Whose preferences matter more? Handling unbalanced panel data for choice modelling

Laurent Cazor*,†, Mirosława Łukawska†, Mads Paulsen†, Thomas Kjær Rasmussen†, and Otto Anker Nielsen†

*Corresponding author (lauca@dtu.dk)
†Technical University of Denmark, Department of Technology, Management and Economics, Bygningstorvet 116b, 2800 Kgs. Lyngby

February 28, 2023

Short summary

The emergence of GPS-enabled smartphones and crowdsourcing tools are unique opportunities for understanding transport behaviour. However, the datasets they generate are often unbalanced, as individuals may use the service collecting data at different frequencies and periods. This raises important questions: are typical discrete choice models robust to this unbalance? Are model estimates biased towards over-represented individuals?

This paper tackles the issue of handling unbalanced panel datasets for route choice modelling. It first develops a simulation experiment to study to which degree Mixed Logit Models with panel effects reproduce the population preferences using unbalanced data. It then investigates bias reduction strategies, using subsampling and likelihood weighting. These strategies are compared to give guidelines that fit the model purpose. We show that weighting and subsampling techniques can reduce the bias when interpreting the model output for tastes. Combining these techniques helps to find an optimal trade-off between bias and variance of the estimates.

Keywords: Unbalanced panel, panel mixed logit model, subsampling, likelihood weighting, bias-efficiency trade-off

1 Introduction

An increasing number of large crowd-sourced datasets are available for choice modelling. For instance, smartphone GPS datasets bring researchers new opportunities when analyzing bicycle traffic. They overcome some limitations related to stated preference data or small sample sizes (Nelson et al., 2021; Lee & Sener, 2021). However, due to their crowd-sourced opt-in nature, such datasets may suffer from having a large proportion of the data collected by only a few active users. This may be problematic if the models estimated on these datasets are used for policy implications or forecasting purposes, for instance if the preferences of these active users differ from the population mean.

The issue of repeated observations per individual in the panel setup for the mixed logit model has been addressed in the literature in the context of stated preference (SP) data. Bliemer & Rose (2010) recognized the advantages of panel setup and studied the construction of an optimal experiment design (in terms of the statistical properties of the model) for stated choice surveys with panel information. Rose et al. (2009) found that adding repeated choice observations per individual improves the model accuracy only until a certain point. Including multiple repeated observations from an individual, which are identical in terms of both the set of attributes and the choice outcome, can be used to account for the effect of e.g. habit (Cherchi & Cirillo, 2014) or correlation patterns (Cherchi et al., 2017).

Yáñez et al. (2011) found that the most significant improvement to the model in terms of fit can be attributed to the introduction of panel correlation. Furthermore, including multiple identical observations should not influence the efficiency of the estimated parameters and does not contribute to the improved capability to retrieve the true parameters. The latter can, however, be improved

by introducing weighting since it increases the influence of the fixed part of the utility over the random part.

Two recent studies (van Cranenburgh & Bliemer, 2019; Ortelli et al., 2022) recognize the challenges of estimating models based on rapidly emerging massive data sources. They propose strategies for a dataset size reduction by optimizing multiple criteria, such as model efficiency, estimation bias, out-of-sample performance, computational time, and value of time for relevant parameters. They propose optimizing the mixed logit model setup based on a simpler multinomial model (MNL).

However, these papers do not answer the question: are estimates biased toward individuals contributing more to a crowdsourced dataset? How to deal with a bias-efficiency trade-off? This paper aims to fill this gap, first testing whether discrete choice models estimated with unbalanced data represent the average tastes of the individuals of the sample population and then finding solutions to eliminate the potential sources of bias.

2 Methods

Estimated model: the Panel Mixed-Logit model (PMXL)

The Mixed Logit model with panel effects, also denoted Panel Mixed-Logit (PMXL), builds on the traditional Multinomial Logit (MNL) model (McFadden et al., 1973; McFadden & Train, 2000; Train, 2009).

The Panel Mixed-Logit model can be derived as follows: a decision maker $n \in \{1, ..., N\}$ has T_n choice situations $t \in \{1, ..., T_n\}$. For each situation t, the decision maker n can choose an alternative i from choice set C_{nt}, whose utility U_{nti} can be written as:

$$U_{nti} = V(\boldsymbol{\beta}_n, X_{nti}) + \epsilon_{nti}$$

$\boldsymbol{\beta}_n$ is a parameter representing the tastes of decision maker n, X_{nti} is a vector of attributes and ϵ_{nti} is a stochastic error term. Under the logit assumption, the ϵ_{nti}'s are independently and identically distributed (iid) according to Gumbel(0,1). The choice probability of alternative i by decision maker n in choice situation t (i.e., the event $(y_{nt} = i)$) is given by:

$$\mathbb{P}(y_{nt} = i|\boldsymbol{\beta}_n, X_{nti}) = \frac{e^{V(\boldsymbol{\beta}_n, X_{nti})}}{\sum\limits_{j \in C_{nt}} e^{V(\boldsymbol{\beta}_n, X_{nti})}} \tag{1}$$

We assume the utility is linear in parameters, i.e. that $V(\boldsymbol{\beta}_n, X_{nti}) = \boldsymbol{\beta}_n^\top X_{nti}$. The PMXL model assumes that tastes vary across individuals and that these tastes $\boldsymbol{\beta}_n \sim f(\boldsymbol{\beta}|\boldsymbol{\theta})$ are iid across decision-makers. The PXML probability of the sequence of choices $\mathbf{i_n} = i_1, ..., i_{T_n}$ for decision maker n is then given by:

$$P_{\mathbf{i_n}}(\boldsymbol{\beta}) = \int \prod_{t=1}^{T_n} \mathbb{P}(y_{nt} = i_t|\boldsymbol{\beta}, X_{nt}) f(\boldsymbol{\beta}|\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \tag{2}$$

This probability is approximated through simulation, using:

$$\hat{P}_{\mathbf{i_n}}(\boldsymbol{\beta}) = \frac{1}{R} \sum_{r=1}^{R} \prod_{t=1}^{T_n} \mathbb{P}(y_{nt} = i_t|\boldsymbol{\beta}_r, X_{nt}) \tag{3}$$

Where the $\boldsymbol{\beta}_r$ are drawn from the distribution $f(\boldsymbol{\beta}|\boldsymbol{\theta})$. The parameter $\boldsymbol{\beta}$ is estimated through Maximum Simulated Likelihood Estimation (MLSE). We define the simulated log-likelihood of the observations $(\mathbf{y}, \mathbf{X})$ as:

$$\mathrm{LL}(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) = \sum_{n=1}^{N} \ln \hat{P}_{\mathbf{i_n}}(\boldsymbol{\beta}) \tag{4}$$

Path-Size correction

Route choice modelling often involves choice sets with overlapping routes, leading to correlated alternatives and violating the independence assumption of irrelevant alternatives (IIA) in the MNL. To account for this, each alternative receives an additional attribute measuring their overlap with other routes of the choice set. For an alternative i of choice situation t or an individual n, the Path-Size correction (Ben-Akiva & Ramming, 1998) is defined as:

$$\mathrm{PS}_{nti} = \sum_{a \in \Gamma_i} \frac{l_a}{L_i} \frac{1}{\sum_{j \in C_{nt}} \delta_{aj}}, \tag{5}$$

where Γ_i is the set of links for alternative i, C_{nt} it the set of alternatives for choice situation t, l_a is the length of link a, L_i is the length of alternative i, and δ_{aj} equals 1 if j includes link a and 0 otherwise. The utility of an alternative U_{nti} can be written as:

$$U_{nti} = V(\boldsymbol{\beta}_n, X_{nti}) + \beta_{PS} \ln \mathrm{PS}_{nti} + \epsilon_{nti} \tag{6}$$

Where β_{PS} is the Path-Size coefficient to estimate.

Model evaluation

To compare model outputs, we need metrics evaluating bias and precision of the estimations. We denote the Maximum Likelihood estimator of a true parameter vector $\boldsymbol{\beta} = (\beta_1 \dots \beta_K)^\top$ of size $(1 \times K)$ as $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1 \dots \hat{\beta}_K)^\top$.

Bias The bias of an estimator is the difference between this estimator's expected value and the true value of the estimated parameter. It is given by $\mathrm{Bias}(\hat{\boldsymbol{\beta}}) = \mathbb{E}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta}$. In the case of a multidimensional estimator, we calculate $\|\mathrm{Bias}(\hat{\boldsymbol{\beta}})\|^2 = \sum_{k=1}^{K} \mathrm{Bias}(\hat{\beta}_k)^2$

D-error The D-error is an efficiency metric commonly used in experimental designs (see Kessels et al. (2006) for an overview). It is defined as the determinant of the AVC matrix, exponentially scaled w.r.t. to the number of parameters. We calculate the AVC matrix $\boldsymbol{\Omega} = \mathbf{H}^{-1}$ as the inverse of the log-likelihood Hessian matrix at the estimates. For $i, j \in \{1, ..., K\}$ the Hessian matrix coefficients H_{ij} are given by,

$$H_{ij}(\hat{\boldsymbol{\beta}}) = \mathbb{E}\left[\frac{\partial^2 LL(\boldsymbol{\beta})}{\partial \beta_i \partial \beta_j}\right]_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$$

Then, the D-error is given by:

$$\text{D-error} = \det \boldsymbol{\Omega}^{1/K} \tag{7}$$

Minimizing the D-error is minimizing variances and covariances of the estimates. Lower D-error values indicate higher efficiency of the estimated parameter results regarding standard errors. However, it does not provide information on the bias of the estimated parameters.

3 Simulation experiment

In order to test if the Panel Mixed Logit model reproduces the population parameters, we need to conduct a simulation experiment with known true parameters and sample composition. The following simulation mimics a route choice modeling framework. For one of the population parameters (linked to Elevation gain), we will assume that the number of observations in our dataset is correlated to the parameter individual value.

Step 1 - Network and attributes: We design a small network, with $p \in \{1, ..., P\}$ Origin-Destination (OD) pairs and choice sets $\mathcal{C}_p$ of routes linking them. These routes have k attributes. For an alternative l of a pair p, we can store these attributes in a vector $X_{l,p} \in \mathbb{R}^k$. We call $\mathbf{X}_p$ the matrix storing the attributes of all the alternatives of $\mathcal{C}_p$.

The network consists of three OD pairs, A-B, B-C, and A-C (see Figure 1), each linked by 9 routes, composed of all the combinations of links going closer to the destination. Each link has

four attributes: Length (associated to parameter β_L), Elevation Gain (β_E), Bicycle Infrastructure (β_I) and Surface type (β_S). A Path-Size correction term (β_{PS}) (see Equation 5) handles the correlation between routes.

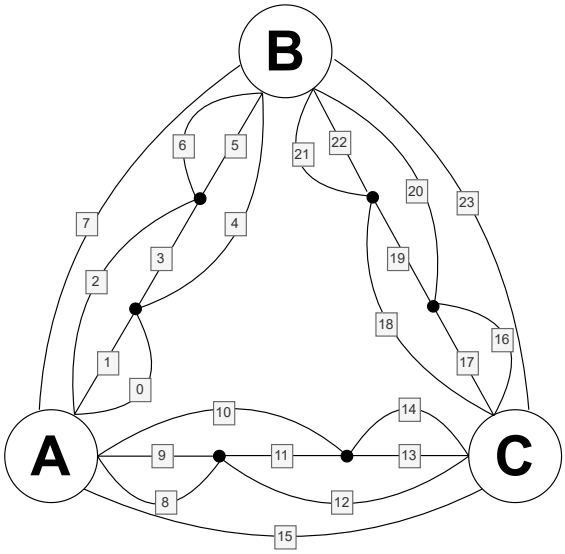

Fig. 1: Network, composed of three ODs and 24 links

Step 2 - Draw population: Assume that the general population follows a multivariate normal distribution for their parameters, i.e., the parameter vector $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, ..., \sigma_k^2)$.

In our simulation, we assume the true parameters values given in Table 1, and that $\sigma_S = \sigma_L = \sigma_{PS} = 0$.

We draw samples of $N = 100$ individuals. The distributions and histograms for the random parameters are plotted below (see Figure 2). We write $\boldsymbol{\beta}_n = \begin{pmatrix} \beta_{L,n} & \beta_{E,n} & \beta_{I,n} & \beta_{S,n} & \beta_{PS,n} \end{pmatrix}^\top$ the individual parameter drawn for individual n.

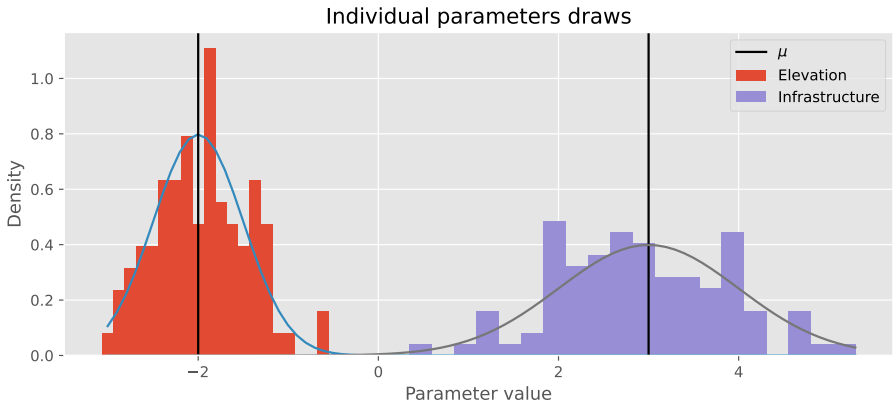

Fig. 2: Histogram of 100 draws (a population sample) for the random parameters β_E and β_I

Step 3a - Draw OD pairs: For each individual n, draw uniformly a permutation π_n of the trip purposes fulfilled on each OD. For each drawn individual, the points A, B, C can either be their "home", "work/study place" or "leisure" place. These are allocated randomly with an equal probability of $\frac{1}{3}$.

Step 3b - Draw number of observations: The number of drawn observations depends on the β_E values. This means individuals with more observations in the dataset are the least sensitive to

elevation gain. For $n \in \{1, ..., N\}$:

$$T_n = \phi(\boldsymbol{\beta}_n) = \lceil a \exp(b * \lambda(\beta_{E,n})) \rceil \tag{8}$$

$\lambda(\beta_{E,n})$ is the index of $\beta_{E,n}$ in the sequence of $\beta_{E,n}$ in increasing order. The maximum number of draws per individual has been set to $n^* = 200$, so we chose $a = \frac{n^*}{\exp(bN)}$ so that $n_1 = 1$ and $n_N = n^*$. $b = 0.075$ is a scaling constant. We draw a total of 2819 observations.

Step 4 - Draw of observations For $n \in \{1, ..., N\}$ and $t \in \{1, ..., T_n\}$ do:

1. Draw the used OD pair $p_{n,t} \sim \text{Categorical}(q_1, ..., q_P|\pi_n)$ using the probability distribution of bicycle trip purposes given by the Danish National Travel Survey (Christiansen & Baescu, 2022). $q_1 = \mathbb{P}(\text{Home-Work}) = 0.46$, $q_2 = \mathbb{P}(\text{Home-Leisure}) = 0.24$, $q_3 = \mathbb{P}(\text{Work-Leisure}) = 0.3$.
2. Draw the chosen alternative $y_{n,t} \sim \text{Categorical}(L(\boldsymbol{\beta}_n, \mathbf{X}_{p_{n,t}}))$, where

$$L(\boldsymbol{\beta}_n, \mathbf{X}_{p_{n,t}}) = \frac{\exp(\boldsymbol{\beta}_n^\top \mathbf{X}_{p_{n,t}})}{\sum_{l \in \mathcal{C}_{p_{n,t}}} \exp(\boldsymbol{\beta}_n^\top X_{l,p_{n,t}})} \tag{9}$$

Step 5: Based on these observations re-estimate a discrete choice model.

Step 4 is repeated $N_{exp} = 100$ times, to account for the randomness of the dataset creation process. The flow-chart (Figure 3) illustrates the described steps.

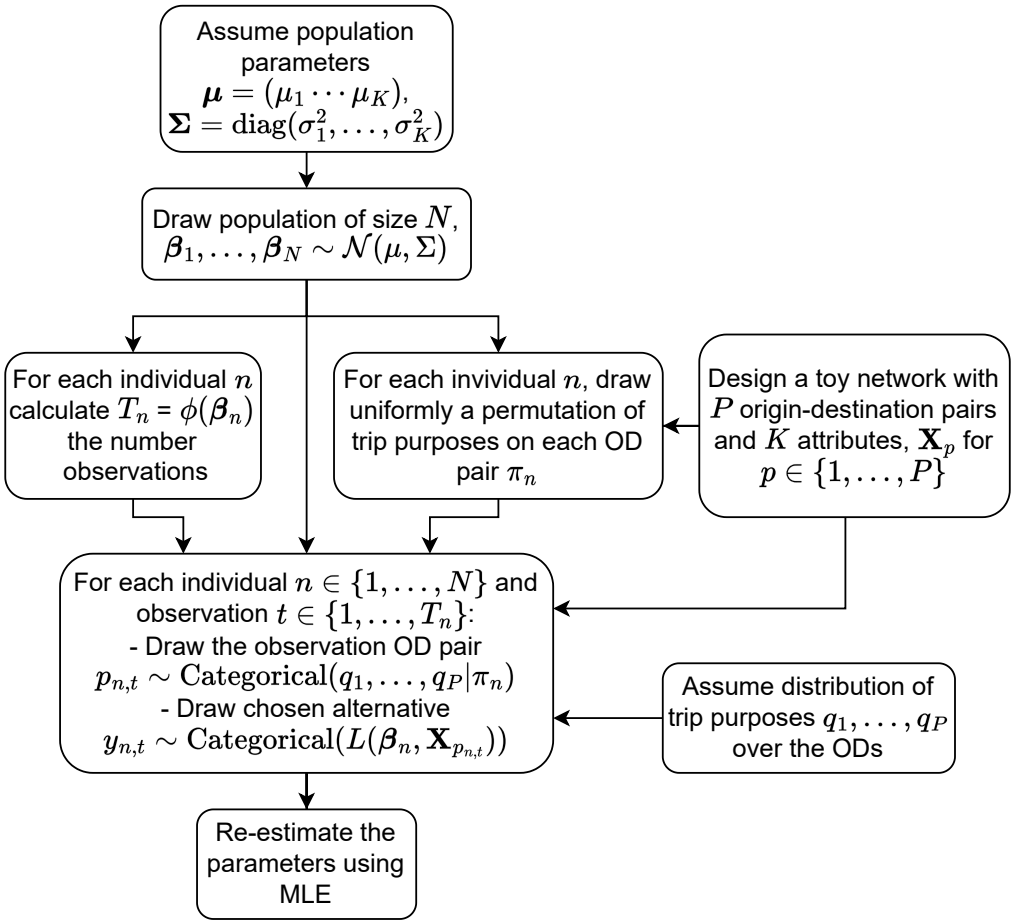

Fig. 3: Flowchart of the simulation experiments

All models are estimated using the Python library *xlogit* (Arteaga et al., 2022).

4 Results: base models

Three base models have been estimated: a Multinomial Logit model (MNL), a Mixed Logit model (MXL) and a Panel Mixed Logit Model (PMXL). The estimated parameters are summarized on Table 1. The estimated distributions are also plotted on Figure 4. We calculate the Marginal rates of substitution (onwards referred to as *tastes*) $\text{Taste}(x) = \frac{\mu_x}{\mu_L}$ as the ratio of any coefficient and the Length coefficient. This allows separating the issue of the estimation of the model scale and the derivation of people's preferences. While the model scale defines one's sensitivity to an attribute change on choice probabilities, tastes allow understanding the relative *value of distance* of each attribute.

The bias of tastes is defined as:

$$\text{Bias of tastes} = \left(\frac{\mu_E}{\mu_L} - \frac{\hat{\mu}_E}{\hat{\mu}_L}\right)^2 + \left(\frac{\mu_I}{\mu_L} - \frac{\hat{\mu}_I}{\hat{\mu}_L}\right)^2 + \left(\frac{\mu_S}{\mu_L} - \frac{\hat{\mu}_S}{\hat{\mu}_L}\right)^2$$

Tab. 1: Estimates, tastes, bias of tastes and D-error for the base models

	μ_L	μ_E	μ_I	μ_S	μ_{PS}	σ_E	σ_I	$\frac{\mu_E}{\mu_L}$	$\frac{\mu_E}{\mu_L}$	$\frac{\mu_E}{\mu_L}$	Bias of tastes	D-error
True value	-10	-2	3	1	1.5	0.5	1	0.2	-0.3	-0.1	-	-
MNL	-5.783	-0.775	1.729	0.486	1.327	-	-	0.133	-0.299	-0.0849	0.0584	0.0011
MXL	-9.938	-1.334	2.991	0.974	1.518	0.345	0.945	0.134	-0.302	-0.0982	0.0657	0.0028
PMXL	-9.964	-1.694	2.986	0.993	1.491	0.484	0.863	0.170	-0.300	-0.0997	0.0296	0.0013

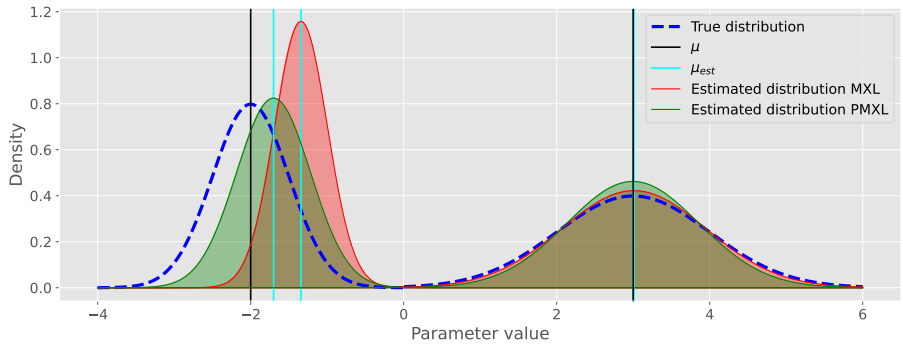

Fig. 4: Base models estimated parameters for elevation (left distribution) and infrastructure (right distribution) or distribution against true distributions

For the MNL model, the model marginal substitution rate is -13.3% for elevation gain, while the actual value is -20%. For bicycle infrastructure, the model outputs a taste of +29.9%, while the actual value is +30%. The MXL shows similar bias. Moreover, the MXL estimates a standard deviation for elevation gain that is way lower than the actual value. The PMXL shows less bias than the other base models, and estimates closer standard deviations to the true values. However it does not represent the average tastes of the individuals in the dataset. The taste for elevation gain is still shifted towards over-represented individuals (see also Figure 4). For the other parameters, however, all models show almost no bias. This is because the individual number of observations is uncorrelated with the other parameter values.

These estimations allow us to search for ways to decrease the taste bias. The developed strategies are presented in the following section.

5 Sampling and weighting strategies

We implement a number of strategies to correct the bias in tastes of the **Panel Mixed Logit model** estimated on unbalanced data. These methods use subsampling, weighting techniques, or a combination of both.

Sampling strategies

To correct the dataset unbalance, several sampling strategies have been tested out. Their purpose is to reduce the bias in the estimated parameters compared to the actual parameters of the population.

Naive subsampling: We reduce the dataset size by randomly drawing a subset of the observation. This method does not aim to reduce the bias, but is used as a benchmark for bias and efficiency.

Pruning: We remove any individual with less then k_0 observations from the dataset.

Uniform random subsampling: To keep the same number of choice experiments for each individual, we choose k as the minimum number of observations for an agent in the dataset and select k observations for each individual randomly. This method is naive, but some extensions could be added, e.g. keeping dissimilar observations.

Uniform random truncation: For many experiments, some individuals will only have one observation in the dataset ($k = 1$). Thus, another method would be to randomly choose $n > 1$ and select n observations per individual. If an individual has less than n observations in the dataset, all their observations would be selected.

Subsampling of repeated observations: Another possibility for subsampling that would keep more variability in the dataset is to have a subsampling method that keeps, for each individual, one observation per choice scenario. The unbalance in the resulting dataset can again be handled by weighting the likelihood function.

A Maximum Weighted Likelihood Estimation (MWLE)

Sampling strategies reduce bias by reducing the potential over-representation of some individuals in the dataset. However, it also leads to lower efficiency of the estimates. Another way to deal with this bias would be to modify the likelihood function so that the estimator accounts for the dataset unbalance by penalizing over-represented individuals.

We implemented a new method to weigh the likelihood function. Let $\boldsymbol{\beta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the model parameters. The goal of the weighting algorithm we describe below is that each individual contributes equally (has the same weight) in the likelihood function. For individuals $n \in \{1, ..., N\}$, we note $\mathbf{w} = (w_1 \ \dots \ w_N)$ the vector of individual weights. We note $\mathbf{LL}(\boldsymbol{\beta}) = \left(\ln \hat{P}_{\mathbf{i_1}}(\boldsymbol{\beta}) \ \dots \ \ln \hat{P}_{\mathbf{i_N}}(\boldsymbol{\beta})\right)$ the vector of individual contributions to the log-likelihood. The weighted likelihood function is thus given by:

$$\mathrm{LL}_{\mathbf{w}}(\boldsymbol{\beta}) = \mathbf{w}^\top \mathbf{LL}(\boldsymbol{\beta}) = \sum_{n=1}^{N} w_n \ln \hat{P}_{\mathbf{i_n}}(\boldsymbol{\beta})$$

Our goal is to find the vector of weights $\mathbf{w}^*$ for which each individual gives the same contribution to the weighted likelihood function evaluated at the estimated parameters. Each iteration calculates weights that are inversely proportional to the weighted likelihood contribution of an individual, using the weights of the previous iteration. This is equivalent to solving the following fixed-point problem: $\mathbf{w}^* = F(\mathbf{w}^*)$, where, for an element w_j of $\mathbf{w}$, we have:

$$F(w_j) = \frac{(w_j \ln \hat{P}_{\mathbf{i_j}}(\hat{\boldsymbol{\beta}}))^{-1}}{\frac{1}{N}\sum_{n=1}^{N}(w_n \ln \hat{P}_{\mathbf{i_n}}(\hat{\boldsymbol{\beta}}))^{-1}}; \ \hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \mathrm{LL}_{\mathbf{w}}(\beta) \tag{10}$$

$\phi = \left(\frac{1}{N}\sum_{n=1}^{N}(w_n \ln \hat{P}_{\mathbf{i_n}}(\hat{\boldsymbol{\beta}}))^{-1}\right)^{-1}$ is a normalizing constant ensuring that $\sum_{n=1}^{N} w_n = N$; which will be useful to compute the AVC matrix of the estimates.

To solve this fixed-point problem, the solution algorithm builds a sequence of weight vectors $\mathbf{w}^{(k)} = \left(w_1^{(k)} \ \dots \ w_N^{(k)}\right)$ which can be described by the pseudo-code below.

Algorithm 1: Algorithm to determine optimal weights; $\mathbf{w}^* = F(\mathbf{w}^*)$

Input: $\mathbf{X}, \mathbf{y}, n_0, \varepsilon$
Result: $\mathbf{w}^*$, the vector of optimal weights
Initialization:
$\mathbf{w}^{(1)} \leftarrow (1 \ \dots \ 1)$
$k \leftarrow 1$
while $\|F(\mathbf{w}^{(k)}) - \mathbf{w}^{(k)}\| > \varepsilon$; `// F also depends on` $\mathbf{X}, \mathbf{y}$
do
  $\hat{\mathbf{w}}^{(k+1)} \leftarrow F(\mathbf{w}^{(k)})$;
  if $k > n_0$ **then**
    $\mathbf{w}^{(k+1)} \leftarrow \lambda_k \hat{\mathbf{w}}^{(k+1)} + (1 - \lambda_k)\mathbf{w}^{(k)}$; `// Method of successive averages`
  else
    $\mathbf{w}^{(k+1)} \leftarrow \hat{\mathbf{w}}^{(k+1)}$;
  end
  $k \leftarrow k + 1$
end

The Method of Successive Averages (MSA) (Robbins & Monro, 1951), ensures the convergence of the sequence. We use $\lambda_k = \frac{1}{k-n_0}$, so that the k^{th} calculated weight vector is the arithmetic mean of the previously computed weights, i.e. $\mathbf{w}^{(k)} = \frac{1}{k-n_0} \sum_{i=n_0}^{k} \hat{\mathbf{w}}^{(i)}$. This method begins to be applied after n_0 iterations, so the first weights are not included in the average.

The simulation experiment described in section 3 is then carried out for the following setups:

1. Whole dataset, weighted
2. Random naive subsampling at 500 observations
3. Pruning individuals with less than 5 observations
4. Randomly truncated at 2, 5, 10, 20, 50 observations, unweighted and weighted
5. Randomly subsampled at the minimum number of observations per individual (equivalent to a truncation at 1 observation with this dataset)
6. Random subsampling of unique observations

Setups 2 to 5 use random subsamples of the generated datasets. The subsampling algorithms are used 100 times for the same dataset, and the results are averaged (the standard deviation is also calculated). With 100 different generated populations and observations, these setups are repeated 10000 times.

6 Results

This section compares two metrics for the different strategies: Bias of tastes and D-error, stored in Table 2. Some insights given by these tables are:

- For the models using the whole datasets, the bias in tastes has been significantly reduced by the weighting algorithm while not affecting the model efficiency (see Figure 5 for a plot of the estimated mixing distributions).
- As expected, the naive and pruning strategies, used as benchmarks, give worse results than the base model in efficiency and bias.
- The unweighted truncation give the best results in terms of *Bias of tastes* for low truncation thresholds (see Figure 6). The more the dataset is truncated, the more it is balanced, and the more the estimates are close to the population's mean. Conversely, lower truncation thresholds also lower the estimated parameters' efficiency, and increase variability between random subsamples. The bias-variance trade-off is higlighted by the green curve on Figure 8.

- The weighted truncation, while slighltly increasing the *D-error* for the same truncation threshold, shows a decrease in bias of tastes when the threshold increases at 50 (see Figure 7). The red curve on Figure 8 shows that the weighting algorithm breaks the bias-variance trade-off when a certain truncation threshold is exceeded.
- Removing repeated observations behave similarly to truncation.

Moreover, random truncation to lower thresholds gives more variability to the model output; it is important to repeat random truncation several times and average the results to lower bias.

Tab. 2: Estimates, tastes, bias of tastes and D-error for the base models

	β_L	β_E	β_I	β_S	β_{PS}	σ_E	σ_I	$\frac{\beta_E}{\beta_L}$	$\frac{\beta_I}{\beta_L}$	$\frac{\beta_S}{\beta_L}$	Bias of tastes	D-error
True value	-10	-2	3	1	1.5	0.5	1	0.2	-0.3	-0.1	-	-
MNL	-5.783	-0.775	1.729	0.486	1.327	-	-	0.133	-0.299	-0.0849	0.0584	0.0011
MXL	-9.938	-1.334	2.991	0.974	1.518	0.345	0.945	0.134	-0.302	-0.0982	0.0657	0.0028
PMXL	-9.964	-1.694	2.986	0.993	1.491	0.484	0.863	0.170	-0.300	-0.0997	0.0296	0.0013
PMXL, w	-10.307	-2.136	3.162	1.021	1.529	0.617	0.960	0.207	-0.307	-0.0991	0.0107	0.0011
Naive	-10.188	-1.538	3.154	1.002	1.530	0.365	0.998	0.150	-0.310	-0.0987	0.0395	0.0093
Pruned	-9.979	-1.509	2.954	0.993	1.495	0.357	0.838	0.151	-0.296	-0.0996	0.0490	0.0013
Trunc 2	-10.333	-1.999	3.109	1.002	1.574	0.423	1.018	0.193	-0.301	-0.0975	0.0071	0.0421
Trunc 5	-10.038	-1.849	3.067	0.982	1.522	0.379	1.003	0.184	-0.306	-0.0981	0.017	0.0143
Trunc 10	-9.989	-1.786	3.082	0.984	1.497	0.369	1.005	0.179	-0.309	-0.0987	0.023	0.0074
Trunc 20	-9.988	-1.749	3.104	0.989	1.491	0.373	1.006	0.175	-0.311	-0.0992	0.0271	0.0042
Trunc 50	-10.001	-1.727	3.094	0.992	1.496	0.402	0.975	0.173	-0.309	-0.0993	0.0288	0.0023
Trunc 2, w	-12.908	-2.600	3.990	1.257	2.073	0.473	1.199	0.201	-0.309	-0.0980	0.0106	0.0703
Trunc 5, w	-11.421	-2.252	3.612	1.113	1.743	0.451	1.126	0.197	-0.316	-0.0978	0.0175	0.0194
Trunc 10, w	-10.927	-2.165	3.501	1.072	1.636	0.465	1.097	0.198	-0.320	-0.0983	0.0214	0.0091
Trunc 20, w	-10.687	-2.161	3.438	1.051	1.589	0.495	1.073	0.202	-0.322	-0.0985	0.0228	0.0048
Trunc 50, w	-10.460	-2.163	3.311	1.033	1.554	0.545	1.015	0.207	-0.317	-0.0988	0.0189	0.0023
Obs	-9.977	-1.781	3.076	0.993	1.524	0.362	1.001	0.178	-0.309	-0.0997	0.0231	0.0075
Obs, w	-11.886	-2.338	3.735	1.169	1.778	0.464	1.148	0.196	-0.314	-0.0989	0.0156	0.0397

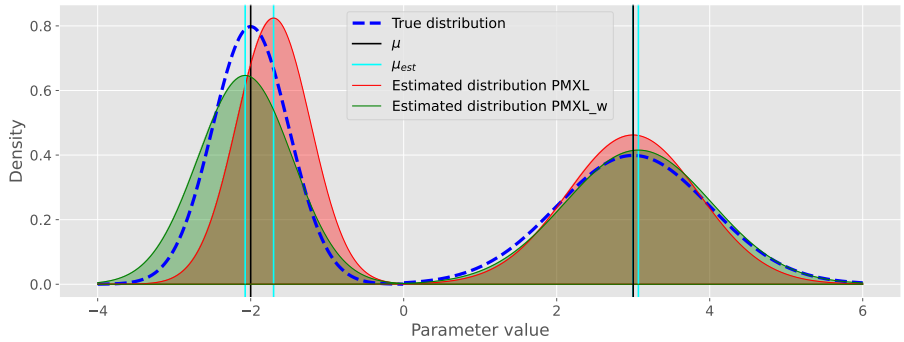

Fig. 5: Estimated distributions for elevation gain (left) and infrastructure (right), PMXL and PMXL weighted

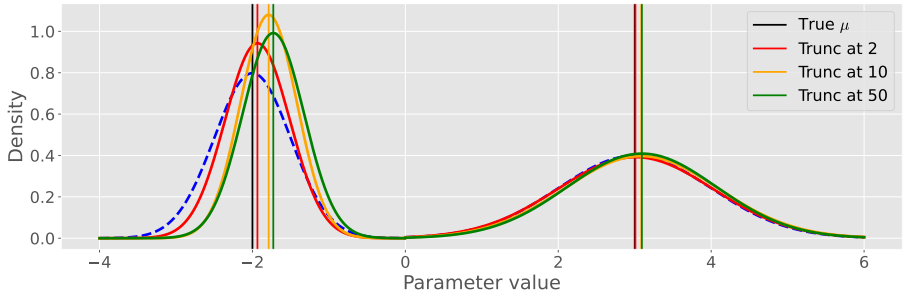

Fig. 6: Estimated distributions for elevation gain (left) and infrastructure (right), Truncated at 2, 10 and 50

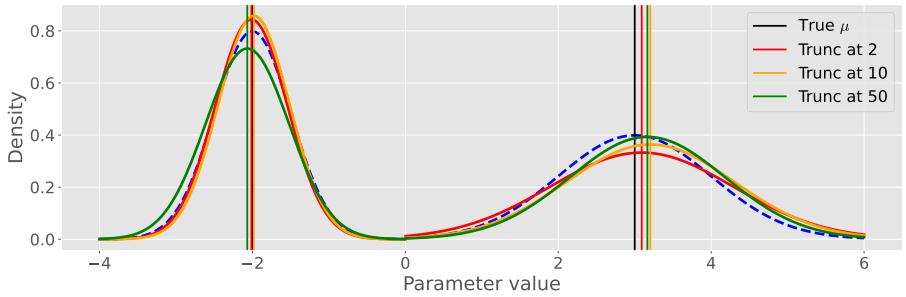

Fig. 7: Estimated distributions for elevation gain (left) and infrastructure (right), Truncated at 2, 10 and 50, weighted

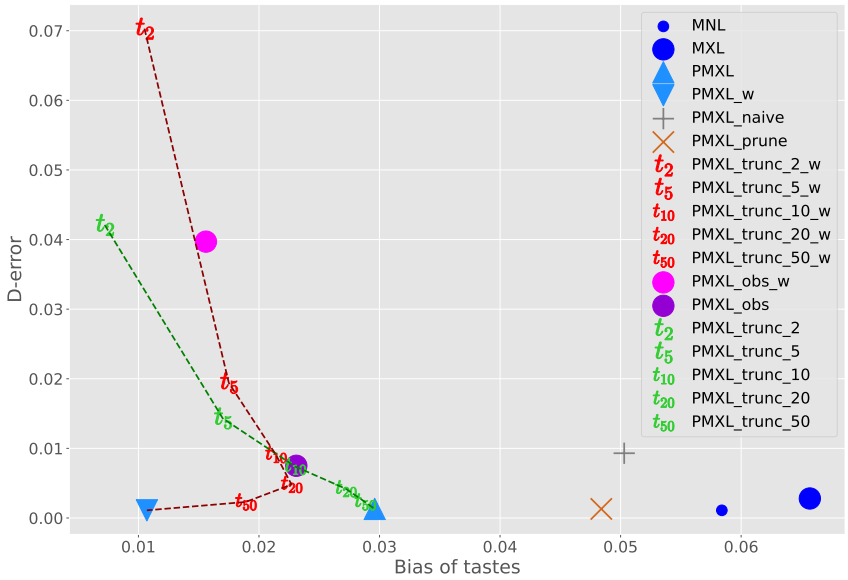

Fig. 8: Bias of tastes vs. D-error for all the different strategies

7 Conclusion and Future work

The simulation study has shown that it is possible to remove bias by applying weighting and subsampling methods. However, it also shows the bias-variance trade-off a modeller may face when choosing an optimal strategy. The newly-developed weighting algorithm breaks this trade-off by evening the contribution of each individual in the likelihood function. This allows maximum potential efficiency, while keeping a reliable explanatory model that is not biased towards over-represented individuals.

This simulation included one mixed parameter that was correlated (β_E) to the number of observation and one that was not (β_I). The results show that the models estimates are mostly unbiased for the parameter that did not correlate. As individual parameters are randomly distributed in

the population, we deduce that if the over-representation of some tastes is randomly distributed, the change in estimation may be negligible. The next step is to test these different strategies on different datasets to see how the preferences may change, i.e. how the number of observations correlates with the individual taste. Future work could also encompass a more thorough analysis on large-scale datasets, showcasing and tackling further challenges, such as bias in scale or computational burden.

Acknowledgements

We acknowledge Prof. Anders Fjendbo Jensen for his help in the early steps of the project.

References

Arteaga, C., Park, J., Beeramoole, P. B., & Paz, A. (2022). xlogit: An open-source python package for gpu-accelerated estimation of mixed logit models. *Journal of Choice Modelling*, *42*, 100339.

Ben-Akiva, M., & Ramming, S. (1998). Lecture notes: Discrete choice models of traveler behavior in networks. *Prepared for Advanced Methods for Planning and Management of Transportation Networks. Capri, Italy*, *25*.

Bliemer, M. C., & Rose, J. M. (2010). Construction of experimental designs for mixed logit models allowing for correlation across choice observations. *Transportation Research Part B: Methodological*, *44*(6), 720–734.

Cherchi, E., & Cirillo, C. (2014). Understanding variability, habit and the effect of long period activity plan in modal choices: a day to day, week to week analysis on panel data. *Transportation*, *41*(6), 1245–1262.

Cherchi, E., Cirillo, C., & de Dios Ortúzar, J. (2017). Modelling correlation patterns in mode choice models estimated on multiday travel data. *Transportation Research Part A: Policy and Practice*, *96*, 146–153.

Christiansen, H., & Baescu, O. (2022). The danish national travel survey: Annual statistical report for denmark for 2021.
doi: 10.11581/dtu:00000034

Kessels, R., Goos, P., & Vandebroek, M. (2006). A comparison of criteria to design efficient choice experiments. *Journal of Marketing Research*, *43*(3), 409–419.

Lee, K., & Sener, I. N. (2021). Strava metro data for bicycle monitoring: a literature review. *Transport reviews*, *41*(1), 27–47.

McFadden, D., et al. (1973). Conditional logit analysis of qualitative choice behavior.

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

Nelson, T., Ferster, C., Laberee, K., Fuller, D., & Winters, M. (2021). Crowdsourced data for bicycling research and practice. *Transport reviews*, *41*(1), 97–114.

Ortelli, N., de Lapparent, M., & Bierlaire, M. (2022). Faster estimation of discrete choice models via dataset reduction.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, 400–407.

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. *Transportation Research Record*(September), 1–15.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press.

van Cranenburgh, S., & Bliemer, M. C. (2019). Information theoretic-based sampling of observations. *Journal of choice modelling*, *31*, 181–197.

Yáñez, M. F., Cherchi, E., Heydecker, B. G., & de Dios Ortúzar, J. (2011). On the Treatment of Repeated Observations in Panel Data: Efficiency of Mixed Logit Parameter Estimates. *Networks and Spatial Economics*, *11*(3), 393–418. doi: 10.1007/s11067-010-9143-6