the one that most closely resembles the sophisticated model that we analyst wishes to estimate. But, given the strong requirement that the choice probabilities of the sampling model need to be closed-form, the number of candidate sampling models is actually fairly limited, most notably the Multinomial Logit (MNL) and the Nested Logit (NL) model. Therefore, choosing the sampling model is typically fairly straightforward. For instance, in case the analyst wishes to estimate a (Panel) Mixed Logit model, the most obvious candidate model for the sampling model is an MNL model. In case the analyst wishes to estimate a choice model containing Error Components (ECs), the NL model seems most appropriate. For other models, such as e.g. Latent Class (LC) choice models or Integrated Choice and Latent Variable (ICLV) models there is no obvious candidate sampling model. In those cases it is probably best to go for an MNL sampling model. However, further research is needed to establish best practices for those types of models.

To parameterise the sampling model so-called prior parameters are needed. One way to obtain these parameter priors is by estimating the sampling model based on the large data set (or, based on a random subset in case the data set is too large to even estimate the sampling model). Since the sampling model has closed-form formula for the choice probabilities, estimation of the sampling model does not consume much time or computational resources. Alternatively, the analyst may turn to the literature to find appropriate priors to parametrise the sampling model or base them on their own expert judgement (Bliemer and Collins, 2016).

2. Sampling of observations

The analyst determines the size of the subset N^*. The size of the subset depends on many aspects, e.g. the application, the desired precision of the retrieved parameter estimates, the number of parameters of the sophisticated model, etc. Based on the authors' experiences, a couple of thousands of observations is usually sufficient, but this may be case specific. Next, the solution space is searched to find a subset of observations that allows for statistically efficient recovery of the model parameters, given the sampling model and the subset size.

3. Specify and estimate the sophisticated choice models based on the subset

Having obtained the subset, the next step is to estimate the desired sophisticated choice model. Typically, the analyst does not know

the exact model specification upfront. Rather, the model specification phase involves an exploratory process to find what model specification best describes the decision making process in the data. Therefore, usually multiple sophisticated choice models are tested. Since the subset is relatively small, this exploratory phase can be done relatively quickly.

4. Repeat

In case the sampling model is adequately parameterised (i.e. the prior parameters of the sampling are close to the model estimates obtained from the sophisticated model), there is no need to repeat steps. Research conducted in the context of SC experimental design suggests that mild parameters misspecifications do not lead to large reductions in statistical efficiency of the subset (Ferrini and Scarpa, 2007). However, it is advisable to repeat Steps 1 to 3 in the following two cases:

I. In case the final model specification of the sophisticated model deviate considerably from the model specification of the sampling model (including its parameters). Research conducted in the context of SC experimental design suggests that large prior parameter misspecifications may lead to substantial reductions in efficiency (Walker et al., 2017), undermining the objective of the sampling method. More precise parameter estimates are likely to be obtained after updating the priors in the sampling model, and repeating steps 1 to 3.

II. In case the standard errors (and associated t-values) are considered too large by the analyst. In that case, smaller standard error can be obtained by increasing the size of the subset and repeat steps 1 to 3.

5. Interpret model and obtain behaviour insights

Derive behavioural insights, or use the estimated model for forecasting, depending on the aim of the research. Notice that SoO is methodologically speaking akin to efficient experimental design (e.g. Huber and Zwerina, 1996; Kanninen, 2002; Rose and Bliemer, 2009, 2013b; de Bekker-Grob et al., 2015). SoO and efficient experimental design both aim to maximize the precision at which the model parameters are retrieved from the data (i.e. small standard errors). A key difference between these two methods however is that SoO is done *after* the data are collected, while experimental design takes place *prior* to the Stated Choice (SC) data collection. Therefore, in experimental design the analyst has full control over the data set that is being collected, in the sense that the analyst decides upon e.g. the number of choice tasks that are being presented to each respondent, the number alternatives in a choice task, the attributes, the attribute levels, etc. This allows the analyst to construct the SC experiment in such a way that the model parameters can be recovered with high precision. In contrast, in case of SoO the analyst has no control over the data: the data are already collected. The objective of SoO is rather to extract a subset of observations from the full data set that allows for statistically efficient recovery of the model parameters.

Importantly, the choice on the sampling model or its prior parameters does not impact on model fit inferences, such as which of a set of competing models best fits the data. In other words, using a simplified version of model A as a sampling model will not increase the probability of finding model A to outperform model B in terms of model fit, in case model B better describes the underlying data generating process. Firstly, SoO is conducted after the data are collected. Therefore, SoO cannot affect the data generating process itself. Secondly, the SoO does not make use of the actual choices to evaluate the 'information value' or to extract observations from the full data set. Therefore, the method is simply not equipped to make one model more likely than another.

2.3. Information measures

The sampling process is based on assessing the information value of the extracted subset of observations. The information value can be measured in different ways. Congruent with practice in efficient experimental design, statistical efficiency measures can be used, such as A-efficiency, C-efficiency, D-efficiency, and S-efficiency (Kessels et al., 2006). These statistical efficiency measures are generally based on the Asymptotic Variance Covariance (AVC) matrix, denoted Ω, of the parameter estimates, which is a function of the pooled data $X \equiv [X_{nt}]$, the (prior) parameters, β, and the model specification of the sampling model. The AVC matrix is equal to the inverse of the Fisher Information matrix I, see Equation (1), where Y denotes the vector of choice observations. The Fisher Information matrix can be computed by taking the second-order derivatives of the Log-Likelihood (LL) function w.r.t. the model parameters and taking the expectation w.r.t. Y.

$$\begin{aligned} &\Omega(\beta|X) = (I(\beta|X))^{-1}, \\ &\text{where} \\ &I(\beta|X) = -E_Y\left(\frac{\partial^2 \log L(\beta|X,Y)}{\partial\beta\partial\beta}\right). \end{aligned} \tag{1}$$

In this research we limit our attention to one type of sampling model: the MNL model. This model is a particularly suitable candidate for the sampling model. The MNL allows for fast computation of choice probabilities, and its second order derivatives of the LL function w.r.t. the model parameters are simple to compute. Importantly, as shown in McFadden (1974), for the MNL model the second-order derivatives of the LL function do not depend on Y (i.e. vector of choice observations). This characteristic is useful not only in experimental design where Y is not yet available, but also in our context as it means that the sampling mechanism is independent of the observed choices. Thus, the actual choices are not used to assess the information value of observations.

Although in this research we focus on the MNL model, other types of sampling models could be used as well. When deciding upon the sampling model the analyst needs to be careful. Two considerations should be kept in mind. Firstly, it is crucially important to avoid endogeneity between the sampling model and the actual choices. That is, a sampling model must never make use of the actual choices. Secondly, it is important to consider the computational speed to compute the Fisher information matrix. Since the impact of many thousands of exchanges on the statistical efficiency need to be evaluated, a closed-form formula for the choice probability generating function and simple second-order derivatives of the *LL* function are highly desirable. A particular promising other candidate sampling model is the Nested Logit model. The NL model has a closed-form formula for the choice probability generating function and has relatively simple second-order derivatives of the LL function. Choices Y do not immediately drop from the second-order derivatives of the *LL* function of the NL model (as they do in the MNL model). However, Fisher information is defined in terms of *expected* choices E(Y) as shown in Eq. (1), which can be replaced by choice probabilities, see Bliemer et al. (2009). By doing so, the Fisher information matrix of the NL model becomes independent of the choices and endogeneity is avoided.[3]

The *LL* function for the MNL model is given by Equation (2), where P_{nit} is the probability that decision-maker n chooses alternative i in choice task t, and is a function of β and data X. Note that X may also include socio-demographic data about the respondent such as age, gender, and income, in contrast to experimental design where such data is not available at the time of creating the survey. Therefore, SoO also enables optimising the mix of respondents included in the subset (and hence the respondents in the subset will in general not be, and does not have to be, a representative sample). In other words, SoO provides opportunities to deal with the case in which the full data set is not entirely representative for the target population, e.g. because the full data set is skewed in terms of gender balance. To deal with such imbalances, socio-demographic variables that are relevant for the choice behaviour need to be incorporated in the utility functions in the sampling model as well as in the sophisticated model (e.g. in the form of interaction terms). Thereby, the effects of socio-demographic variables are accounted for during the sampling stage, and the remaining unexplained variance in the sophisticated model is essentially noise.

$$\log L(\beta|X) = \sum_{n=1}^{N} \sum_{i=1}^{J_n} \sum_{t=1}^{T_n} Y_{nit} \log P_{nit}(X_{nt}|\beta) \tag{2}$$

Using Equation (1) to derive the Fisher information matrix (assuming a linear-additive utility function), the value in the cell in row k and column l can be calculated as:

$$(I(\beta|X))_{\beta_k\beta_l} = \sum_{n=1}^{N} \sum_{i=1}^{J_n} \sum_{t=1}^{T_n} P_{nsi} x_{nitk} \left(x_{nitl} - \sum_{j=1}^{J_n} P_{njt} x_{nitl} \right), \tag{3}$$

where x_{nitk} is the level of attribute k in alternative j in choice task t observed by respondent n.

Finally, in the context of this research we focus on one particular measure: the D-error statistic. The D-error statistic is the most commonly used measure of efficiency in experimental design practice, and is calculated by taking the determinant of the AVC matrix, see Equation (4). Accordingly, finding the subset involves searching for that set of observations that minimises the D-error.[4,5]

$$D = \det(\Omega(X|\beta)) \tag{4}$$

2.4. Searching the solution space

To find the subset we need to search the solution space. Numerous so-called exchange algorithms have been developed to solve this type of combinatoric problems, e.g. Federal, k-exchange, RSC (Randomisation, Swapping, and Cycling) algorithms, etc. These methods differ from one another in terms of sophistication and the speed at which they are able to find good candidate solutions, see Cook and Nachtsheim (1980) for an overview. Since we are not designing the data but rather just selecting observations from existing data, we can only use so-called row-based algorithms that keep the data in each row intact without modifications imposed by randomisation or swapping. Importantly, a row here refers to an entire choice set complemented with the observed choice from that choice set (note that this data structure is different from some estimation packages, e.g. Nlogit, Latent Gold, that use multiple rows per choice set). We implemented a classic Federov exchange algorithm (Fedorov, 1972). This algorithm is a widely used row based algorithm which is relatively easy to implement. Future research may however be targeted to find smarter search algorithms to more efficiently solve the sampling of observations optimization problem.

The Federov exchange algorithm is shown in Fig. 3. The Federov algorithm starts by generating a starting subset, by randomly drawing N^* observations from the full data set. After that, the algorithm selects the first observation x_1 in the subset and starts evaluating the effect of exchanging x_1 with candidate observations from the full data set x_j on the D-error. The exchange pair $x_1 \leftrightarrow x_j$ with the largest decrease on the D-error statistic is selected, and the exchange of observations is performed. In case multiple exchanges lead to the same

[3] Note that this approach is common practice when generating in the efficient designs for NL and (Panel) Mixed Logit models.

[4] Note that technically speaking it is computationally more efficient to maximize the determinant of the Fisher Information matrix. This avoids the need to take the inverse of the AVC matrix.

[5] Note that commonly in the experimental design literature the D-error is normalized by the power (1/M), where M is the number of parameters. However, from an optimization perspective this scaling factor is inconsequential, and we omit this normalization here.

```
Create starting subset by drawing N* observations from the full data set;
Calculate Fisher information matrix and D-error;
while exchange pairs are found that decrease the D-error do
    for observation X1 to observation X_{N*} do
        calculate the effect of exchanging observations on the D-error;
        if exactly one exchange pair results in the largest decrease in
         D-error then
            exchange the selected observation;
        else
            randomly select one of the exchange pairs that result in the
             largest decrease in D-error;
            exchange the selected observation;
        end
        update Fisher information matrix and D-error;
    end
end
```

Fig. 3. Federov exchange algorithm.

decrease in D-error, randomly one exchange pair is selected. The algorithm continues by evaluating the effect of exchanges of the next observation x_2, and goes on until all observations have been exchanged. Note that due to the additivity of Fisher information as shown in Equation (3) (w.r.t. the observations), evaluating the effect of an exchange $x_i \leftrightarrow x_j$ on the D-error is computationally cheap: it only involves subtracting Fisher information related to original observation x_i and adding Fisher information related to candidate observation x_j from the total Fisher information in the subset.

In case the data set is not too large, it is possible to evaluate all possible exchanges of $x_i \leftrightarrow x_j$. In that case, one take is enough: repeating the exchange procedure (i.e. starting at x_1 again) does not further decrease the D-error of the subset. Hence, in total $N^* \cdot (N - N^*)$ exchanges need to be evaluated. However, in case the data set is (very) large – e.g. consisting of 1 million or more observations – it is computationally burdensome to evaluate $N^* \cdot (N - N^*)$ exchanges. In that case, the analyst can start with a random starting set and makes exchanges from a randomly selected set of candidate observations which is drawn from the full data set. When this approach is used, repeating the whole exchange procedure – thus starting with new starting sets and exchanging observations from new sets of candidate observations – typically will help to further decrease the D-error as the algorithm may get stuck in local minima. Finally, it is important to note that we do not need to find the optimal design. Rather, we are looking for a sufficiently efficient design. Therefore, for instance, also greedy algorithms could be used.

3. Monte Carlo analyses

This section puts the proposed information-theoretic based sampling method to a test. To be able to draw conclusions regarding the merits of the method we conduct a series of Monte Carlo experiments. Ultimately, we are interested in the reduction of the estimation time that can be achieved using the method, while taking into account the statistical cost of the method (in terms of the precision at which the parameter estimates are recovered). Furthermore, although there is a priori no theoretical ground to expect the method to affect the properties of the estimator, e.g. cause biased estimates, or affect consistency, we also explore these properties. Specifically, in sections 3.1–3.3 we look into the situation in which the true DGP and the sampling model correspond nicely. In section 3.4 we investigate the less neat case in which the true DGP and the sampling model are substantially off. We show that even in that case the sampling method does not result in misguided outcomes, e.g. results in confirming the sampling model, or its prior parameters.

3.1. Data

The proposed sampling method is particularly suited for large Revealed Preference (RP) data sets. These data are typically characterized by high collinearity and low signal-to-noise ratios (hence, low ρ^2). Further, this type of data often consists of multiple observations per individual (i.e., panel data). To adequately test SoO, we need a data set that possesses these characteristics. As generating a synthetic data set possessing these properties is rather difficult, we instead use a real-world car route choice data set. However, in order to test bias and consistency we need to be sure on the true Data Generating Process (DGP). Therefore, rather than using the actual observed choices, we synthetically generated the choices. Using this approach, we have data set that allows assessing reductions in computation efforts in a realistic setting as well as testing for bias and consistency of the estimator.

The route choice data set contains route choices of drivers, living in the Greater Copenhagen area. The vehicles of these drivers were equipped with GPS devices, which collected traces of their routes during a number of weeks in 2011. The traces were mapped onto the road network. A doubly stochastic route generation method was used to generate the choice sets (Nielsen, 2000; Bovy and Fiorenzo-Catalano, 2007).[6] Up to a 100 route alternatives routes were generated to complement the route choice observations. In the post processing phase a substantial number of choice observations were removed from the final data set because they only contained one

[6] Note that the generation of the choice sets is unrelated to the proposed sampling method.

Table 1
Correlation between attribute levels of alternatives within choice sets.

	Free flow travel time	Congested travel time	Travel Time variability	Travel cost	Left turns	Right turns
Free flow travel time	1.00	0.62	0.30	0.98	0.28	0.30
Congested travel time	0.62	1.00	0.44	0.60	0.24	0.27
Travel Time variability	0.30	0.44	1.00	0.27	0.19	0.20
Travel cost	0.98	0.60	0.27	1.00	0.16	0.17
Left turns	0.28	0.24	0.19	0.16	1.00	0.60
Right turns	0.30	0.27	0.20	0.17	0.60	1.00

feasible route alternative. The final data set consist of 14,449 choice observations, from 2 to 100 route alternatives, see Prato et al. (2014) for a more detailed description of the data and the choice set generation process. However, although this data set is already quite large, for the purposes of this study we replicated the data 7 times, such that we end up with a data set consisting of a little more than 100,000 observations. A data set of this size is large enough to run directly into memory problems when estimating sophisticated choice models. Therefore, by doing so, we have created a data set of the type and size for which our proposed method is particularly valuable.

Each alternative consists of six generic attributes: *Free flow travel time*, *Congested travel time, Travel Time variability, Travel cost, Left turns, and Right turns*. Travel cost is calculated based on distance. As is typical in many RP data sets, the attribute levels in this data set are strongly correlated. Table 1 shows the Pearson linear correlation coefficients. As can be seen, especially attribute levels of free flow travel time and travel cost are highly correlated ($\rho = 0.98$). All correlations are highly statistically significant ($p < 0.000$).

For these data we created 10,115 decision-makers, each making 10 route choices. Decision-makers are assumed to make decisions based on Random Utility Maximization (RUM) principles. Utility is assumed to be linear and additive (see Equation (5)). In route choice models, commonly path-size correction factors are used to account for correlation in unobserved utilities between partially overlapping routes. However, in this study we ignore these correction factors. We do so, in order to demonstrate our model in a more general choice modelling context. Unobserved taste heterogeneity is assumed to be present for 3 attributes: Free flow travel time, Congested travel time, and Travel time variability. Specifically, β_{TTF}, β_{TTC} and β_{TTV} are assumed to be normally distributed across decision-makers (but stable across choices of the same decision-maker).

$$U_{in} = \beta_{TTF} TTF_{in} + \beta_{TTC} TTC_{in} + \beta_{TTV} TTV_{in} + \beta_C C_{in} + \beta_L L_{in} + \beta_R R_{in} + \varepsilon_{in} \tag{5}$$

where

TTF_{in} = Free flow travel time [min]
TTC_{in} = Congested travel time [min]
TTV_{in} = Travel time variability [−]
C_{in} = Travel cost [DKr]
L_{in} = Number of left turns [−]
R_{in} = Number of right turns [−]
ε_{in} = Unobserved utility

Table 2 shows the parametrisation of the DGP. These values were taken from (Prato et al., 2014), although we rounded-off the parameter estimates for the sake of simplicity. To generate the choices every decision-maker is attributed a draw from the associated normal densities, for each marginal utility (β_{TTF}, β_{TTC} and β_{TTV}), and pseudo-random draws are generated from the Extreme Value Type I distribution for every alternative in each choice observation. In line with RUM principles, the alternative with the highest total utility is chosen. Finally, in order to avoid the risk of presenting results that are caused by a particular set of draws, we created 100 data sets. This enables us to investigate properties of the estimator, such as unbiasedness and consistency.

3.2. Sampling of observations procedure

To assess the effect of the size of the subset we sampled 6 subsets varying in size from 1000 to 5000 observations. For the sampling model we used a linear-additive RUM-MNL model, specified as shown in Equation (5). To obtain the priors for this model, we estimated the sampling model based on the full data set. The following point estimates were obtained: $\beta_{\mathrm{TTF}} = -0.61$, $\beta_{\mathrm{TTC}} = -0.26$, $\beta_{\mathrm{TTV}} = -0.18$, $\beta_{\mathrm{C}} = -0.30$, $\beta_{\mathrm{L}} = -0.70$ and $\beta_{\mathrm{R}} = -0.49$ (Appendix A shows the full estimation results of the sampling model). Comparing these sampling model estimates with the true model parameters in Table 2 shows that the sampling model provides fairly accurate prior parameters, albeit the means of the parameters β_{TTF}, β_{TTC} and β_{TTV} are somewhat underestimated.

The sophisticated model that we aim to estimate is a Panel Mixed Logit model. The utility specification of this Panel Mixed Logit model is congruent with the true DGP, as shown in Equation (5). The Panel Mixed Logit models are estimated using Simulated Maximum Likelihood. Given the large number of Panel Mixed Logit models that estimated (6 subsets × 100 repetitions of the data), we used a somewhat low number of draws, 250. Specifically, we used Modified Latin Hypercube Sampling (MLHS) draws, as these performs well with low numbers of draws (Hess et al., 2006).

With regard to the sampling procedure, we tried restricted and unrestricted sampling of observations. In the restricted case, additional constraints were added to the SoO procedure. Specifically, in light of the results of Rose et al. (2009) – who showed that when estimating Panel Mixed Logit models, having more than one observations per individual improves the statistical efficiency of retrieving

Table 2
Parametrisation of DGP.

Parameter	True value/distribution
β_{TTF}	N (-0.7,0.3)
β_{TTC}	N (-0.3,0.3)
β_{TTV}	N (-0.2,0.1)
β_{C}	−0.3
β_{L}	−0.7
β_{R}	−0.5

the standard deviations of random parameters – we tried a restricted sampling procedure in which a minimum of 2 observations per respondent were sampled. However, the merits of adding this constraint (as compared to the unrestricted sampling), were found to be relatively small. Our tentative conclusion is that the number of observations sampled per individual may provide the analyst an additional knob to tweak the sampling algorithm such that it balances the precision at which the non-random parameters and the random parameters are recovered (although based on our data set its merits were relative small). However, further research is needed to explore the costs and benefits of such restriction in more detail. For reasons of clarity, in our results section we report the results of the unrestricted sampling procedure only. Finally, we also have tested the impact of using A-efficiency instead of D-efficiency. Results of these analyses (not shown for reasons of brevity) are rather similar to those that we present in the next section based of D-efficiency.

3.3. Results

This subsection investigates the efficacy of the SoO method. To assess the efficacy of the method we compare SoO with a naïve random sampling approach, in which a subset of N^* observations is randomly taken from the full data set. Although other sampling techniques and methods to deal with large data sets have been devised, in particular in the field of Machine Learning, we believe using naïve random sampling as a benchmark to compare with is most insightful as this approach is de facto the standard practice in choice modelling. Accordingly, to make these comparisons, we estimated another 6×100 Panel Mixed Logit models.

3.3.1. D-error

Fig. 4 shows the mean D-error as a function of the number of observations in the subset across the 100 repetitions of the data, obtained using SoO (red) and using a naïve sampling method (blue). The D-errors, depicted on the y-axis, are computed based on the AVC matrix of the estimated Panel Mixed Logit models. Furthermore, power functions are fitted to the data points.

The results in Fig. 4 show that SoO is able to reduce the number of observations at relatively limited statistical cost (note the logarithmic scale on the x-axis). Reducing the number of observations using a naïve random sampling approach results in substantially larger D-errors, and hence comes at a substantially higher statistical cost.

To further highlight the gains in terms of data reduction that can be achieved by using SoO as compared to a naïve random sampling approach, Fig. 5 shows the number of observation sampled using SoO as a function of the number of naïvely sampled observations that result in the same statistical efficiency (D-error). For instance, it show that the statistical efficiency of 4000 naively sampled observations is equivalent to that of 1000 observations sampled using the proposed method. Hence, using the proposed method the data can be reduced by a factor 4. We believe the data reductions of that order of magnitude can be considered worthwhile. As computational effort

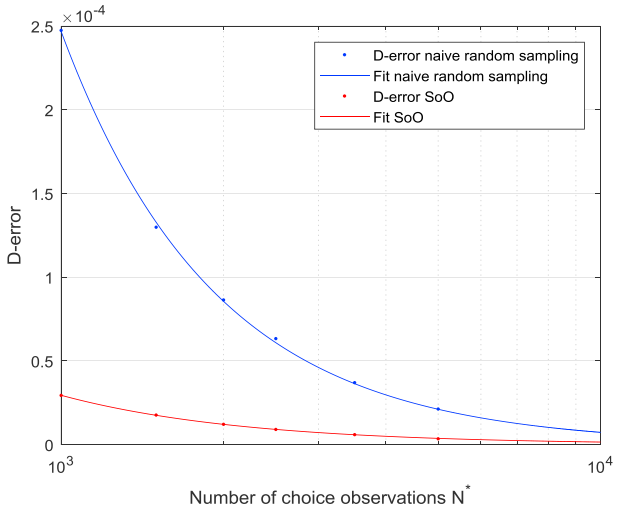

Fig. 4. D-error as a function of the sample size.

and memory use of sophisticated discrete choice models typically scale linear at best with the number of observations, at least similarly sized reductions of computational times can be obtained. Moreover, when the data set is substantially larger than the one we use here, it is even likely that even larger reductions can be achieved by means of information-theoretic sampling of observations.

3.2.2. Parameter estimates

The scatter plots in Fig. 6 and Fig. 7 show respectively the point estimates of the taste parameters and of the standard deviations (σ_{TTF}, σ_{TTV}, σ_{TTC}) associated with the random parameters, based on the 100 repetitions of the data. Fig. 6 and Fig. 7 show the results for a subset consisting of $N^* = 1000$ observations. The *x*-axis shows the parameter estimates obtained using a naïve random sampling approach; the *y*-axis shows the parameter estimates obtained using SoO. The vertical and horizontal dotted black lines depict the true value of the parameter. Finally, the black star indicates the mean of the parameter estimates obtained using the naïve sampling method (*x*-coordinate) and obtained using SoO (*y*-coordinate), across the 100 repetitions of the data.

Based on Fig. 6 and Fig. 7 two inferences can be made. Firstly, the point clouds are horizontally stretched. This is in line with expectations and reflects the fact that when using SoO the parameters are recovered with a higher precision (i.e. smaller standard errors). The horizontally stretched clouds are more prominently present in Fig. 6 than in Fig. 7. Hence, the increase in statistical efficiency due to using SoO (as compared to the naïve random sampling approach) is relatively larger for the taste parameters than for the standard deviations. This is in line with intuition. Since the MNL sampling model does not account for the presence of unobserved taste heterogeneity, it is not equipped to optimise the retrieval of these standard deviations. Secondly, there is no sign of systematic bias due to the information-theoretic based sampling methodology. As can be seen, the mean of the parameter estimates (i.e. the black star) typically tend to be very close to the true parameter values (indicated by the black dotted lines).

To give a more complete picture of the performance of SoO, Table 3 reports for each subset the following statistics, based on the 100 repetitions of the data.

- o Average. This is the arithmetic mean of the parameter estimates across the 100 repetitions of the data.
- o *Bias.* This is the mean of the differences between the true parameter and the parameter estimates across the 100 repetitions of the data.
- o *Sampling standard deviation.* This is standard deviation of the parameter estimates across the 100 repetitions of the data.
- o *Root Mean Square Error (RMSE).* This is the root of the mean of the errors squared, where the error is the difference between true and predicted parameter values. A smaller RMSE indicates better small sample efficiency.
- o *t-test.* The *t*-test gives the ratio between the bias and the sampling standard deviation of parameter estimates. This statistic is used to test whether the mean of the sampling distribution is equal to the true value. A *t*-statistic larger than 1.96 indicates that the hypothesis that the mean of the sampling distribution is equal to the true value is rejected at a significance level of $\alpha = 0.05$.

Table 3 indicates that SoO does not bias the estimation results. All *t*-statistics are much smaller than 1.96. This shows that the differences between the means of the sampling distribution and the true values are not significantly different from one another. Furthermore, the RMSEs of SoO are substantially smaller than those of the naïve random sampling approach. This shows that SoO improves the efficiency of the estimator (as compared to the naïve random sampling approach). Finally, Table 3 shows that the RMSE decreases with an increasing number of observations sampled. This indicates that SoO does not impact on the consistency of the estimator.

3.3.3. Standard error

Fig. 8 & 9 depict the mean asymptotic standard error of the parameters across the 100 repetitions of the data, based on the SoO

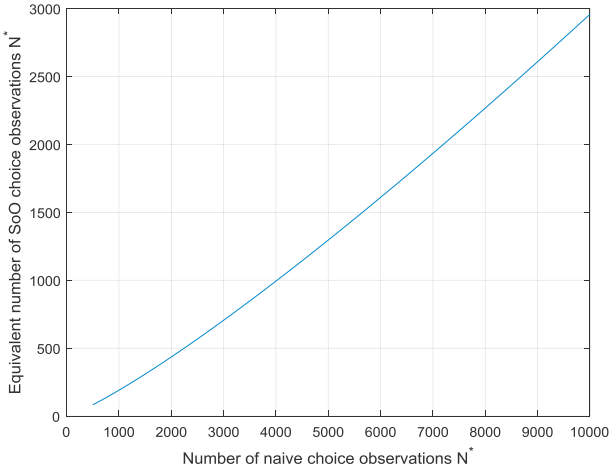

Fig. 5. Data reduction: SoO compared to naïve random sampling.

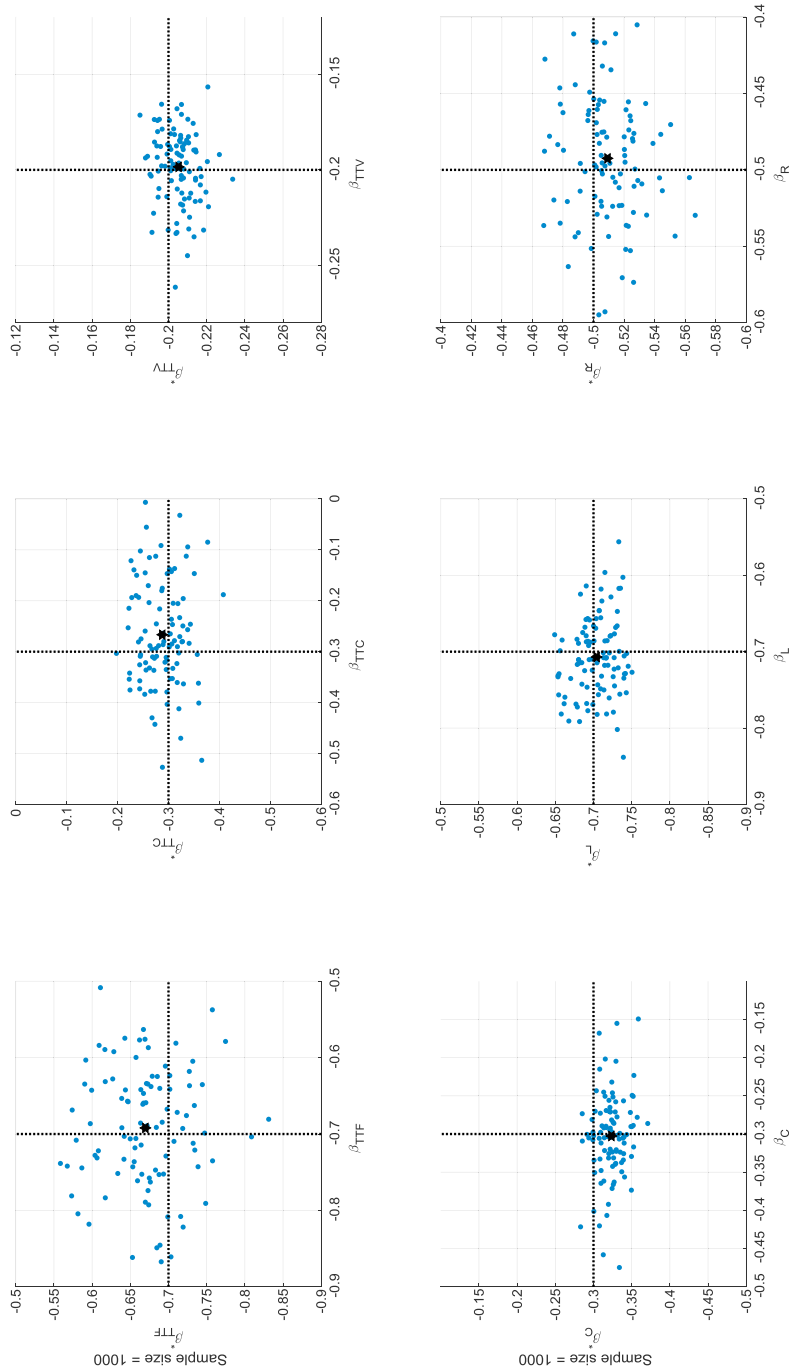

Fig. 6. Point estimates of β; $N = 1000$.

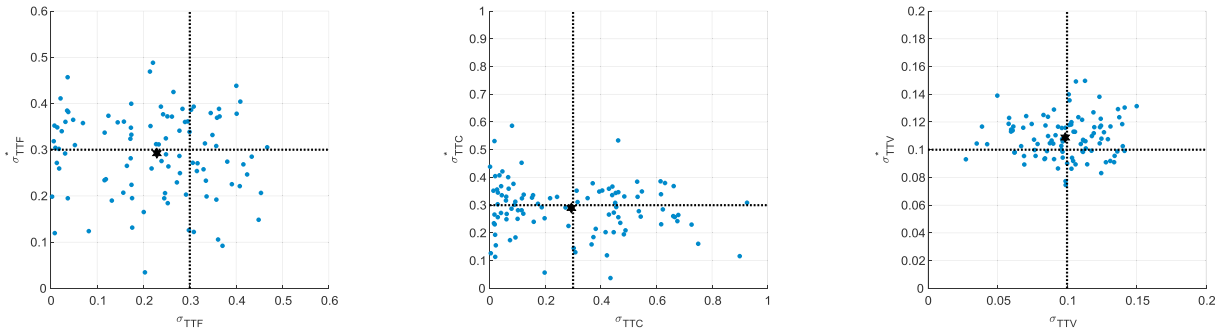

Fig. 7. Point estimates of σ; $N = 1000$.

method (red) and on a naïve random sampling approach (blue). Power functions ($f = a{\cdot}N^b$) are fitted to the data points to assess the general trend. Fig. 8 & 9 show that the standard errors obtained using SoO are considerable smaller than those obtained using the naïve sampling strategy (blue), over the whole range of N^*. This again highlights that SoO enables recovering of the model parameters with a considerably higher precision, as compared to naïve random sampling.

3.4.4. Computational efforts

The ultimate goal of SoO is to reduce the computational burden to estimate sophisticated discrete choice models. Therefore, it is essential that the computational efforts to down-sample the data are small as compared to the computational efforts to estimate sophisticated discrete choice models.

Table 4 reports the computation times to construct the subset (i.e. the down sampling time) and to estimate the Panel Mixed Logit models. The sampling method as well as the estimations are implemented in MATLAB R2016b, and all estimations and sampling are conducted on a High Performance Computing (HPC) cluster, consisting of 16 cores. Although we used a HPC cluster, it is important to note that SoO is very suitable for desktops computing. The main reason why we used a HPC cluster is that we repeated the SoO procedure and model estimation 100 times for each sampled subset, in order to investigate properties of the estimator. So, we utilized the HPC cluster to estimate 16 models at the same time, rather than to estimate them sequentially.

Table 4 shows that down-sampling is effective in reducing the time to estimate the Panel Mixed Logit models. Let us give an example to interpret this table. In case we down-scale this data set to $N^* = 1000$ observations – which takes just 12 min – estimation of the potentially numerous Panel Mixed Logit models during the model specification phase take just 18 min each. This means that we can easily test 10 competing specifications in half a day. In contrast, in case we do not down-scale the data each Panel Mixed Logit model takes about 2 h to estimate as in order to attain the parameters with the same precision we would need a naïve subset consisting of $N^* = 4000$ observations (see Fig. 5). In practice, this means it would take multiple days to test 10 competing model specifications. Therefore, we can conclude that SoO is a potentially valuable approach to reduce computationally times of estimating sophisticated discrete choice models in the context of large choice data.

Although we illustrate SoO in the context of a moderately sized large data set ($N = 1e^5$), we are confident that the method will perform well in the context of much larger data sets too. It goes without saying that down sampling time increases with data size. But, there are also clever ways to make SoO also work in the context of even larger data sets. As discussed in section 2.4, it is possible to execute the exchanges of pairs of observations in the Federov algorithm based on a randomly selected set of candidate observations, rather than based on all observations in the remaining data. By doing so, the computational efforts of down sampling remain manageable in the context of very large data.

3.4. Mismatch between DGP and sampling model

In sections 3.1–3.3 we have considered the 'ideal' situation in which the true DGP and the sampling model correspond nicely. That is, the sampling model is well-chosen: it relatively accurately reflects the true DGP. In this section we investigate the opposite situation, i.e. we look into the situation in which the true DGP and the sampling model are considerably off. We do so to understand how sensitive the method is towards such mismatches, and in particular to see whether, or not, a mismatches may result in misguided outcomes, e.g. leading to a tendency to 'confirm' the 'wrong' sampling model.

3.4.1. Two cases

There are two situations in which a mismatch between DGP and sampling model is likely to occur. The first situation occurs when the analyst starts exploring the data and has limited knowledge on all the attributes that are relevant to the choice. This means that the sampling model is likely to be a curtailed version on the true underlying DGP. That is, the sampling model does not contain of all attributes that are relevant to explain the choices. The second situation is the opposite of the first situation and is particularly prone to happen when the analyst has collected the data with one or more hypotheses in mind. In that case the analyst may 'overspecify' the

Table 3
Average, bias, RMSE and *t*-test based on 100 repetitions of the data.

Data size	N = 101,113											
Subset size		$N^* = 1000$		$N^* = 1500$		$N^* = 2000$		$N^* = 2500$		$N^* = 3500$		$N^* = 5000$	
Sampling method		Naïve	SoO	Naïve	SoO	Naïve	SoO	Naïve	SoO	Naïve	SoO	Naïve	SoO
	True												
β_{TTF}													
average	−0.7	−0.692	−0.670	−0.693	−0.675	−0.692	−0.671	−0.682	−0.680	−0.688	−0.682	−0.692	−0.684
bias		0.008	0.030	0.007	0.025	0.008	0.029	0.018	0.020	0.012	0.018	0.008	0.016
sampling std dev		0.078	0.053	0.065	0.039	0.056	0.031	0.052	0.024	0.045	0.022	0.041	0.021
RMSE		0.078	0.060	0.065	0.047	0.056	0.042	0.055	0.031	0.046	0.029	0.041	0.026
t-test		0.102	0.575	0.113	0.643	0.148	0.935	0.350	0.842	0.259	0.822	0.195	0.764
β_{TTC}													
average	−0.3	−0.267	−0.288	−0.281	−0.294	−0.274	−0.292	−0.275	−0.292	−0.293	−0.292	−0.275	−0.298
bias		0.033	0.012	0.019	0.006	0.026	0.008	0.025	0.008	0.007	0.008	0.025	0.002
sampling std dev		0.103	0.040	0.073	0.033	0.087	0.028	0.071	0.026	0.050	0.024	0.047	0.024
RMSE		0.108	0.042	0.075	0.033	0.090	0.029	0.075	0.027	0.050	0.025	0.053	0.024
t-test		0.323	0.304	0.266	0.198	0.294	0.284	0.358	0.305	0.138	0.319	0.527	0.094
B_{TTV}													
average	−0.2	−0.198	−0.205	−0.196	−0.201	−0.198	−0.198	−0.200	−0.198	−0.197	−0.198	−0.198	−0.197
bias		0.002	−0.005	0.004	−0.001	0.002	0.002	0.000	0.002	0.003	0.002	0.002	0.003
sampling std dev		0.019	0.009	0.013	0.007	0.012	0.005	0.011	0.005	0.009	0.004	0.007	0.004
RMSE		0.019	0.010	0.014	0.007	0.012	0.005	0.010	0.005	0.009	0.005	0.007	0.005
β_C													
t-test		0.082	−0.603	0.299	−0.143	0.140	0.319	0.026	0.410	0.292	0.439	0.272	0.729
average	−0.3	−0.303	−0.323	−0.302	−0.316	−0.303	−0.314	−0.309	−0.310	−0.308	−0.310	−0.303	−0.308
bias		−0.003	−0.023	−0.002	−0.016	−0.003	−0.014	−0.009	−0.010	−0.008	−0.010	−0.003	−0.008
sampling std dev		0.055	0.017	0.046	0.013	0.043	0.011	0.039	0.010	0.034	0.009	0.025	0.008
RMSE		0.055	0.029	0.046	0.020	0.043	0.018	0.040	0.014	0.035	0.014	0.025	0.012
t-test		−0.048	−1.346	−0.040	−1.243	−0.070	−1.185	−0.239	−0.993	−0.227	−1.153	−0.108	−1.031
β_L													
average	−0.7	−0.707	−0.704	−0.693	−0.705	−0.698	−0.702	−0.704	−0.704	−0.701	−0.703	−0.702	−0.699
bias		−0.007	−0.004	0.007	−0.005	0.002	−0.002	−0.004	−0.004	−0.001	−0.003	−0.002	0.001
sampling std dev		0.051	0.025	0.037	0.021	0.036	0.020	0.032	0.019	0.026	0.017	0.022	0.015
RMSE		0.051	0.026	0.037	0.022	0.036	0.020	0.032	0.019	0.025	0.017	0.022	0.015
t-test		−0.137	−0.162	0.179	−0.250	0.042	−0.121	−0.115	−0.207	−0.035	−0.178	−0.084	0.071
β_R													
average	−0.5	−0.492	−0.509	−0.506	−0.499	−0.499	−0.501	−0.501	−0.502	−0.499	−0.501	−0.496	−0.501
bias		0.008	−0.009	−0.006	0.001	0.001	−0.001	−0.001	−0.002	0.001	−0.001	0.004	−0.001
sampling std dev		0.041	0.020	0.033	0.017	0.026	0.015	0.025	0.015	0.020	0.012	0.016	0.010
RMSE		0.042	0.022	0.033	0.017	0.026	0.015	0.025	0.015	0.020	0.012	0.016	0.010
t-test		0.183	−0.450	−0.179	0.029	0.023	−0.084	−0.021	−0.114	0.062	−0.100	0.268	−0.134
σ_{TTF}													
average	0.3	0.229	0.292	0.267	0.298	0.270	0.284	0.263	0.286	0.273	0.281	0.276	0.294
bias		−0.071	−0.008	−0.033	−0.002	−0.030	−0.016	−0.037	−0.014	−0.027	−0.019	−0.024	−0.006
sampling std dev		0.136	0.092	0.122	0.075	0.103	0.068	0.101	0.063	0.079	0.057	0.066	0.037
RMSE		0.153	0.092	0.126	0.075	0.107	0.069	0.107	0.064	0.083	0.060	0.070	0.037
t-test		−0.522	−0.082	−0.267	−0.023	−0.296	−0.236	−0.368	−0.231	−0.343	−0.329	−0.360	−0.170
σ_{TTC}													
average	0.3	0.294	0.290	0.323	0.288	0.293	0.273	0.320	0.267	0.313	0.262	0.327	0.281
bias		−0.006	−0.010	0.023	−0.012	−0.007	−0.027	0.020	−0.033	0.013	−0.038	0.027	−0.019
sampling std dev		0.239	0.095	0.212	0.095	0.198	0.082	0.172	0.080	0.166	0.078	0.136	0.060
RMSE		0.238	0.095	0.212	0.096	0.197	0.086	0.172	0.086	0.166	0.087	0.138	0.062
t-test		−0.026	−0.106	0.110	−0.127	−0.038	−0.334	0.119	−0.415	0.081	−0.487	0.197	−0.319
σ_{TTV}													
average	0.1	0.099	0.109	0.093	0.109	0.098	0.110	0.100	0.110	0.099	0.107	0.099	0.104
bias		−0.001	0.009	−0.007	0.009	−0.002	0.010	0.000	0.010	−0.001	0.007	−0.001	0.004
sampling std dev		0.026	0.015	0.021	0.013	0.019	0.011	0.018	0.010	0.013	0.009	0.010	0.008
RMSE		0.026	0.017	0.022	0.016	0.019	0.015	0.018	0.014	0.013	0.011	0.010	0.009
t-test		−0.050	0.592	−0.324	0.677	−0.118	0.932	0.022	1.035	−0.076	0.807	−0.074	0.492

sampling model in the sense that the sampling model includes attributes that actually have no explanatory power for the choices.

Case 1: The sampling model is curtailed.

To investigate this situation we use the same Panel Mixed Logit DGP as before (see equation (5)) with the parametrisation as given in Table 2. The analyst is however not aware at the moment (s)he conducts the down-sampling step that the number of left and right turns

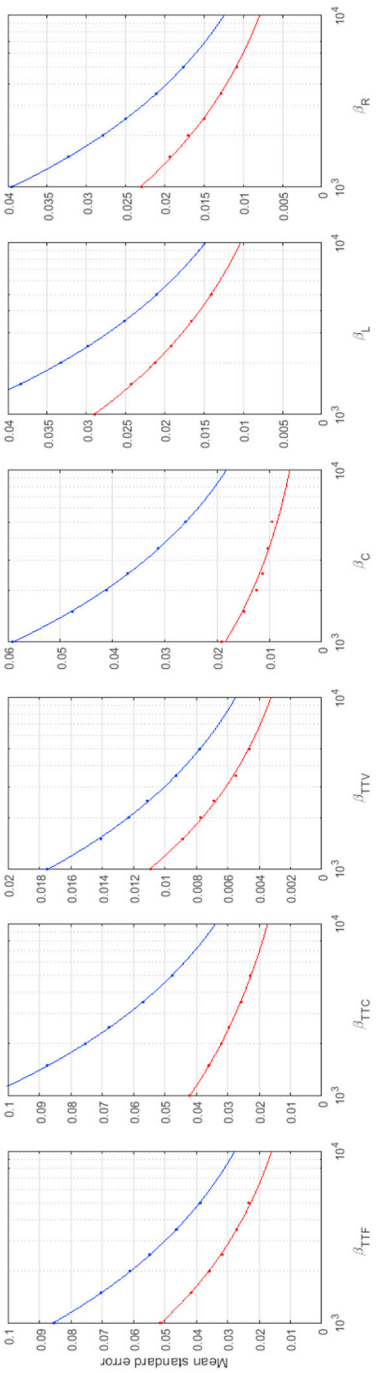

Fig. 8. Standard error of the model's taste parameters as a function of the sample size N^*.

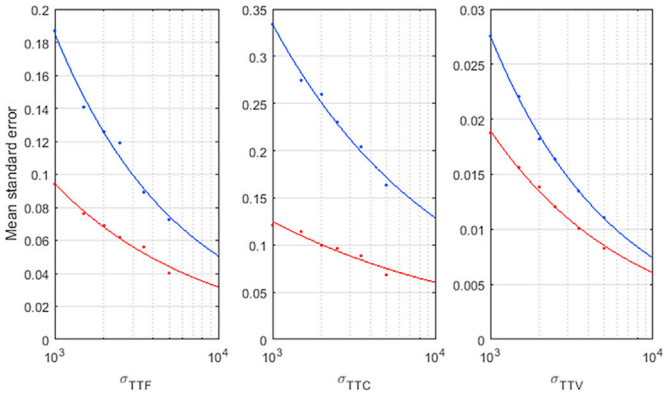

Fig. 9. Standard error of σ_{TTF}, σ_{TTC} and σ_{TTV} as a function of the sample size N^*.

Table 4
Computational time for estimation and down sampling.

Subset size N*	Estimation time Panel Mixed Logit model [min]	Down sampling time [min]
1000	18	12
1500	32	18
2000	45	24
2500	55	31
3500	91	48
5000	130	70
7500	215	98
10000	280	120

are relevant explanatory variables for the observed route choices. Therefore, the utility function of the sampling model does not include β_L and β_R, see equation (6). After the down sampling the analyst tests a number of competing specifications, including a specification that is equal to the true DGP which includes the number of left and right turns as explanatory variables. In this case we are specifically interested to see that the analyst is able to recover the true DGP, i.e. recover the correct sizes β_L and β_R. More generally speaking, we would like to establish that the estimator is unbiased, regardless of the (poor) specification of the sampling model.

$$U_{in} = \beta_{TTF} TTF_{in} + \beta_{TTC} TTC_{in} + \beta_{TTV} TTV_{in} + \beta_C C_{in} + \varepsilon_{in} \tag{6}$$

Case 2: The sampling model is over specified

To investigate this situation we use the Panel Mixed Logit DGP as before (equation (5)) with the parametrisation as given in Table 2, but this time we set the parameters associated with left and right turns (β_L and β_R) to zero. The analyst in this case however believes that the parameters associated with left and right turns (β_L and β_R) are non-zero. Therefore, the analyst uses non-zero prior parameters in the sampling model associated with left and right turns: $\beta_L = -0.7$ and $\beta_R = -0.5$. After the sampling stage the analyst will test these hypotheses. To do so (s)he will estimate a Panel Mixed Logit model which includes β_L and β_R. In this case we are specifically interested to see that the analyst is able to accept the null-hypotheses, i.e. find that β_L and β_R are not significantly different from zero. In other words, we would like to establish that the estimator is unbiased in the sense that it does create a tendency to confirm the 'hypotheses' that were used to parameterise the sampling model.

3.4.2. Results

Table 5 shows the Panel Mixed Logit estimation results for case I and II. It shows the average, bias, RMSE, sampling standard deviation and a *t*-test, based on 100 repetitions of the data. Let's first look at the results for case I. Table 5 shows that a curtailed sampling model does not hamper the analyst to recover the true DGP. Specifically, the average values of β_L and β_R are very close to the true values (respectively -0.7 and -0.5). The associated t-values indeed show that β_L and β_R are not significantly different from their true values. So, the estimator is unbiased, despite the use of a poorly specified (curtailed) sampling model. Moreover, looking at the sampling standard deviations and the RMSEs we see that despite the fact that β_L and β_R are not taken into account during the sampling stage, their sizes are recovered at a considerable higher precision than in case a naïve sampling method would have been used.

Looking at the results for case II, we can see that an overspeficied sampling model does not lead to a confirmation of the sampling

Table 5
Average, bias, RMSE and *t*-test based on 100 repetitions of the data.

$N = 101{,}143$						
Case			I		II	
Sampling method			Naïve	SoO	Naïve	SoO
		True				
β_{TTF}						
	average	−0.7	−0.692	−0.675	−0.703	−0.676
	bias		0.008	0.025	−0.003	0.024
	sampling std dev		0.078	0.060	0.070	0.037
	RM SE		0.078	0.064	0.069	0.044
	t-test		0.102	0.422	−0.042	0.644
β_{TTC}						
	average	−0.3	−0.267	−0.289	−0.268	−0.285
	bias		0.033	0.011	0.032	0.015
	sampling std dev		0.103	0.042	0.094	0.041
	RM SE		0.108	0.044	0.099	0.043
	t-test		0.323	0.254	0.341	0.371
β_{TTV}						
	average	−0.2	−0.198	−0.206	−0.201	−0.202
	bias		0.002	−0.006	−0.001	−0.002
	sampling std dev		0.019	0.011	0.015	0.008
	RM SE		0.019	0.013	0.015	0.008
	t-test		0.082	−0.556	−0.038	−0.287
β_{C}						
	average	−0.3	−0.303	−0.327	−0.303	−0.317
	bias		−0.003	−0.027	−0.003	−0.017
	sampling std dev		0.055	0.018	0.056	0.021
	RM SE		0.055	0.033	0.056	0.027
	t-test		−0.048	−1.492	−0.048	−0.790
β_{L}						
	average	−0.7/0	−0.707	−0.695	0.004	0.000
	bias		−0.007	0.005	0.004	0.000
	sampling std dev		0.051	0.033	0.033	0.021
	RM SE		0.051	0.033	0.033	0.021
	t-test		−0.137	0.142	0.111	−0.012
β_{R}						
	average	−0.5/0	−0.492	−0.518	0.001	−0.006
	bias		0.008	−0.018	0.001	−0.006
	sampling std dev		0.041	0.032	0.026	0.013
	RM SE		0.042	0.036	0.026	0.015
	t-test		0.183	−0.550	0.021	−0.429
σ_{TTF}						
	average	0.3	0.229	0.275	0.294	0.288
	bias		−0.071	−0.025	−0.006	−0.012
	sampling std dev		0.136	0.089	0.106	0.054
	RM SE		0.153	0.092	0.105	0.055
	t-test		−0.522	−0.280	−0.055	−0.223
σ_{TTC}						
	average	0.3	0.294	0.301	0.272	0.278
	bias		−0.006	0.001	−0.028	−0.022
	sampling std dev		0.239	0.090	0.221	0.101
	RM SE		0.238	0.090	0.221	0.103
	t-test		−0.026	0.012	−0.125	−0.216
σ_{TTV}						
	average	0.1	0.099	0.100	0.099	0.103
	bias		−0.001	0.000	−0.001	0.003
	sampling std dev		0.026	0.020	0.022	0.012
	RM SE		0.026	0.020	0.021	0.012
	t-test		−0.050	0.015	−0.035	0.232

model. Specifically, the average values of β_L and β_R are very close to the true values (zero). The associated t-values indicate that β_L and β_R are not significantly different from zero. Thus, despite the use of non-zero priors for β_L and β_R in the sampling model, the analyst will not erroneously reject the null-hypotheses when testing them based on the sampled data.

4. Conclusions and discussion

This paper presented a new method to lower the computational burden of estimating sophisticated discrete choice models based on

large data sets. This method – which we call Sampling of Observations (SoO) – scales down the size of the choice data set based on information-theoretic principles. SoO extracts a subset of observations from the full data set which is much smaller in volume than the original data set, yet produces nearly identical result. The method is inspired by, and closely related to, efficient experimental design theory. SoO and efficient experimental design have in common that they both aim to maximize the precision at which the model parameters are retrieved from the data. A fundamental difference between SoO and efficient experimental design is however that SoO takes place *after* the data are collected, while experimental design takes place prior to the data collection.

Our results show that the SoO method is promising. Specifically, using Monte Carlo analyses, we show that: 1) SoO is able to scale down data at relatively little statistical cost, 2) SoO does not bias parameter estimates, even when the sampling model does not accurately reflect the true DGP, and 3) SoO can substantially reduce the computational burden to estimate sophisticated discrete choice models based on large data sets. The sampling of observations takes only a fraction of the time that is needed to estimate a sophisticated discrete choice model. Given that discrete choice modelling typically involves estimating numerous competing model specifications, we believe the reductions in computational efforts achieved by SoO are worthwhile. However, notwithstanding those advantages, as long as SoO is not available in off-the-shelf software packages, applying SoO requires considerable efforts on the side of the analyst.

Despite these promising results, we believe it is important to point out that analysts need to be cautious when applying this method. It is particularly important to be aware of potential endogeneity issues.[7] One particular example of this occurs when there is association between preferences and choice sets. In such a situation, the method is no longer 'guaranteed' to result in unbiased outcomes. Association between preferences and choice sets may, for instance, occur in a mode choice setting in which people with low preference for Public Transport (PT) live in areas without PT access. As a consequence, their (imputed) choice sets do not contain PT alternatives – creating association between preferences and choice sets composition. Although the relation between the sampling model (including its prior parameters) and the sampled observations is highly complex, and presumably weak, it is certainly not ruled out that such association between preferences and choice sets will not create distortions in the parameter estimates. Further research is needed to look into this type of situations in more detail.

The methodology presented in this paper provides ample scope for further research. Firstly, our analysis is based on one combination of sampling model and sophisticated model and on one data set. Further research is needed to investigate the efficacy of SoO in the context of other choice models (e.g. Latent class and Integrated Choice and Latent Variable models) and other data sets (e.g. data sets which are different from one another in terms of number of observations, number of alternatives and the number of explanatory variables). More experience with the method would also help to devise guidelines for the method, e.g. on when to use what type of sampling model and on the required number of sampled observations. Also, comparisons of the performance of this method with other sampling or estimation techniques developed in the data sciences to work with large data sets (e.g. gradient descent techniques) would be insightful. Secondly, in our analyses we assumed that there is no association between preferences and choice sets. Further research may investigate the effect of such association on the methodology. Thirdly, in this research we focussed mainly on one efficiency measure, namely D-efficiency. Exploring the use of e.g. Bayesian efficiency measures (Sandor and Wedel, 2005) to account for uncertainty on the model parameters on the side of the analyst, or composite efficiency measures (Rose and Bliemer, 2013a; Van Cranenburgh et al. 2018) to account for uncertainty on the model specification (e.g. in terms of the underlying decision rule) are promising avenues for further research. Finally, we hope this research inspires further work and discussions on methods to work with large choice data obtained in a new data landscape.

Statement of contribution

Due to the surge in the amount of data that are being collected, analysts are increasingly faced with very large data sets. Estimation of sophisticated discrete choice models (such as Mixed Logit models) based on these large data sets can be computationally burdensome, or even technically infeasible. This research contributes to the state-of-the-art on challenges in the field of choice modelling associated with utilizing the full potential of emerging data sources. It develops a new sampling method to lower the computational burden of estimating sophisticated discrete choice models based on large data sets. This method – which we call Sampling of Observations (SoO) – scales down the size of the choice data set based on information-theoretic principles.

Acknowledgements

The authors would like to thank Prof. Carlo G. Prato, dr. Thomas K. Rasmussen and Prof. Otto A. Nielsen for sharing their data with us.

Appendix B. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jocm.2018.02.003.

[7] We thank an anonymous reviewer for pointing this out.

Appendix A. Sampling model estimation results

Model	MNL			
No.observation	101,143			
No.individuals	10,115			
Null log-Likelihood	−277,631			
Final-Likelihood	−188,034			
ρ^2	0.32			
Estimation time [min]	2.4			
	Est	SE	t-stat	p-val
β_{TTF}	−0.605	0.0069	−87.45	0.00
β_{TTC}	−0.262	0.0095	−27.43	0.00
β_{TTV}	−0.178	0.0014	−131.48	0.00
β_c	−0.300	0.0055	−54.85	0.00
β_L	−0.698	0.0045	−153.81	0.00
B_R	−0.494	0.0038	−130.24	0.00

References

Arnaiz-González, Á., Díez-Pastor, J.-F., Rodríguez, J.J., García-Osorio, C., 2016. Instance selection of linear complexity for big data. Knowl. Base Syst. 107, 83–95.

Bliemer, et al., 2009. Efficient stated choice experiments for estimating nested logit models. Transp. Res. Part B Methodol. 43, 19–35.

Bliemer, M.C., Collins, A.T., 2016. On determining priors for the generation of efficient stated choice experimental designs. J. Choice Model. 21, 10–14.

Bliemer, M.C.J., Rose, J.M., Chorus, C.G., 2017. Detecting dominancy in stated choice data and accounting for dominancy-based scale differences in logit models. Transp. Res. Part B Methodol. 102, 83–104.

Bovy, P.H.L., Fiorenzo-Catalano, S., 2007. Stochastic route choice set generation: behavioral and probabilistic foundations. Transportmetrica 3 (3), 173–189.

Cook, R.D., Nachtsheim, C.J., 1980. A comparison of algorithms for constructing exact d-optimal designs. Technometrics 22 (3), 315–324.

Daly, A., Zachery, S., 1978. Improved multiple choice models. In: Hensher, D. (Ed.), Determinants of Travel Choice. Saxon House, Sussex.

de Bekker-Grob, E.W., Donkers, B., Jonker, M.F., Stolk, E.A., 2015. Sample size requirements for discrete-choice experiments in healthcare: a practical guide. The Patient - Patient-Cent. Outcomes Res. 8 (5), 373–384.

Farooq, B., Beaulieu, A., Ragab, M., Ba, V.D., 2015. Ubiquitous Monitoring of Pedestrian Dynamics: exploring Wireless ad hoc Network of Multi-sensor Technologies. 2015 IEEE SENSORS.

Fedorov, V.V., 1972. Theory of Optimal Experiments. Elsevier.

Ferrini, S., Scarpa, R., 2007. Designs with a priori information for nonmarket valuation with choice experiments: a Monte Carlo study. J. Environ. Econ. Manag. 53 (3), 342–363.

Fisher, R.A., 1925. Statistical Methods for Research Workers. Genesis Publishing Pvt Ltd.

Hess, S., Train, K.E., Polak, J.W., 2006. On the use of a modified Latin Hypercube sampling (MLHS) method in the estimation of a mixed logit model for vehicle choice. Transp. Res. Part B Methodol. 40 (2), 147–163.

Huber, J., Zwerina, K., 1996. The importance of utility balance in efficient choice designs. J. Market. Res. 33 (3), 307–317.

Iqbal, M.S., Choudhury, C.F., Wang, P., González, M.C., 2014. Development of origin–destination matrices using mobile phone call data. Transport. Res. C Emerg. Technol. 40, 63–74.

Jánošíková, Ľ., Slavík, J., Kohání, M., 2014. Estimation of a route choice model for urban public transport using smart card data. Transport. Plann. Technol. 37 (7), 638–648.

Kanninen, B.J., 2002. Optimal design for multinomial choice experiments. J. Market. Res. 39 (2), 214–227.

Kessels, R., Goos, P., Vandebroek, M., 2006. A comparison of criteria to design efficient choice experiments. J. Market. Res. 43 (3), 409–419.

Loyola, R.D.G., Pedergnana, M., Gimeno García, S., 2016. Smart sampling and incremental function learning for very large high dimensional data. Neural Network. 78, 75–87.

McFadden, D.L., 1974. Conditional logic analysis of qualitative choice behavior. In: Zarembka, P. (Ed.), Frontiers in Econometrics. Academic Press, New York, pp. 105–142.

McFadden, 1978. Modelling the choice of residential location. Transport. Res. Rec. 673.

Nielsen, O.A., 2000. A stochastic transit assignment model considering differences in passengers utility functions. Transp. Res. Part B Methodol. 34 (5), 377–402.

Prato, C., Rasmussen, T., Nielsen, O., 2014. Estimating value of congestion and of reliability from observation of route choice behavior of car drivers. Transport. Res. Rec.: J. Transport Res. Board 2412, 20–27.

Revelt, D., Train, K., 1998. Mixed logit with repeated choices: households' choices of appliance efficiency level. Rev. Econ. Stat. 80 (4), 647–657.

Rieser-Schüssler, N., 2012. Capitalising modern data sources for observing and modelling transport behaviour. Transport. Lett. 4 (2), 115–128.

Rose, J.M., Bliemer, M.C.J., 2009. Constructing efficient stated choice experimental designs. Transport Rev.: A Transnat. Transdiscipl. J. 29 (5), 587–617.

Rose, J.M., Bliemer, M.C.J., 2013a. Incorporating analyst uncertainty in model specification of respondent processing strategies into efficient designs for logit models. In: The 59th World Statistics Congress, 25–30 August 2013, hong Kong.

Rose, J.M., Bliemer, M.C.J., 2013b. Sample size requirements for stated choice experiments. Transportation 40 (5), 1021–1041.

Rose, J.M., Hess, S., Bliemer, M.C.J., Daly, A., 2009. The impact of varying the number of repeated choice observations on the mixed multinomial logit model. In: European Transport Conference, Leeuwenhorst, The Netherlands, pp. 5–7.

Sandor, Z., Wedel, M., 2005. Heterogeneous conjoint choice designs. J. Market. Res. 42 (2), 210–218.

Shannon, C.E., Weaver, W., 1949. The Mathematical Theory of Communication. University of Illinois Press, Urbana,.

Van Cranenburgh, S., Rose, J.M., Chorus, C.G., 2018. On the robustness of efficient experimental designs towards the underlying decision rule. Transport. Res. Pol. Prac 109, 50–64.

Vlahogianni, E.I., Park, B.B., van Lint, J.W.C., 2015. Big data in transportation and traffic engineering. Transport. Res. C Emerg. Technol. 58 (Part B), 161.

Walker, J.L., Wang, Y., Thorhauge, M., Ben-Akiva, M., 2017. D-efficient or deficient? A robustness analysis of stated choice experimental designs. Theor. Decis. (84), 215–238. http://dx.doi.org/10.1007/s11238-017-9647-3.

Witlox, F., 2015. Beyond the data smog? Transport Rev. 35 (3), 245–249.

Wong, M., Farooq, B., Bilodeau, G.A., 2017. Latent Behaviour Modelling Using Discriminative Restricted Boltzmann Machines. ICMC 2017, Cape Town.