Revealing and reducing bias when modelling choice behaviour on imbalanced panel datasets

Łukawska, Mirosława; Cazor, Laurent; Paulsen, Mads; Rasmussen, Thomas Kjær; Nielsen, Otto Anker

Published in:
Journal of Choice Modelling

Link to article, DOI:
10.1016/j.jocm.2024.100471

Publication date:
2024

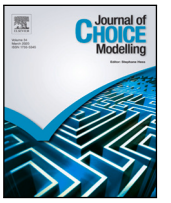
Document Version
Publisher's PDF, also known as Version of record

Journal of Choice Modelling 50 (2024) 100471

Contents lists available at ScienceDirect

Journal of Choice Modelling

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



Revealing and reducing bias when modelling choice behaviour on imbalanced panel datasets

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

Technical University of Denmark, Department of Technology, Management and Economics, Bygningstorvet 116b, 2800 Kgs. Lyngby, Denmark

ARTICLE INFO

Keywords:
Imbalanced panel
Panel mixed multinomial logit model
Subsampling
Weighting
Bias-efficiency trade-off

ABSTRACT

The emergence of modern tools and technologies gives a unique opportunity to collect large amounts of data for understanding behaviour. However, the generated datasets are often imbalanced, as individuals might contribute to the datasets at different frequencies and periods. Models based on these datasets are challenging to estimate, and the results are not straightforward to interpret without considering the sample structure. This study investigates the issue of handling imbalanced panel datasets for modelling individual behaviour. It first conducts a simulation experiment to study to which degree mixed logit models with and without panel reproduce the population preferences when using imbalanced data. It then investigates how the application of bias reduction strategies, such as subsampling and likelihood weighting, influences model results and finds that combining these techniques helps to find an optimal trade-off between bias and variance of the estimates. Considering the conclusions from the simulation study, a large-scale case study estimates bicycle route choice models with different correction strategies. These strategies are compared in terms of efficiency, weighted fit measures, and computational burden to provide recommendations that fit the modelling purpose. We find that the weighted panel mixed multinomial logit model, estimated on the entire dataset, performs best in terms of minimising the bias-efficiency trade-off in the estimates. Finally, we propose a strategy that ensures equal contribution of each individual to the estimation results, regardless of their representation in the sample, while reducing the computational burden related to estimating models on large datasets.

1. Introduction

An increasing number of large crowdsourced datasets for analysing behaviour provides researchers with new opportunities and overcomes some of the existing limitations related to stated preference data or small sample sizes (Nelson et al., 2021; Lee and Sener, 2021). However, due to their crowdsourced opt-in nature, such panel datasets may suffer from having a large proportion of the data being collected by only a few very active individuals on different time periods. If the modelling purpose is to retrieve the average preferences of individuals in a dataset, model estimations using such imbalanced panel datasets risk being biased towards the preferences of individuals with many observations, which might be undesirable for, e.g., policy implications or forecasting purposes. When it comes to modelling behaviour with such data, researchers have the liberty to influence which part of the dataset to use and how to use it. The goal of this study is to propose correction strategies, such as weighting and subsampling methods, that allow reducing the bias inherent in models based on imbalanced panel datasets.

* Corresponding author.
E-mail address: mirlu@dtu.dk (M. Łukawska).

https://doi.org/10.1016/j.jocm.2024.100471
Received 31 May 2023; Received in revised form 28 November 2023; Accepted 14 January 2024
Available online 18 January 2024

The issue of repeated observations per individual in the panel setup for the mixed logit (MXL) models has been addressed in the literature in the context of stated preference (SP) data. Bliemer and Rose (2010) recognised the advantages of the 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 and Cirillo, 2014) or correlation patterns (Cherchi et al., 2017).

Yáñez et al. (2011) investigated how repeated observations per individual influenced the statistical properties and the outcome of the model. They found that the greatest improvement to the model in terms of the measure of fit can be attributed to the introduction of panel correlation. Furthermore, including multiple identical observations (either in general or per individual) should not influence the efficiency of the estimated parameters and does not contribute to the improved capability to retrieve the true parameters. Two recent studies (van Cranenburgh and Bliemer, 2019; Ortelli et al., 2022) recognised the challenges resulting from estimating models based on rapidly emerging new big data sources. They proposed strategies for reduction of the dataset size by maximising model efficiency (van Cranenburgh and Bliemer, 2019) or taking multiple criteria into account, such as efficiency, estimation bias, out-of-sample performance, computational time, and value of time for relevant parameters (Ortelli et al., 2022). They proposed to optimise the model setup for the MXL model, relying on a much simpler multinomial logit (MNL) model. However, none of the studies considered the panel setup and its influence on the sample composition. A subsampling strategy for the panel MXL model, that preserves the inherent inter-respondent heterogeneity in the best possible way, was also outside the scope of these studies.

Recently, several new software packages for estimating choice models have been released, e.g. Arteaga et al. (2022), Molloy et al. (2021). These packages are extremely useful in estimating MXL models on large datasets, thanks to introducing, e.g. mini-batches and GPU acceleration (Arteaga et al., 2022), or multiprocessing (Bierlaire, 2020; Molloy et al., 2021). However, these implementations do not take into account that an extremely large number of observations per individual can lead to intractable numerical problems when calculating the probabilities for each individual while estimating a panel MXL (PMXL) model. We found that estimating a full PMXL model on a large dataset with several hundred observations for some of the individuals is not possible with these packages.

Therefore, the potential of the large-scale crowdsourced datasets is not entirely released, as models based on these datasets are challenging to estimate and the results are not straightforward to interpret without considering the sample structure. The behaviour of a large number of individuals is inherently very heterogeneous (e.g. Krueger et al., 2021), and hence, there is a need for tools to handle and reveal this heterogeneity in the models. Moreover, different strategies should be applied for either sampling the data or defining the model specification, depending on the modelling purpose and if the interest is in the estimation of preferences on the individual level or on the observational level. Finally, a thorough understanding of the processing tools as well as the availability of the computational resources are necessary to handle a dataset of such size.

In this paper, we make several contributions which address the above-mentioned challenges. Firstly, with a simulation experiment, we demonstrate that the bias-efficiency trade-off is important to consider when sampling from data with an imbalance in terms of the number of observations per individual. Secondly, we propose and explore various subsampling and weighting strategies to address this trade-off and provide recommendations for estimating equitable models, describing the behaviour of the individuals in the available dataset. Thirdly, this study tackles the issue of estimating a PMXL model on a dataset with a very large number of observations per individual and describes the required mathematical reformulations. Finally, we estimate a PMXL bicycle route choice model, applying the strategy that equalises the contribution of each individual to the model results. For this purpose, we utilise a crowdsourced dataset of observed route choices of cyclists to showcase and evaluate the large-scale performance of the strategies from the simulation study. This large-scale example suits the purpose of the study, as the used dataset indicates high dynamics in the opt-in participation of the users and the route choice preferences are expected to vary between individuals.

It should be mentioned that, while the strategies proposed and employed in this paper reduce the bias caused by the imbalance in the number of observations per individual, they do not address the issue of potential lack of representativity of the sample population from the data w.r.t. the general population. When defining the modelling framework, relevant aspects of sample stratification, sample bias, or further transferability of the results to the general population should still be considered. Readers interested in the topic of sample design and balancing are further referred to seminal publications such as Manski and Lerman (1977), Manski and McFadden (1981).

The remainder of this paper is structured as follows. Section 2 describes the structure of the PMXL model, and introduces the correction strategies applied in this study. Section 3 describes a simulation experiment, conducted to understand the bias-efficiency trade-off present in the model estimates. Section 4 includes a large-scale case study, estimating bicycle route choice models with different correction strategies. Section 5 formulates several recommendations to reduce bias when modelling choice behaviour on imbalanced panel datasets and Section 6 concludes the paper. Finally, the paper includes four appendices with mathematical derivations and reformulations, additional information for the simulation experiment, and detailed model estimation results.

2. Method

This section provides a relevant theoretical background for the concepts applied in this paper. Section 2.1 describes the MXL model and its extension – the PMXL model – where panel setup is included in the estimation. We also describe the maximum likelihood estimation (MLE) procedure for the PMXL model in Section 2.2. Furthermore, in Section 2.3, we introduce the definitions of model bias and model efficiency - evaluation criteria used in the case studies. Finally, Section 2.4 proposes alternative weighting and sampling strategies to manipulate the estimation dataset, which will be further applied to investigate the behaviour of bias and efficiency values.

2.1. Panel mixed multinomial logit model

The mixed multinomial logit (MXL) model (McFadden and Train, 2000) builds on the traditional multinomial logit (MNL) model (McFadden et al., 1973). If the utility expression U_{it}, associated with alternative i in choice situation t, is decomposed into the systematic (observed) part V_{it} and the random part ϵ_{it}, i.e. $U_{it} = V_{it} + \epsilon_{it}$, the logit model is obtained by assuming that the random part ϵ_{it} is i.i.d. and follows the type-I extreme value distribution (Train, 2009). In the MNL model, the probability P_{it}^{MNL} of choosing alternative i in choice situation t is defined as

$$P_{it}^{\text{MNL}} = P\left(i|C_t\right) = \frac{\exp\left(V_{it}\right)}{\sum_{i \in C_t} \exp\left(V_{it}\right)}, \tag{1}$$

where C_t denotes the set of all alternatives for choice situation t. We further assume, without loss of generality, that the systematic part of utility is linear-in-parameters, i.e. $V_{it} = \boldsymbol{\beta} X_{it}$.

The MXL model accounts for the unobserved inter-heterogeneity by assuming that the parameters follow a certain distribution. This distribution is commonly assumed to be a parametric distribution, and its shape is chosen a priori in the model specification. We further assume that the set of parameters β consists of *fixed* parameters α, that are constant across all observations, and *random* parameters ζ, drawn from a certain *mixing* distribution with parameters Ω and a probability distribution function $f(\zeta|\Omega)$. Then, for the MXL model, we can define the probability $P_{it}(\Omega)$ of choosing the alternative i in the choice situation t as an integral of standard logit probabilities over a density of parameters $f(\zeta|\Omega)$ as

$$P_{it} = \int_{\zeta} P_{it}^{\text{MNL}}(\alpha, \zeta)\, f(\zeta|\Omega)\, d\zeta. \tag{2}$$

In the panel setup, we have additional information about observations belonging to specific individuals. In this case, we need to consider the correlation between observations from the same individual. This information can be reflected in the formulation for mixed logit model with panel (PMXL) (Revelt and Train, 1998), assuming inter-respondent heterogeneity (and intra-respondent homogeneity, a condition later relaxed by Hess and Train (2011)). In such case, the probability of the individual n making a series of T_n choices $\mathbf{i}_n = \{i_{n1}, i_{n2}, \ldots, i_{nT_n}\}$ can be expressed as

$$P_n = \int_{\zeta} \prod_{i \in \mathbf{i}_n} P_i^{\text{MNL}}\left(\alpha, \zeta_n\right) f(\zeta|\Omega)\, d\zeta. \tag{3}$$

The random parameters ζ_n are individual-specific, drawn from the mixing distribution with a probability distribution function $f(\zeta|\Omega)$. T_n refers to the number of choice observations for individual n.

Path-size logit. Route choice modelling often involves choice sets with overlapping routes (routes sharing links of the underlying network), leading to correlated alternatives and violating the MNL assumption of the independence of irrelevant alternatives (IIA). The path-size logit (PSL) model (Ben-Akiva and Ramming, 1998) builds on the MNL model, where the utility expression V_{it} is extended by adding a (logarithm of) path-size coefficient PS_{it} (and the respective parameter β_{PS}), correcting for the overlap between alternatives (Ben-Akiva and Bierlaire, 1999). The path-size term for the alternative i in choice situation t is defined as

$$\text{PS}_{it} = \sum_{a \in \Gamma_i} \frac{l_a}{L_i} \frac{1}{\sum_{j \in C_t} \delta_{aj}}, \tag{4}$$

where Γ_i is the set of links for alternative i, C_t is 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.

2.2. Maximum likelihood estimation

The parameter estimates for the MXL model with panel setup are found by maximising the following product of the probabilities of the series of chosen alternatives $P_{\mathbf{i}_n}$ (Eq. (3)) for all N individuals in the sample:

$$\text{L}(\alpha, \zeta) = \prod_{n=1}^{N} P_n\left(\alpha, \zeta_n\right). \tag{5}$$

Computing the term $\text{L}(\alpha, \zeta)$ poses several challenges. Firstly, the integral over the parameter density in Eq. (3) is intractable. Secondly, the computation of likelihood as the product of probabilities can suffer from numerical underflow issues very fast. The latter problem is even more pronounced for the PMXL model than for the simple MXL model because of the double product (across all choice observations per individual T_n and all individuals N).

A solution to the intractability problem relies on approximating the integral over a density with an average across draws (Train, 2009). The simulated probability $\hat{P}_n$ is approximated by taking the average of the probabilities for each of the random draws $r \in R$ from the mixing distribution of parameters ζ:

$$\hat{P}_n(\alpha, \zeta) = \frac{1}{R} \sum_{r=1}^{R} \prod_{i \in \mathbf{i}_n} P_i^{\text{MNL}}\left(\alpha, \zeta_{n,r}\right). \tag{6}$$

The computational issue is usually tackled by changing the objective function of the optimisation algorithm from the likelihood to its natural logarithm (log-likelihood, see Eq. (7)), resulting in a change from a product of probabilities to a sum:

$$\mathrm{LL}(\alpha,\zeta) = \ln\left(\prod_{n=1}^{N} \hat{P}_n\left(\alpha,\zeta_n\right)\right) = \sum_{n=1}^{N} \ln \hat{P}_n\left(\alpha,\zeta_n\right). \tag{7}$$

However, it turns out that in some situations this approach is not enough to ensure computational stability of the estimation procedure with panel setup. In particular, if a sample includes a very high number of choice observations T_n for some individuals, the simulated probability from Eq. (6) can suffer from underflow issues for these individuals, making a direct computation of the log-likelihood expression impossible. We encounter this issue with several available packages for estimating discrete choice models (Arteaga et al., 2022; Bierlaire, 2020; Molloy et al., 2021). In Appendix A we propose a reformulation to an equivalent expression for the log-likelihood with a higher numerical stability. The reformulation is applied in our case study using a large-scale dataset in Section 4.

2.3. Model evaluation criteria

To evaluate and compare model results in the simulation study, we employ bias and efficiency metrics. 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 $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1 \dots \hat{\beta}_K)^\top$. The *bias* of an estimator is defined as the difference between its expected value and the true value of the estimated parameter: $\mathrm{Bias}(\hat{\boldsymbol{\beta}}) = \mathbb{E}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta}$. In the case of a multidimensional estimator, we calculate $\|\mathrm{Bias}(\hat{\boldsymbol{\beta}})\|^2 = \sum_{k=1}^{K} \mathrm{Bias}(\hat{\beta}_k)^2$.

As a second criterion, we seek to minimise the standard errors of the estimates, and we measure the overall statistical efficiency of the model by means of the D-optimality ((Kessels et al., 2006), as recommended by Yáñez et al. (2011)). The *D-error* is defined as the determinant of the inverse of the log-likelihood Hessian matrix at the estimates $\mathbf{H}$, exponentially scaled w.r.t. to the number of parameters. If by H_{ij} we denote the coefficients of the Hessian matrix:

$$H_{ij}(\hat{\boldsymbol{\beta}}) = \mathbb{E}\left[\frac{\partial^2 LL(\boldsymbol{\beta})}{\partial\beta_i \partial\beta_j}\right]_{\beta=\hat{\beta}}, \tag{8}$$

for $i, j \in \{1, \dots, K\}$, then the definition of the D-error can be formalised as:

$$\text{D-error} = \det \boldsymbol{\Omega}^{1/K}, \tag{9}$$

where $\boldsymbol{\Omega} = \mathbf{H}^{-1}$ is the variance–covariance matrix. D-error is insensitive to parameter scaling, and a sheer minimisation of the D-error ignores the variety in the underlying true preferences and leads to keeping only similar observations in the sample (Ortelli et al., 2022). Therefore, in the simulation study in Section 3 we apply both measures complementarily and analyse the trade-off between them.

2.4. Correction strategies

Manipulating the structure of a dataset for the model estimation is expected to alter the above-mentioned criteria. We aim to minimise bias in tastes and compensate for the influence of overrepresented individuals in the sample while retaining high model efficiency. The strategies that we explore to tackle the bias-efficiency trade-off are *subsampling*, where only a subset of the dataset is used for estimation, *weighting*, where the importance of certain observations or individuals can be manipulated, or a combination of both.

2.4.1. Subsampling

Subsampling of observations reduces the sample size, making some observations entirely non-existent to the model. Subsampling can be performed either on the individual level, where all individuals are purposely retained in the sample, or at the observational level, ignoring the panel effect in the data when sampling from the observations. The approaches have different influences on the model. While the first one only restricts the impact of an individual on model estimates, the second may entirely remove some individuals and their choice preferences from the analysis. As in this paper, we consider models explaining the preferences of all individuals in a given sample, we mostly focus on subsampling strategies that keep all individuals in the sample. However, we also define a few benchmark strategies, such as *naive* subsampling or *pruning*, operating on the observational level. All subsampling methods applied in this study are summarised in Table 1.

2.4.2. Weighting

Sampling strategies reduce bias by diminishing the potential over-representation of some individuals in the sample. However, these strategies also ignore some preferences setups and lead to lower efficiency of the estimates, due to the reduction in the sample size.

Another way to tackle the bias issue is to restrict the influence of some observations or individuals on the model while retaining all observations in the sample. When certain groups of individuals are over- or underrepresented in the sample (compared to the general population and within the sample), weights can be assigned to diminish or increase the contribution of those individuals to the model. The weights are then used in the estimation procedure to redefine the objective function, leading to model results which better reflect the preferences of the individuals in the dataset.

Table 1
Subsampling strategies. M denotes any (but fixed) positive integer. All random subsampling methods are without replacement.

Subsampling strategy	Description
Removing individuals:	
Naive subsampling	Random draw of a subset of observations
Pruning (at M)	Removal of all individuals with less than M observations
Keeping all individuals:	
Uniform random subsampling	For all N individuals: Random draw of $\min_{n\in\{1,\dots,N\}} T_n$ observations
Uniform random truncation (at M)	For each individual n: Random draw of $\min_{n\in\{1,\dots,N\}}(M, T_n)$ observations
Truncation of repeated observations	Random draw of one observation per unique choice scenario

The goal of this study is to define strategies for PMXL models that equally account for the preferences of all individuals in the sample. Therefore, we are interested in a weighting strategy that equalises the contribution of each individual to the model results. Although it would seem intuitive to weight by the inverse of the number of observations $\frac{1}{T_n}$, this approach leads to overcompensation and in turn favours less represented individuals, as illustrated in Appendix B. In the next paragraph, we propose heuristics to find the optimal set of individual weights to ensure an equal contribution of each individual to the log-likelihood objective function from Eq. (7).

Maximum weighted likelihood estimation (MWLE). We propose a method to find the vector of weights $\mathbf{w}^*$, such that all elements of the sum from Eq. (10) are equal (corresponding to an equal contribution of each individual). By $\mathbf{w} = \begin{pmatrix} w_1 & \dots & w_N \end{pmatrix}$, we denote the vector of individual weights, i.e. w_n is a weight assigned to individual n, $n \in \{1, \dots, N\}$. For the set of parameters $\boldsymbol{\beta} = (\alpha, \zeta)$, the weighted log-likelihood function is thus given by:

$$\mathrm{LL}_{\mathbf{w}}(\boldsymbol{\beta}) = \mathbf{w}^{\top}\mathbf{LL}(\boldsymbol{\beta}) = \sum_{n=1}^{N} w_n \ln \hat{P}_n(\boldsymbol{\beta}), \tag{10}$$

where $\mathbf{LL}(\boldsymbol{\beta}) = \begin{pmatrix} \ln \hat{P}_1(\boldsymbol{\beta}) & \dots & \ln \hat{P}_N(\boldsymbol{\beta}) \end{pmatrix}$ denotes the vector of individual contributions.

The proposed procedure is iterative and calculates weights that are inversely proportional to the weighted likelihood contribution of an individual at each iteration, 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:

$$F(w_j) = \frac{\left(\ln \hat{P}_j(\hat{\boldsymbol{\beta}})\right)^{-1}}{\frac{1}{N}\sum_{n=1}^{N}\left(\ln \hat{P}_n(\hat{\boldsymbol{\beta}})\right)^{-1}} \tag{11}$$

with

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \mathrm{LL}_{\mathbf{w}}(\boldsymbol{\beta}). \tag{12}$$

$\left(\frac{1}{N}\sum_{n=1}^{N}(\ln \hat{P}_n(\hat{\boldsymbol{\beta}}))^{-1}\right)^{-1}$ is a normalising constant ensuring that $\sum_{n=1}^{N} w_n = N$ and allowing for a direct comparison of variance–covariance matrices (and the resulting D-errors) between weighted and unweighted models.

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 pseudocode below.

The Method of Successive Averages (MSA, Robbins and 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 is applied after n_0 iterations, causing the first weights not to be included in the average. The procedure terminates once the difference in weights between two consecutive iterations falls below a predefined threshold τ.

In a simulation study in Section 3 and a real-life large-scale case study in Section 4, we apply various combinations of the above-described sampling and weighting strategies, and evaluate and compare the model results.

3. Simulation experiment

In this section, we conduct a Monte-Carlo simulation experiment with known true parameters and sample composition, to test how different models reproduce the population parameters and perform in terms of bias and efficiency. The simulation mimics a route choice modelling framework. Importantly, we set up the experiment, such that for one of the variables the parameters for each individual are correlated with the corresponding number of observations for each individual. Correlation patterns of this type have previously been witnessed in the literature, for instance by Hood et al. (2011), Alizadeh et al. (2019), Łukawska et al. (2023), who all estimated route choice models where some parameter estimates differ across groups with different trip frequencies.

In the following, we describe in detail how the dataset is generated and present the results for base models (without any interference in the sample), and for models where we apply the strategies from Section 2.4. All models are estimated using the Python library *xlogit* (Arteaga et al., 2022), with functions from objects *MultinomialLogit* and *MixedLogit*.

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

Input: $\mathbf{X}, \mathbf{y}, n_0, \tau$
Result: $\mathbf{w}^*$, the vector of optimal weights
Initialisation:
$\mathbf{w}^{(1)} \leftarrow (1 \ \dots \ 1)$;
$k \leftarrow 1$;
while $\|F(\mathbf{w}^{(k)}) - \mathbf{w}^{(k)}\| > \tau$; // 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

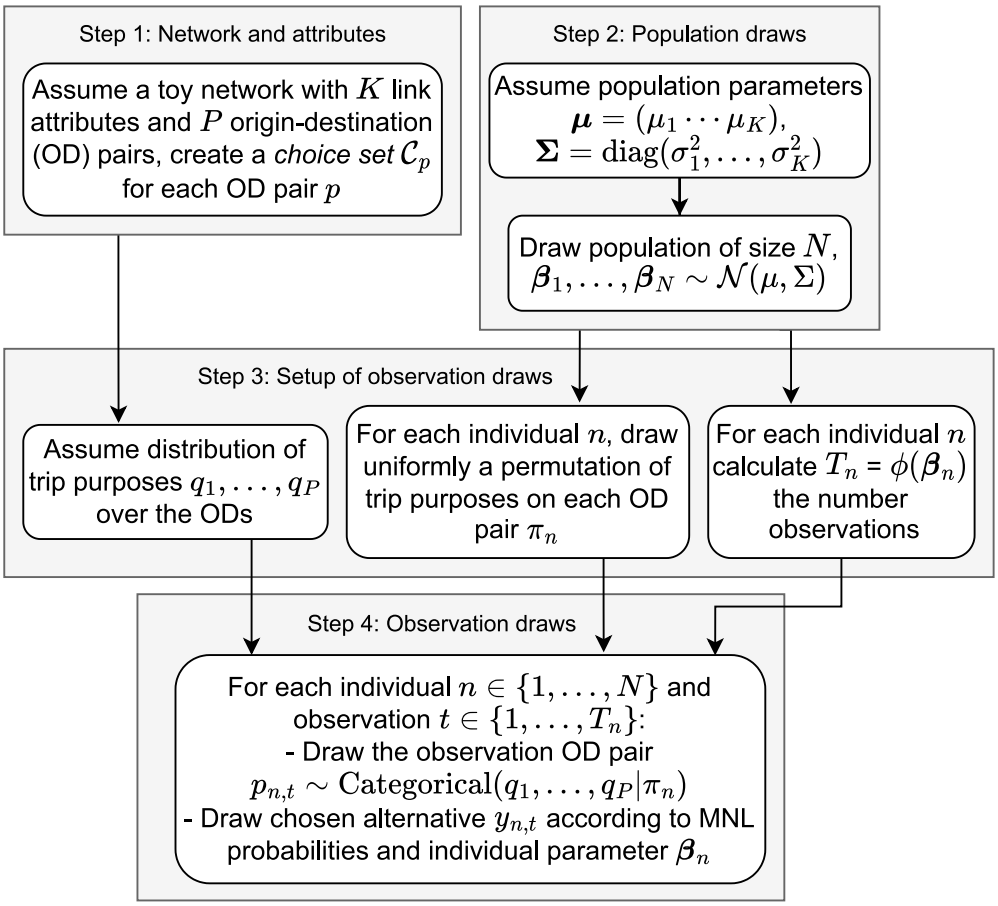


Fig. 1. Flowchart of the data generation process for the simulation experiments. The process is repeated 100 times.

3.1. Data generation

The data generation process consists of four main steps. They are illustrated in the flowchart in Fig. 1, and we describe all steps of the data generation process in the following in turn.

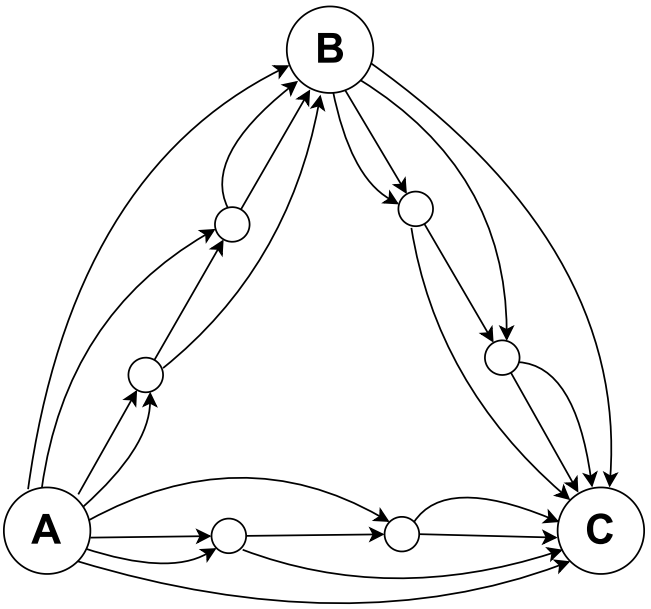


Fig. 2. Network composed of three OD pairs and 24 links. The link attributes are given in Appendix C.

Table 2
True values for the population parameters.

	μ_L	μ_E	μ_I	μ_S	μ_{PS}	σ_E	σ_I	$\frac{\mu_E}{\mu_L}$	$\frac{\mu_I}{\mu_L}$	$\frac{\mu_S}{\mu_L}$
True value	−10	−2	3	1	1.5	0.5	1	0.2	−0.3	−0.1

Step 1: Network and attributes. We design a small network with $p \in \{1, \dots, P\}$ origin–destination (OD) pairs and choice sets C_p consisting of routes linking each pair. Each of the alternatives l for an OD pair p has a set of k attributes $X_{l,p} \in \mathbb{R}^k$, and by $\mathbf{X}_p$ we denote the matrix of the attributes for all the alternatives of C_p.

The network consists of three OD pairs, A → B, B → C, and A → C (see Fig. 2), each linked by nine directed routes. Each link has four attributes: length (associated with the parameter β_L), elevation gain (β_E), bicycle infrastructure (β_I) and non-smooth surface (β_S). The parameter for the path-size correction term (β_{PS}, see Eq. (4)) captures the correlation between routes.

Step 2: Population. We assume that the preference parameters of the general population follow a multivariate normal distribution, i.e. the parameter vector $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \dots, \sigma_k^2)$. The true mean and standard deviation values are given in Table 2. Moreover, we assume a point distribution for the following parameters: β_L, β_S, and β_{PS}, i.e. $\sigma_S = \sigma_L = \sigma_{PS} = 0$.

We draw a sample of $N = 100$ individuals. The histograms of the distributions for the random parameters are displayed in Fig. 3. We denote the individual parameter drawn for the individual n as $\boldsymbol{\beta}_n = \begin{pmatrix} \beta_{L,n} & \beta_{E,n} & \beta_{I,n} & \beta_{S,n} & \beta_{\text{PS},n} \end{pmatrix}^\top$.

Step 3a: OD pairs. For each individual n, we draw uniformly a permutation π_n of the trip purposes of each OD-pair. For each individual, points A, B, and 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: Number of observations. We assume that the number of observations per individual is correlated to the individual value for the β_E parameter. With this assumption, we mimic a situation where the individuals with more observations in the dataset – who thus cycle more often – are less sensitive to the elevation gain on the route. The number of observations T_n for the individual n, $n \in \{1, \dots, N\}$ is defined as:

$$T_n = \phi(\boldsymbol{\beta}_n) = \lceil a \exp(b * \lambda(\beta_{E,n})) \rceil, \tag{13}$$

where $\lambda(\beta_{E,n})$ denotes the rank of $\beta_{E,n}$ in the sequence of $\beta_{E,n}$ in increasing order. The maximum number of draws per individual is set to $n^* = 200$, and we calibrate the value of $a = \frac{n^*}{\exp(bN)}$ so that $n_1 = 1$ and $n_N = n^*$. $b = 0.075$ is a scaling constant. This leads to a total of 2819 observations (see Fig. 4 for the distribution of observations across individuals).

Step 4: Observations. For each individual $n \in \{1, \dots, N\}$ and observation $t \in \{1, \dots, T_n\}$ we apply two steps:

(1) Firstly, we draw an OD pair $p_{n,t} \sim \text{Categorical}(q_1, \dots, q_P|\pi_n)$ with the probability distribution of bicycle trip purposes given by the Danish National Travel Survey (Christiansen and Baescu, 2022): $q_1 = P(\text{Home} \to \text{Work}) = 0.46$, $q_2 = P(\text{Home} \to \text{Leisure}) = 0.24$, $q_3 = P(\text{Work} \to \text{Leisure}) = 0.3$.

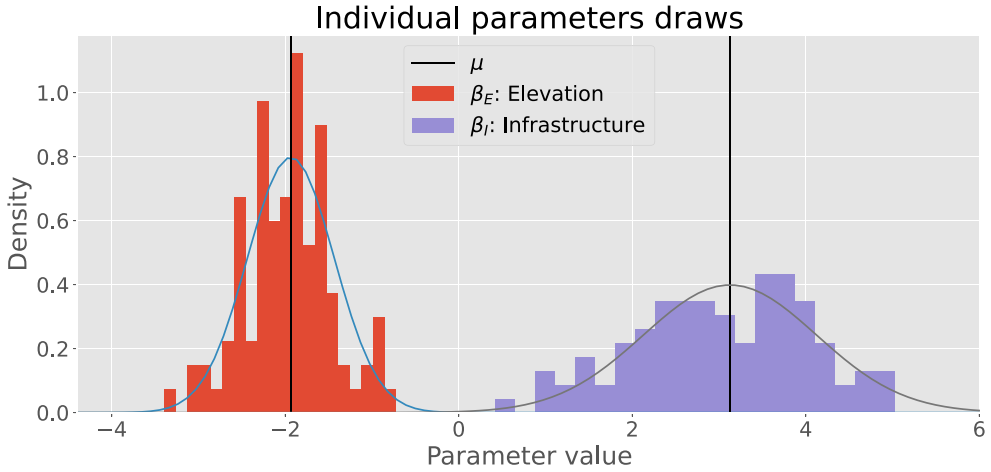


Fig. 3. Histogram of 100 draws (a population sample) for the random parameters β_E and β_I.

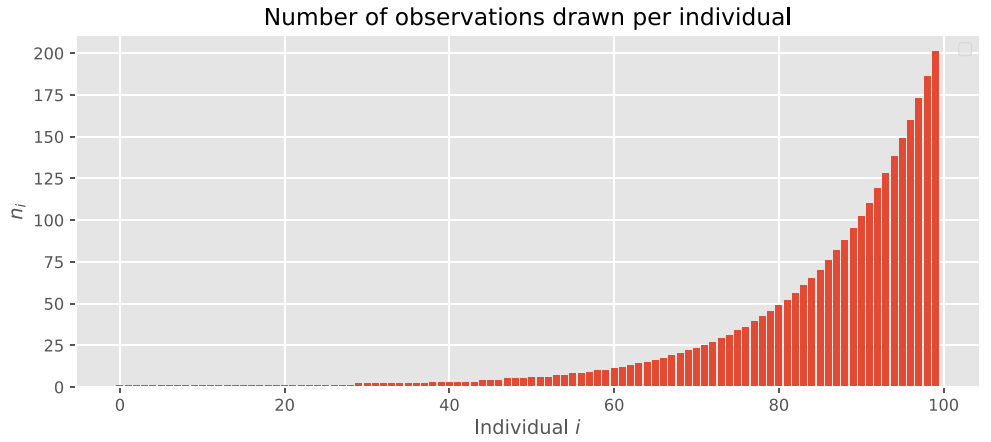


Fig. 4. Number of draws n_i for each individual i.

(2) Secondly, we draw the chosen alternative $y_{n,t}$ based on MNL probabilities (Eq. (1)) with individual parameters $\boldsymbol{\beta}_n$.

We repeat Steps 2–4 a total of 100 times, to account for the randomness of the dataset creation process.

3.2. Model estimation

On each of the samples from Section 3.1, we estimate three base models: a MNL model, a MXL model, and a PMXL model. We calculate the marginal rates of substitution (MRS) for the length attribute, onwards referred to as *tastes*, as ratios between the marginal utilities of the network attributes and the marginal utility of length, i.e. $\text{Taste}(x) = \frac{\mu_x}{\mu_L}$. This allows for separating the issue of the estimation of the model scale and the derivation of the actual preferences. While the model scale defines the individual sensitivity to an attribute change on choice probabilities, tastes allow understanding the relative value-of-distance of the attribute.

For models in this simulation experiment, the bias in tastes is defined as:

$$\text{Bias}_{\text{taste}} = \sum_{\mu \in \{\mu_E, \mu_I, \mu_S\}} \left(\frac{\mu}{\mu_L} - \frac{\hat{\mu}}{\hat{\mu}_L} \right)^2, \tag{14}$$

where μ and $\hat{\mu}$ denote the true parameters from Table 2 and their estimators, respectively.

3.3. Results

This section presents the results of the simulation study. We apply the correction strategies from Section 2.4 to evaluate the behaviour of two indicators jointly: bias of tastes and efficiency. The models with the correction strategies are benchmarked against simpler base models without any inference to the estimation dataset.

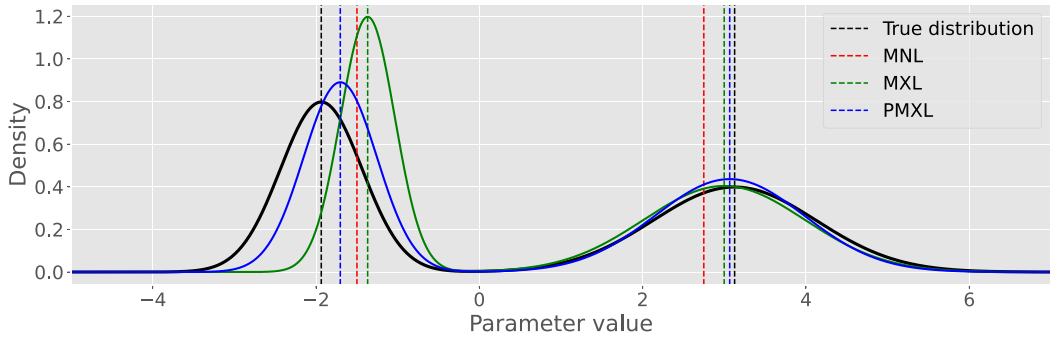


Fig. 5. Distributions for the elevation gain (left curve) and the bicycle infrastructure (right curve) attributes, estimated in the base models.

Table 3
Estimates, tastes, bias of tastes and D-error of the base models.

	μ_L	μ_E	μ_I	μ_S	μ_{PS}	σ_E	σ_I	$\frac{\mu_E}{\mu_L}$	$\frac{\mu_I}{\mu_L}$	$\frac{\mu_S}{\mu_L}$	$\text{Bias}_{\text{taste}}$	D-error
True value	−10	−2	3	−1	1.5	0.5	1	0.2	−0.3	0.1	–	–
MNL	−7.099	−1.064	1.951	−0.719	1.706	–	–	0.149	−0.274	0.102	0.058	0.002
MXL	−10.070	−1.377	3.019	−0.998	1.562	0.333	0.988	0.136	−0.300	0.099	0.064	0.005
PMXL	−10.080	−1.717	3.091	0.992	1.564	0.448	0.991	0.170	−0.307	0.099	0.031	0.002

3.3.1. Base models

The estimated distributions of the mixed parameters are plotted in Fig. 5. The MNL model indicates the most bias among the three base models and for the MXL and PMXL models, we observe no bias for the bicycle infrastructure parameter. On the other hand, a considerable bias is visible regarding the elevation gain parameter — the parameter which we correlated with the number of observations per individual.

The estimated parameters for each of the models are summarised in Table 3. For the simplest of the three base models, the MNL model, the marginal substitution rates for the elevation gain and the bicycle infrastructure are +0.149 and −0.274 (true values: +0.2 and −0.3). The MXL model, accounting for the parameters' heterogeneity, estimates an even more biased taste for elevation gain (+0.136). It also underestimates the standard deviation value for elevation gain (0.333 for the true value of 0.5). The PMXL, accounting for the panel effect, shows less bias than the other base models and estimates values of the standard deviations that are closer to the true value (0.448 and 0.991, for the true values of 0.5 and 1, respectively). However, even the PMXL model does not exactly represent the average tastes of the individuals in the dataset. The taste parameter for the elevation gain attribute is still shifted towards individuals overrepresented in the sample (see also Fig. 5). For the other parameters, however, all models show almost no bias. This is because the individual number of observations is uncorrelated with the true values for these parameters.

These estimations are the starting point for searching for strategies to decrease the bias in tastes in a model. The subsequent section presents the results of the models, where correction strategies are applied.

3.3.2. Corrected models

Table 4 reports the results for all considered strategies and the trade-off between the bias of tastes and the efficiency for all tested strategy is shown in Fig. 6. We discuss the results of different strategies in turn.

Fig. 6 implies that for the PMXL models estimated on the entire datasets, the weighting strategy reduces the bias significantly, while not affecting the model efficiency. This approach performs (by far) best among all tested modelling approaches. On the other hand, MNL and MXL without panel indicate the highest bias, as they ignore the panel effect while remaining highly efficient since they include all observations from the dataset. As expected, PMXL models with the benchmark strategies (naive subsampling and pruning) give worse results than the base model in terms of both efficiency and bias.

For the uniform random truncation strategy, where the number of observations per individual is randomly cut at a threshold M, a lower truncation threshold means more balance in the resulting dataset. This, in turn, results in estimates with means closer to those of the underlying population. Even though the unweighted models with truncation at only a few observations per individual perform well in terms of taste bias, applying low truncation thresholds increases variability between random subsamples and worsens the efficiency of the models.

In the case of weighted models with truncated datasets, we observe the opposite effect, namely that low truncations increase the bias in tastes. Thus, for the weighted models, increasing the truncation values does not cause a trade-off between those two metrics, but instead improves both bias of tastes and efficiency. This makes the combination of two strategies, weighting and truncating at high thresholds, attractive when considering a joint minimisation of these two criteria, eventually reaching a Pareto-optimal performance unreachable without employing weights.

Fig. 7 presents plots of the estimated mixing distributions for all discussed correction strategies.

Table 4
Estimates, tastes, bias of tastes and D-error of and the corrected models (and the base models for comparison). The subsampling strategies with randomness factor are applied 100 times for the same dataset, and the results are averaged. For truncation, with 100 different generated populations and observations, these setups are repeated 10,000 times. All weighted approaches use the optimal weights from Algorithm 1 (with delay $n_0 = 3$ and tolerance $\tau = 10^{-6}$) and all subsampling strategies are described in Table 1.

Criterion	β_L	β_E	β_I	β_S	β_{PS}	σ_E	σ_I	$\frac{\beta_E}{\beta_L}$	$\frac{\beta_I}{\beta_L}$	$\frac{\beta_S}{\beta_L}$	$\text{Bias}_{\text{taste}}$	D-error
True value	−10	−2	3	−1	1.5	0.5	1	0.2	−0.3	0.1	–	–
Base models												
MNL	−7.099	−1.064	1.951	−0.719	1.706	–	–	0.149	−0.274	0.102	0.058	0.002
MXL	−10.070	−1.377	3.019	−0.998	1.562	0.333	0.988	0.136	−0.300	0.099	0.064	0.005
PMXL	−10.080	−1.717	3.091	0.992	1.564	0.448	0.915	0.170	−0.307	0.099	0.031	0.002
Corrected PMXL models												
Entire dataset (W)	−12.160	−2.436	3.763	−1.306	1.614	0.444	0.991	0.200	−0.311	0.107	0.012	0.001
Naive	−10.188	−1.538	3.154	−1.002	1.530	0.365	0.998	0.150	−0.310	0.0987	0.040	0.009
Pruning	−10.060	−1.529	3.068	−0.992	1.521	0.314	0.877	0.152	−0.305	0.0987	0.049	0.002
Random	−11.020	−2.193	3.306	−1.117	1.596	0.543	1.112	0.198	−0.301	0.101	0.003	0.162
Trunc 2	−10.300	−2.005	3.109	1.037	1.616	0.398	1.041	0.194	−0.303	0.101	0.007	0.058
Trunc 5	−10.080	−1.864	3.072	−1.006	1.568	0.375	1.025	0.184	−0.305	1.000	0.017	0.020
Trunc 10	−10.050	−1.804	3.077	−1.000	1.558	0.368	1.024	0.179	−0.307	0.099	0.022	0.011
Trunc 20	−10.040	−1.762	3.088	−0.998	1.559	0.368	1.024	0.175	−0.308	0.099	0.026	0.006
Trunc 50	−10.060	−1.734	3.120	−0.995	1.555	0.384	1.005	0.172	−0.310	0.099	0.030	0.004
Trunc 2 (W)	−16.162	−3.004	5.122	−1.871	1.807	0.714	1.391	0.184	−0.319	0.116	0.030	0.219
Trunc 5 (W)	−13.535	−2.563	4.292	−1.495	1.787	0.544	1.215	0.188	−0.319	0.110	0.024	0.042
Trunc 10 (W)	−12.765	−2.449	4.042	−1.387	1.680	0.527	1.113	0.191	−0.318	0.109	0.022	0.016
Trunc 20 (W)	−12.359	−2.415	3.894	−1.330	1.628	0.496	1.071	0.195	−0.316	0.108	0.018	0.006
Trunc 50 (W)	−12.139	−2.409	3.808	−1.301	1.588	0.457	1.026	0.198	−0.315	0.107	0.016	0.002
Trunc of repeated	−10.120	−1.906	3.075	−1.027	1.578	0.410	1.031	0.187	−0.304	0.102	0.014	0.004
Trunc of repeated (W)	−11.060	−2.070	3.418	−1.161	1.598	0.451	1.080	0.186	−0.310	0.105	0.018	0.049

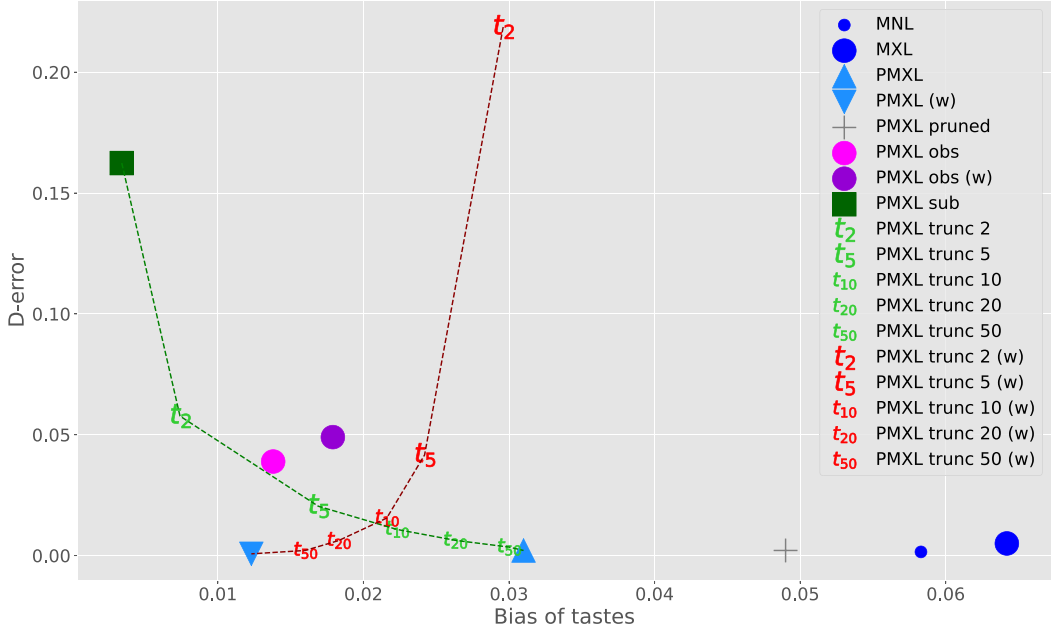
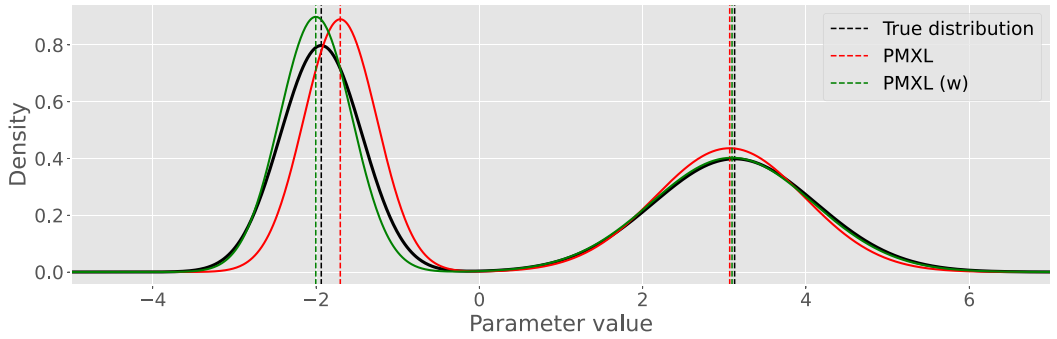
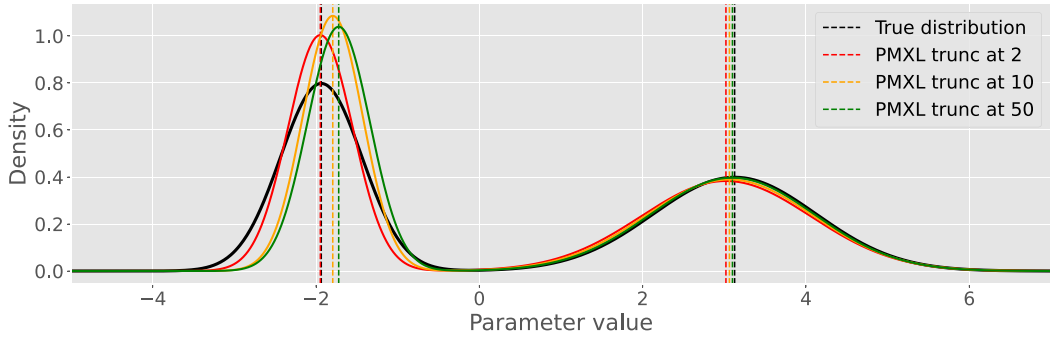


Fig. 6. Bias of tastes vs. D-error for all applied strategies.

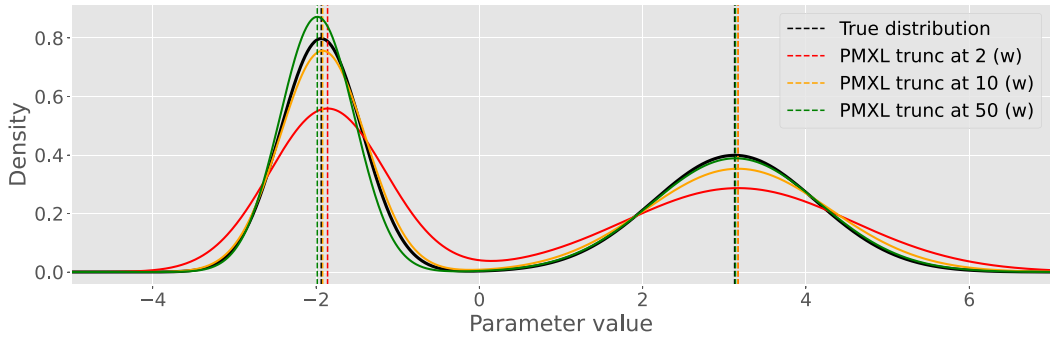
While considering dataset truncation, an important point is that the estimated parameters and tastes may vary from one random subsample to another. Therefore, for each dataset and truncation technique, we generate 100 random subsamples and average the estimates and the values of the evaluation criteria. Fig. 8 shows that the results of the weighted models are less precise than those from unweighted models for the same dataset size, both for scale and for tastes. To circumvent this issue, and to avoid the variability in the model output resulting from the random truncation, it is preferable to repeat random truncation several times and average the results for consistency.



(a) PMXL and PMXL (W).



(b) PMXL with truncations at 2, 10, and 50.



(c) PMXL (W) with truncations at 2, 10, and 50.

Fig. 7. Estimated distributions for elevation gain (left) and infrastructure (right).

Removing repeated observations behaves similarly to truncation. Due to the high preprocessing burden of the former (for some applications, detecting repeated observations might not be a trivial task) and to the similar performance of both approaches, random truncation is preferred for further analyses.

Before considering recommendations for strategies that reduce bias in the model estimations, we now turn to evaluating some of the above-discussed correction strategies in a large-scale setting. Based on joint conclusions from both case studies, the recommendations are formulated and discussed in Section 5.

4. Large-scale case study

This case study utilises a large-scale dataset to estimate a bicycle route choice model and to showcase how different combinations of subsampling and weighting approaches can influence model results, considering conclusions from the simulation study in Section 3. We aim to determine a panel configuration for a bicycle route choice model that corrects the inherent bias caused by the imbalanced number of observations per individual while retaining high model efficiency and reasonable computational time. The

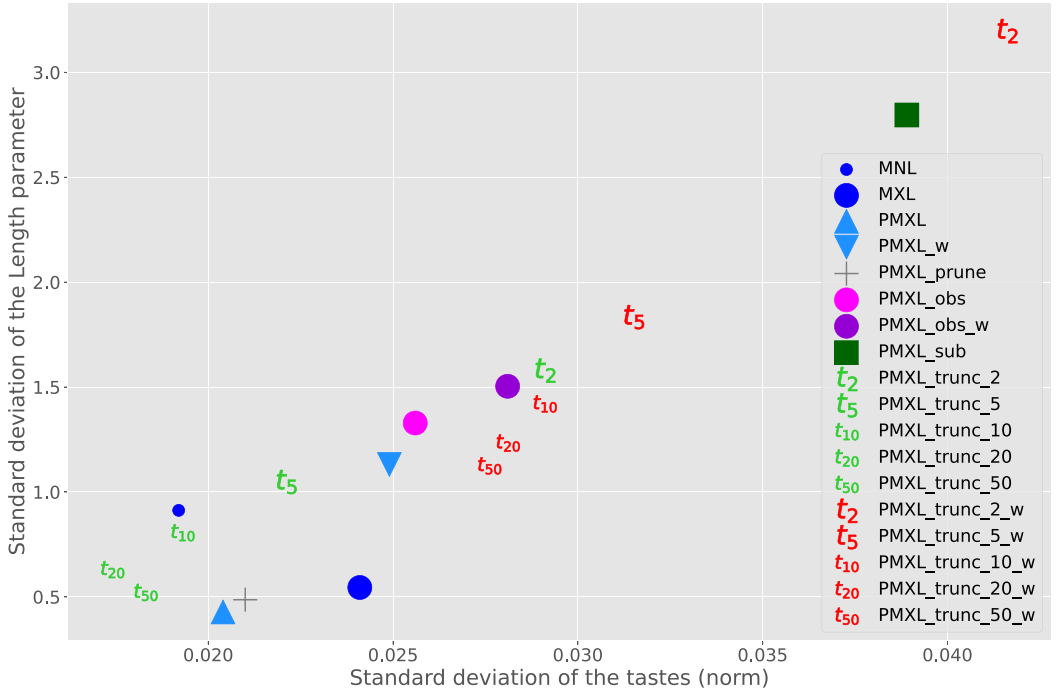


Fig. 8. Standard deviation of the tastes (norm) vs. Standard deviation of the Length parameter.

focus of this case study is to evaluate the strategies considered in previous sections, not to interpret behavioural aspects of the route choice models themselves.

4.1. Data

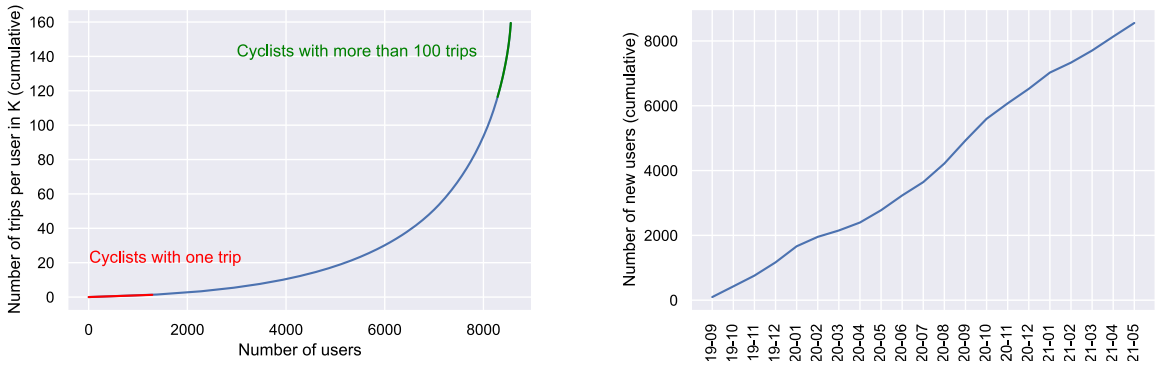
We utilise a large-scale crowdsourced dataset of bicycle GPS trajectories, received from Hövding.[1] The original dataset covers the entire Copenhagen metropolitan area in the period from September 16, 2019, until May 31, 2021 and consists of 365,813 trips from 10,049 individuals. For a detailed description of the data, the bicycle network, and the algorithms applied for data processing, we refer to a study dedicated entirely to estimating elaborated route choice models based on this dataset (Łukawska et al., 2023). The final dataset used for model estimation consists of a large number of 159,451 trips made by 8555 cyclists.

4.1.1. Panel imbalance

The number of trips per individual in the entire period of the Hövding dataset highly varies — from only one trip up to 555 trips per cyclist (see Fig. 9(a) for the cumulative distribution.). In total, 77% of the trips have been cycled by less than one-fourth of individuals in the sample (23%). This disparity can be linked to the cycling frequency, but might as well be related to the duration for which the device has been available for each individual (or both). The number of cyclists actively using the Hövding airbag for their trips has been systematically increasing each month within the data period (see Fig. 9(b)). Therefore, the sheer number of trips per individual in this dataset (and in many crowdsourced datasets if we consider a fixed time period) is not necessarily an indicator of cycling experience or good network knowledge.

It becomes apparent that the estimation of a model with a "flat" structure (for example, MNL or MXL without panel) will favour route choice preferences of these cyclists who are overrepresented in the sample. In the case of the considered data, that would mean that the contribution of 1291 cyclists with only one trip (15% of all individuals) would be almost equalised by only two cyclists with the most trips (555, 517, respectively). While this compensation might serve a purpose for some datasets or modelling applications (for example, for predicting bicycle demand in the network), in this study we aim to propose a modelling strategy that assures an equal contribution of each individual and preserves the inherent taste heterogeneity of the sample. We will therefore further apply some of the strategies from the simulation study to tackle the problem of panel imbalance.

[1] https://hovding.com



(a) Cumulative number of trips per user

(b) Cumulative sum of new users per month

Fig. 9. User patterns in the Hövding dataset.

4.2. Setup

This section describes the model setup of the large-scale models, the procedure of applying correction strategies to the Hövding dataset, and the evaluation criteria for the model results. All models presented in this section are estimated using the *xlogit* package in Python (Arteaga et al., 2022) with the correction from Appendix A to enable the estimation on the large dataset and to ensure the comparability of the execution time across models. The models are run on a machine with a Threadripper 3960X (24-core processor), 128 GB of RAM and an Nvidia RTX 3080 TI GPU.

4.2.1. Attributes and utility specification

In the models estimated in this study, we evaluate the influence of the bicycle network attributes in the Copenhagen metropolitan area on the route choice preferences of cyclists. We assume that each of the attributes contributes to the utility function linearly, and we include the attributes expressed in the form of tastes (in length unit in relation to distance, value-of-distance space), to enable a comparison of preferences across models. Besides the length of the total route, the attribute categories modelled in value-of-distance space include infrastructure type, land use, surface type, cycle superhighways classification, as well as information about the allowed direction of traffic. Attributes modelled in the preference space (i.e. that are not scaled by length) include vertical elevation, intersection information, and the number of stair segments. For more details about the bicycle network and the description of the attribute categories, we again refer to Łukawska et al. (2023).

Domain knowledge suggests that there is high heterogeneity in the route choice preferences for the land-use attributes (Prato et al., 2018). As we intend to reveal this heterogeneity, we include the land-use attributes as mixed parameters in our models, and we assume the parameters for these attributes to be normally distributed across individuals. All the remaining parameters are assumed constant across individuals.

4.2.2. Correction strategies

We set up a semi-automatic procedure to investigate multiple combinations of sampling and weighting strategies to reduce the bias in the model. The procedure consists of several steps. First, we determine a set of truncation thresholds $\mathcal{M} = \{1, 2, 5, 10, 20, 50, 100, 200\}$, and for each truncation threshold $M \in \mathcal{M}$, we create 20 subsamples $S_{M,j}$, $j \in \{1, \dots, 20\}$, applying the uniform random truncation (see Section 2.4.1). For each of the subsamples $S_{M,j}$, we estimate a base PMXL model. Finally, for each of the subsamples $S_{M,j}$, we find the optimal set of weights $\mathbf{w}^*_{M,j}$ with Algorithm 1 (with input parameters $n_0 = 1$ and $\tau = 10^{-2}$) and we estimate a weighted PMXL model with these weights. We present the results and evaluate all models in Section 4.3.

4.2.3. Evaluation criteria

We employ several criteria to evaluate the application of strategies from the simulation study on the large dataset. As we aim to ensure that each of the individuals contributes equally to the estimated model, we consider each of the measures in both the standard (unweighted) and the weighted version, with the optimal weights from Algorithm 1. Besides the weighted log-likelihood (Eq. (10)), we also compute the weighted McFadden's pseudo-R-squared to enable a comparison between all the models, regardless the sample structure. McFadden's pseudo-R-squared (McFadden et al., 1973) is defined as

$$R^2 = 1 - \frac{\mathrm{LL}}{\mathrm{LL}_0}, \tag{15}$$

where LL denotes the log-likelihood of the evaluated model, and LL_0 denotes the log-likelihood of the "null" model (a model without independent variables).

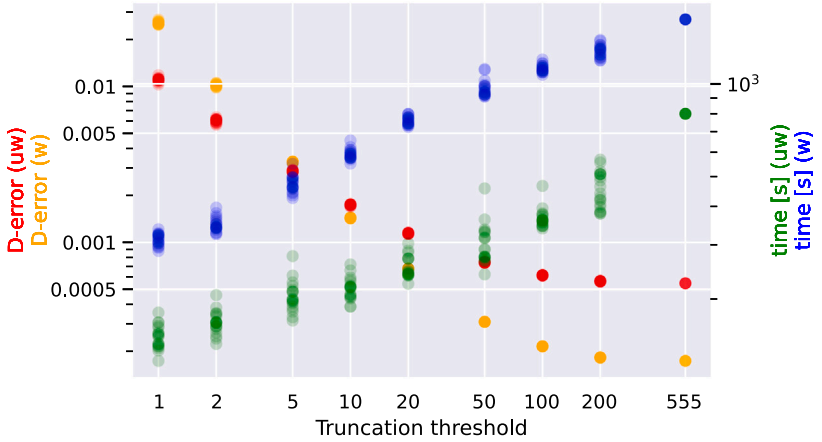


Fig. 10. Log–log plot of the D-error values versus evaluation time for the large-scale models.

To compute the weighted McFadden's pseudo-R-squared R^2_w, we extend the above definition by applying weights to the individual contributions, i.e.:

$$R^2_w = \sum_{n=1}^{N} w_n^{*\top} \left(1 - \frac{\mathrm{LL}_n}{\mathrm{LL}_{n,0}}\right), \tag{16}$$

where $w_n^{*\top}$ denotes the element of the vector of optimal weights $\mathbf{w}^*$ (Algorithm 1) referring to the individual n. LL_n and $\mathrm{LL}_{n,0}$ denote the contribution of the individual n to the total log-likelihood in the evaluated model, and in the "null" model, respectively (see Eq. (10)).

As in the simulation study in Section 3, we measure model efficiency by means of the D-error. Additionally, with a dataset with such a multidimensional complexity – 159,451 observations, 8555 individuals, 38 network attributes, and up to 31 alternatives per choice situation – the magnitude of computational estimation time becomes a non-negligible criterion and is considered in the evaluation.

4.3. Results

This section investigates some of the strategies from the simulation study on a large dataset and their influence on the model results.

4.3.1. Evaluation of the strategies

Fig. 10 shows the development of D-error and computational time for both weighted and unweighted models in relation to truncation thresholds. For both curves, the trends are as expected, and we make several observations based on the plots. In the case of weighted models, it is the estimation of the initial solution (out of the multiple estimations needed to determine the optimal weights in Algorithm 1) that contributes the most to the total computation time. Additional analysis reveals that if we consider two models with the same truncation threshold, one with and one without weights, the difference in the sheer estimation time of the final model (after the weights have been determined) is minor. The efficiency of the weighted models is worse than in the case of the unweighted models, until a certain threshold. Above this threshold weighted models are more efficient, a trend that we also observe in the simulation study in Section 3. Multiple observations per individual contribute relevantly to the total estimation time which results from both the number of iterations before the model reaches convergence and the time to compute the individual probability from Eq. (6). The whole dataset consists of only 2.6% more trips than the subsample truncated at 200 trips per individual; however, it takes (on average) twice the time to estimate a model on the entire dataset, in the case of the unweighted model. Also, the benefit in the D-error value is marginal.

Combining the findings from Section 3 (and, more precisely, from Fig. 6) with the results displayed in Fig. 10 suggests that the PMXL model with weights from Algorithm 1 best reflects the underlying individual preferences of an imbalanced sample. Additional consideration of the computational time criterion suggests that the optimal strategy to reveal the heterogeneity in the large-scale dataset and to assure that the route choice model is equitable, is truncating at around 20 observations per individual, and estimating a weighted PMXL model with weights from Algorithm 1. We present and discuss the results of this model in the next section, benchmarking the results and evaluation criteria against other approaches.

Table 5
Estimations of random parameters and values of evaluation criteria for bicycle route choice models.

	PMXL T1		PMXL T1 (W)		PMXL T20		PMXL T20 (W)		PMXL		PMXL (W)		MXL		MNL	
	coeff.	z-value	coeff.	z-value	coeff.	z-value	coeff.	z-value	coeff.	z-value	coeff.	z-value	coeff.	z-value	coeff.	z-value
Fixed parameters																
See Appendix D																
Random parameters																
Land-use (right-hand side)																
High-rise urban areas	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Green areas	−0.03	2.33	−0.02	0.92	−0.01	0.54	−0.02	6.67	−0.01	2.62	−0.02	14.41	−0.02	8.37	−0.07	23.49
Areas near water	−0.12	4.42	−0.12	4.16	−0.08	11.02	−0.11	20.25	−0.12	26.03	−0.11	42.57	−0.12	20.57	−0.18	38.90
Industrial areas	0.03	−1.78	0.01	−0.75	0.04	−8.30	0.03	−8.78	0.02	−7.05	0.03	−18.07	0.01	−3.30	−0.03	8.43
Low-rise urban areas	−0.02	1.99	−0.02	1.40	0.01	−3.07	−0.01	4.73	0.02	−8.89	−0.01	9.84	0.00	0.31	−0.03	13.67
Open landscape	0.01	−0.46	−0.01	0.11	0.02	−3.67	0.02	−3.22	0.04	−10.44	0.02	−7.38	−0.01	2.22	−0.06	13.55
Green areas (sd)	1.66	11.97	2.27	9.18	2.36	69.27	2.33	62.47	2.64	119.51	2.33	121.85	1.68	54.62		
Areas near water (sd)	2.15	9.12	2.46	6.66	3.04	40.07	3.04	50.40	2.12	48.07	2.98	98.52	1.43	23.37		
Industrial areas (sd)	2.20	13.03	2.72	9.28	2.89	65.23	2.88	68.32	2.78	105.79	2.88	133.29	2.12	59.52		
Low-rise urban areas (sd)	1.39	9.83	1.82	7.23	2.26	67.08	2.17	60.65	2.59	120.73	2.15	116.93	1.65	58.56		
Open landscape (sd)	2.65	10.49	3.36	7.29	3.61	57.26	3.67	51.94	3.67	93.27	3.67	100.58	2.41	45.92		
Evaluation																
Number of observations	8,555		8,555		80,035		80,035		159,451		159,451		159,451		159,451	
Number of individuals	8,555		8,555		8,555		8,555		8,555		8,555		8,555		8,555	
Number of parameters	38		38		38		38		38		38		38		33	
Final log likelihood	−13,996.5		−14,146.8		−125,459.3		−125,896.6		−244,695.4		−245,693.5		−257,966.5		−261,831.2	
Final log likelihood (W)	−11,897.4		−11,788.5		−118,336.2		−118,114.1		−235,612.0		−234,899.1		−239,904.5		−243,425.5	
McFadden's pseudo-R2	0.312		0.305		0.347		0.344		0.364		0.361		0.329		0.319	
McFadden's pseudo-R2 (W)	0.333		0.339		0.335		0.336		0.335		0.338		0.324		0.314	
D-error	1.11×10^{-2}		2.57×10^{-2}		1.16×10^{-3}		6.78×10^{-4}		5.47×10^{-4}		1.74×10^{-4}		6.07×10^{-4}		5.01×10^{-4}	
Evaluation time (GPU; in s)*	150.9		312.8		254.7		760.0		799.2		1617.8		153.4		259.3	

*Time for weighted models includes both the algorithm to find the optimal weights and the model estimation.

4.3.2. Bicycle route choice models

We present the results of the route choice model with the strategy determined as optimal, namely truncating at 20 observations per individual and weighting. We include several further models for comparison: the unweighted model for truncation at 20, both unweighted and weighted model for truncation at 1, as well as models estimated on the entire dataset: MNL model, MXL model, and PMXL models, both unweighted and weighted.

For models where truncation is performed, we compute the "average" model (both weighted and unweighted), based on 20 subsamples to eliminate the randomness of the truncation step. For each of the models, we compute the following evaluation criteria (again as average): log-likelihood (Eq. (6)), weighted log-likelihood (Eq. (10)), McFadden's pseudo-R-squared (Eq. (15)), weighted McFadden's pseudo-R-squared (Eq. (16)), and D-error (Eq. (9)). We also compute the evaluation time for all estimated models. Please note that we refrain from performing a hold-out sample validation since the goal of the models with correction strategies is not to assure high performance in terms of predicting choices on the individual level, but rather to describe the preferences of an "average" choice maker from the dataset.

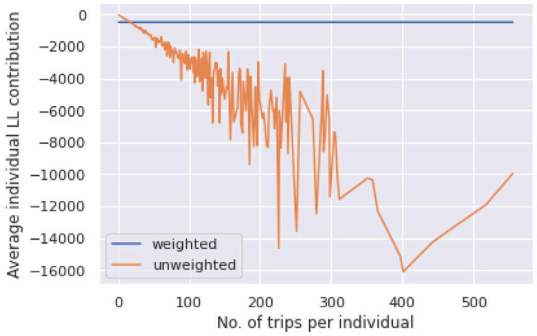
We report the results for the random parameters in the MXL models as well as the evaluation criteria in Table 5. Additionally, Appendix D includes the estimation results for the full set of parameters.

The results of the mixed models indicate a very high heterogeneity for all land-use attributes. The estimated mean values oscillate around zero for all land-use attributes and are even insignificant for some of the strategies. An exception to this is areas near water that are clearly preferred over the reference category for all the models. For the attributes open landscape and low-rise urban areas, the coefficients change sign depending on the model type (e.g. 0.021 for PMXL, −0.014 for PMXL W, and −0.034 for MNL for low-rise urban areas), causing different qualitative interpretations of the model results. The estimated standard deviation values of mixing distributions are high, and even higher when the panel effect is taken into account in the model.

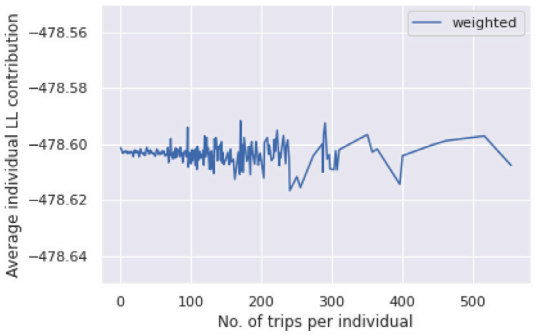
The results from Section 3 suggest that MNL and MXL models are biased towards the overrepresented individuals. In the large-scale experiment, we observe that including the panel effect changes the signs of some of the parameters, but we do not observe a systematic bias for any of the attributes in the large-scale models. Fig. 11 compares the individual contributions to the log-likelihood expression for both the unweighted and weighted models on the entire dataset. It is surprising that the difference in the results between these two models is not more pronounced, despite the notable dominance of the overrepresented individuals in the sample. This can be attributed to the complexity of the route choice models in general and might indicate that there is no strong correlation between individual model parameters and the number of observations per individual.

5. Recommendations for reducing bias in imbalanced panel datasets

This study seeks to reveal and reduce the bias when modelling choice behaviour using panel datasets that are imbalanced in terms of the number of observations per individual. The simulation study finds that models without panel structure (MNL and MXL without panel) indicate the highest bias of tastes among all tested approaches. The findings of the simulations suggest that the combination of two strategies, weighting and truncating at high thresholds, performs best when jointly minimising the bias of tastes and maximising the efficiency of the PMXL model. It is recommended to repeat the random truncation several times and average the results to avoid variability in the model output. Additionally, the analysis in the large-scale study reveals that a large number of observations per individual contributes greatly to the total estimation time of the model, and already a minor reduction of this imbalance causes considerable savings in the computational time.



(a) Individual log-likelihood contribution (weighted versus unweighted model)



(b) Individual log-likelihood contribution (weighted model, zoom)

Fig. 11. Individual log-likelihood contribution in the PMXL models on the whole dataset.

Combining the findings from the simulation study and the large-scale study suggests that, if computational resources are available, the PMXL model with weights from Algorithm 1 best describes the underlying preferences of an imbalanced sample. However, if taking a subsample should decrease the computational time significantly, or even make it possible to estimate a model in the first place (due to hardware restrictions), the recommended strategy is to truncate at a high truncation threshold and apply the weighting. Results achieved with this approach differ negligibly from the estimates of the model that performed best in terms of bias-efficiency trade-off (weighted PMXL), as suggested by both the simulation study and the large-scale study.

Restricting the influence of overrepresented individuals by applying low truncation thresholds reduces the bias significantly, but performs very poorly in terms of efficiency. If truncating at a low threshold is necessary, for example, due to the lack of computational resources, it is recommended to refrain from weighting and instead to perform multiple truncations and average the results. If the data availability does not allow for more than a few observations per individual, weighting only worsens the efficiency.

In terms of the evaluation criteria, we aim to optimise the weighted measures (marked by "(W)") as they reflect the equitable models and consider the optimal weights. Based on these evaluation criteria, the weighted models performed better than the unweighted models, by design. By including both types of measures, we aim to showcase that the goodness-of-fit measures applied to the available dataset, and traditionally optimised in the process of determining the *best* model, might in fact not optimally reflect the preferences of the individuals in the dataset.

6. Conclusion

In the emerging large crowdsourced platforms, the individuals contribute to the collected datasets with a varying number of observations. This can be owing to the different frequencies of the actual contribution, different periods of active data delivery, and other reasons that are not directly related to the frequency of the actual activity in the data collection process. This study addressed the problem of choice models based on such imbalanced datasets, where individuals might have a large variety of the number of observations.

Our study revealed the existing bias in individual tastes when applying simpler methods, such as MNL or MXL models, and showed that it is possible to reduce this bias by applying weighting and subsampling strategies. The simulation experiment showed the bias–variance trade-off a modeller might face when choosing the optimal strategy. We proposed a weighting algorithm inverting this trade-off by balancing the contribution of each individual to the log-likelihood function. This allows maximum potential efficiency while keeping a reliable explanatory model that is not biased towards individuals overrepresented in the sample. We found that the weighted panel MXL model estimated on the entire dataset performs best in terms of minimising the bias-efficiency trade-off in the estimates.

On an example of a bicycle route choice problem, the large-scale case study estimated and evaluated a model that ensures equal contribution of each individual to the estimation results, regardless of their representation in the sample. We showed that the combination of truncating at a high threshold and weighting can result in a model as reliable as the most effective weighted PMXL model estimated on the whole dataset, with a considerable reduction in the computational burden.

This study is the first to address and tackle the bias-efficiency trade-off that will become increasingly prevalent with a growing number of large, imbalanced datasets from emerging sources. The results of this study revealed that the model estimates are most unbiased for the parameters 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. In future research, it would be relevant to test the proposed correction strategies on different datasets to see how the preferences might change, i.e. how the number of observations correlates with the individual taste. In our study, we assumed normal distribution for the random parameters in all estimated MXL models, in both the simulation experiment and large-scale case study. Analysing the bias and

testing the proposed methods for other, skewed distributions (e.g., log-normal), is a natural direction for future work. However, it is deemed beyond the scope of this paper, as it would require further simulations and considerations, and potentially alternative weighting-truncating correction strategies.

We investigated the issue of imbalanced panel datasets by considering route choice problems; however, the findings are transferable to other problems in the transport field as well as other domains that apply choice modelling to understand individual behaviour.

CRediT authorship contribution statement

Mirosława Łukawska: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – original draft, Visualization. **Laurent Cazor:** Methodology, Software, Formal analysis, Investigation, Writing – original draft, Visualization. **Mads Paulsen:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision. **Thomas Kjær Rasmussen:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision. **Otto Anker Nielsen:** Conceptualization, Validation, Writing – review & editing, Supervision.

Declaration of competing interest

None.

Acknowledgment

The authors wish to acknowledge Prof. Anders Fjendbo Jensen for his contribution at the early stage of the study.

Appendix A. Reformulations of the log-likelihood expressions

In this appendix, we present the reformulations of the expressions for log-likelihood that are necessary for successful estimation of the model on the whole dataset from the large-scale case study (Section 4). For readability reasons, we refer to the MNL probability (Eq. (1)) of the chosen alternative in the choice situation t as $P_t(\beta)$.

$$\begin{aligned}
\mathrm{LL}(\beta) &= \sum_{n=1}^{N} \ln \hat{P}_n \\
&= \sum_{n=1}^{N} \ln\left(\frac{1}{R}\sum_{r=1}^{R}\prod_{t=1}^{T_n} P_t(\beta_{n,r})\right) \\
&= \sum_{n=1}^{N} \ln\left(C_n^{T_n}\frac{1}{R}\sum_{r=1}^{R}\prod_{t=1}^{T_n}\frac{P_t(\beta_{n,r})}{C_n}\right) \\
&= \sum_{n=1}^{N}\left(T_n\cdot\ln(C_n)+\ln\left(\frac{1}{R}\sum_{r=1}^{R}\prod_{t=1}^{T_n} P_t(\beta_{n,r})\right)\right) \\
&= \sum_{n=1}^{N}\left(T_n\cdot\ln(C_n)+\ln\left(\sum_{r=1}^{R}\exp\left(\sum_{t=1}^{T_n}\ln\left(\frac{P_t(\beta_{n,r})}{C_n}\right)\right)\right)-\ln(R)\right) \\
&= \sum_{n=1}^{N}\left(T_n\cdot\ln(C_n)+\mathrm{logsumexp}\left(\sum_{t=1}^{T_n}\ln\left(\frac{P_t(\beta_{n,r})}{C_n}\right)\right)-\ln(R)\right),
\end{aligned} \tag{17}$$

with $C_n = \exp\left(\frac{1}{T_n}\frac{1}{R}\sum_{r=1}^{R}\sum_{t=1}^{Tn}\ln\left(P_t(\beta_{n,r})\right)\right)$, motivated by a geometric mean across individuals. The *logsumexp* (Gundersen, 2020) is a normalisation strategy often used as an implementation trick to avoid under- or overflow problems in the evaluation of the exponential function.

All the computations and adjustments are performed in the xlogit package (Arteaga et al., 2022) which employs gradient-based iterative methods to solve the optimisation problem for the likelihood. Hence, the reformulation of the gradient function is also necessary to estimate a PMXL model on the large dataset. Relying on the source code from xlogit, as well as the expression derived by Krueger et al. (2021), we reformulate the expression for the log-likelihood gradient (for an arbitrary but fixed variable $\tilde{\beta} \in \beta$) as follows:

$$\begin{aligned}
\frac{\partial \mathrm{LL}(\beta)}{\partial\tilde{\beta}} &= \sum_{n=1}^{N}\frac{\sum_{r=1}^{R}\frac{\partial}{\partial\tilde{\beta}}\prod_{t=1}^{T_n}P_t(\tilde{\beta}_{n,r})}{\sum_{r=1}^{R}\prod_{t=1}^{T_n}P_t(\tilde{\beta}_{n,r})} \\
&= \sum_{n=1}^{N}\frac{\sum_{r=1}^{R}\left(\prod_{t=1}^{T_n}P_t(\tilde{\beta}_{n,r})\cdot\sum_{t=1}^{T_n}\left(P_t(\tilde{\beta}_{n,r})\cdot\overbrace{-X_{d,t}\cdot\exp\left(-X_{d,t}\tilde{\beta}_{n,r}\right)}^{Y_{i,r,t}}\cdot D_f\left(\tilde{\beta}\right)\right)\right)}{\sum_{r=1}^{R}\prod_{t=1}^{T_n}P_t(\tilde{\beta}_{n,r})}
\end{aligned} \tag{18}$$

$$
\begin{aligned}
&= \sum_{n=1}^{N} \frac{\sum_{r=1}^{R} \left(\prod_{t=1}^{T_n} \frac{P_t(\tilde{\beta}_{n,r})}{C_{2,n}} \cdot \sum_{t=1}^{T_n} \left(P_t(\tilde{\beta}_{n,r}) \cdot Y_{i,r,t} \right) \right)}{\left(\frac{C_{1,n}}{C_{2,n}} \right)^{T_n} \cdot \sum_{r=1}^{R} \left(\prod_{t=1}^{T_n} \frac{P_t(\tilde{\beta}_{n,r})}{C_{1,n}} \right)} \\
&= \sum_{n=1}^{N} \frac{\sum_{r=1}^{R} \left(\prod_{t=1}^{T_n} \frac{P_t(\tilde{\beta}_{n,r})}{C_{2,n}} \cdot \sum_{t=1}^{T_n} \left(P_t(\tilde{\beta}_{n,r}) \cdot Y_{i,r,t} \right) \right)}{\exp\left(\ln\left(\frac{C_{1,n}}{C_{2,n}} \right)^{T_n} \right) \cdot \exp\left(\ln\left(\sum_{r=1}^{R} \prod_{t=1}^{T_n} \frac{P_t(\tilde{\beta}_{n,r})}{C_{1,n}} \right) \right)} \\
&= \sum_{n=1}^{N} \frac{\sum_{r=1}^{R} \left(\prod_{t=1}^{T_n} \frac{P_t(\tilde{\beta}_{n,r})}{C_{2,n}} \cdot \sum_{t=1}^{T_n} \left(P_t(\tilde{\beta}_{n,r}) \cdot Y_{i,r,t} \right) \right)}{\exp\left(T_n \left(\ln(C_{1,n}) - \ln(C_{2,n}) \right) + \ln\left(\sum_{r=1}^{R} \exp\left(\ln \prod_{t=1}^{T_n} \frac{P_t(\tilde{\beta}_{n,r})}{C_{1,n}} \right) \right) \right)} \\
&= \sum_{n=1}^{N} \frac{\sum_{r=1}^{R} \left(\prod_{t=1}^{T_n} \frac{P_t(\tilde{\beta}_{n,r})}{C_{2,n}} \cdot \sum_{t=1}^{T_n} \left(P_t(\tilde{\beta}_{n,r}) \cdot Y_{i,r,t} \right) \right)}{\exp\left(T_n \left(\ln(C_{1,n}) - \ln(C_{2,n}) \right) + \text{logsumexp}\left(\ln \prod_{t=1}^{T_n} \frac{P_t(\tilde{\beta}_{n,r})}{C_{1,n}} \right) \right)}
\end{aligned}
$$

with $C_{1,n} = C_n, C_{2,n} = \frac{\max(P_{t,r})}{\sqrt{R}}$. $D_f\left(\tilde{\beta}\right)$ refers to the derivative based on the mixing distribution of the parameter $\tilde{\beta}$ with a density function f. If the considered parameter is a fixed parameter, i.e. $\tilde{\beta} \in \alpha$, the component $D_f\left(\tilde{\beta}\right)$ equals 1. $X_{d,t}$ refers to a "normalised" attribute matrix (the difference between non-chosen and chosen alternatives) for choice situation t.

We perform sanity check computations to compare the estimation results of MXL models performed with the original implementation in the xlogit, and with the above-described reformulations, respectively. The results in these two methods are replicated exactly.

Implementing the above reformulations enabled us to estimate a MXL model on the large-scale dataset of 159,451 observations from 8555 individuals, with a maximum number of choice observations per individual equal to 555. The results and evaluation of this model (in the unweighted and weighted versions) are included in Section 4.

Appendix B. Alternative weighting for the PMXL model

Without loss of generality, let us consider two individuals I_1 and I_2, both facing the same choice situation and making the same choice T_1 and T_2 times, respectively. Let us further assume that $T_1 > T_2$. By $\boldsymbol{\beta}$ we denote the vector of parameters describing the preferences for this choice, and let $(\beta_1 \dots \beta_R)^\top$ be the matrix of $R > 1$ draws from the distribution of $\boldsymbol{\beta}$. The individual logit probability of each of the draws is denoted by $P_r := P(\beta_r)$.

We assume that the contribution of both individuals to the final log-likelihood is weighted by the inverse number of observations, i.e. $\frac{1}{T_1}$ and $\frac{1}{T_2}$, respectively. Therefore, the individual contribution $\text{LL}_{w,i}$ can be expressed as follows:

$$
\text{LL}_{w,i} = \frac{1}{T_i} \ln\left(\frac{1}{R} \sum_{r=1}^{R} \prod_{t=1}^{T_i} P_r \right) = \ln\left(\left(\frac{1}{R} \sum_{r=1}^{R} P_r^{T_i} \right)^{\frac{1}{T_i}} \right), \quad i \in \{1, 2\}.
$$

By $\mathbf{P}$ we denote the vector of the probabilities of the chosen alternatives from the choice situations of individuals I_1 and I_2, given the draw R, i.e. $\mathbf{P} = \left(P_1 \dots P_R\right)^\top$. We define the L_p norm of a vector $\mathbf{x} = \left(x_1 \ \dots \ x_n\right)^\top$ as:

$$
\|\mathbf{x}\|_p = \left(\sum_{k=1}^{n} x_k^p \right)^{\frac{1}{p}}.
$$

As a consequence of Jensen's inequality, the following holds:

$$
\|\mathbf{x}\|_q \leq \|\mathbf{x}\|_p \leq n^{\frac{1}{p} - \frac{1}{q}} \|\mathbf{x}\|_q \quad \text{for} \quad p < q.
$$

Applying the above inequality to the probability vector $\boldsymbol{P}$, we obtain:

$$
\begin{aligned}
\|\mathbf{P}\|_{T_1} &\leq \|\mathbf{P}\|_{T_2} \\
\left(\sum_{r=1}^{R} P_r^{T_1} \right)^{\frac{1}{T_1}} &\leq \left(\sum_{r=1}^{R} P_r^{T_2} \right)^{\frac{1}{T_2}} \\
\ln\left(\left(\frac{1}{R} \sum_{r=1}^{R} P_r^{T_1} \right)^{\frac{1}{T_1}} \right) &\leq \ln\left(\left(\frac{1}{R} \sum_{r=1}^{R} P_r^{T_2} \right)^{\frac{1}{T_2}} \right) \\
\text{LL}_{w,1} &\leq \text{LL}_{w,2},
\end{aligned}
$$

with the equality holding if and only if $P_1 = \dots = P_r$; otherwise, the inequality is strict.

As a consequence of the above inequality, we conclude that weighting by the inverse of the number of observations $\frac{1}{T_i}$ overcompensates the log-likelihood contribution of the less represented individuals in the dataset, instead of equalising the contribution for all individuals as intended.

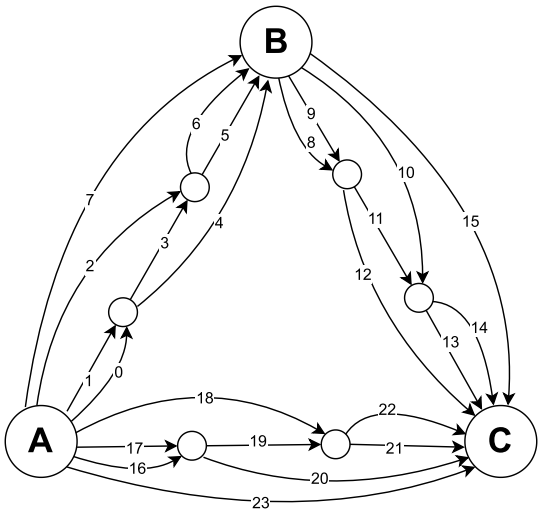


Fig. 12. Toy network with link labels.

Table 6
Table of link attributes for the toy network in the simulation experiment.

OD	*Link ID*	*Length*	*Length with steep elevation*	*Length with bicycle infrastructure*	*Length with non-smooth surface*
AB	0	2.0	2.0	0.0	0.0
AB	1	2.5	2.0	2.5	1.0
AB	2	6.0	0.5	5.0	2.0
AB	3	2.5	2.0	2.0	0.0
AB	4	4.0	4.0	0.0	1.5
AB	5	2.25	1.5	0.0	2.25
AB	6	3.0	0.0	2.0	0.0
AB	7	7.75	0.5	0.0	4.5
BC	8	1.0	1.0	0.0	1.0
BC	9	1.5	0.0	0.0	0.5
BC	10	5.0	0.0	4.0	2.0
BC	11	2.0	1.0	1.0	0.0
BC	12	4.0	4.0	0.0	2.5
BC	13	2.75	1.0	0.5	2.75
BC	14	3.5	1.0	3.5	1.0
BC	15	6.75	6.0	6.0	0.75
AC	16	2.0	2.0	0.0	1.5
AC	17	2.5	2.0	2.5	0.5
AC	18	5.0	0.0	5.0	3.5
AC	19	1.5	0.0	0.5	0.5
AC	20	4.0	4.0	0.0	0.5
AC	21	3.0	3.0	3.0	0.0
AC	22	4.0	0.0	4.0	2.0
AC	23	7.5	4.0	5.5	3.5

Appendix C. Link attributes in the simulation experiment

Table 6 presents the attribute values for the links from the toy network in the simulation experiment in Section 3. Fig. 12 shows the network with link labels.

Appendix D. Bicycle route choice model: results

See Tables 7 and 8.

Table 7
Bicycle route choice models.

	PMXL T1		PMXL T1 (W)		PMXL T20		PMXL T20 (W)		PMXL all		PMXL all (W)		MXL all		MNL all	
	coeff.	z-value	coeff.	z-value	coeff.	z-value	coeff.	z-value	coeff.	z-value	coeff.	z-value	coeff.	z-value	coeff.	z-value
Parameters in preference space																
Length	−8.01	−45.37	−10.41	−40.32	−8.70	−159.02	−9.37	−223.38	−8.72	−228.58	−9.34	−439.69	−8.35	−203.92	−6.86	−168.55
Measure of overlap																
ln(PS)	0.70	13.64	0.80	15.06	0.65	38.24	0.73	77.55	0.64	53.01	0.73	154.25	0.63	52.16	0.63	52.75
Elevation gradient																
Flat or downhill	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Steep uphill (10–35 m/km)	−0.08	−7.99	−0.11	−6.13	−0.08	−24.65	−0.10	−34.96	−0.07	−31.46	−0.10	−68.65	−0.08	−32.69	−0.06	−27.02
Very steep uphill (>35 m/km)	−0.15	−9.26	−0.20	−7.15	−0.15	−27.97	−0.18	−39.32	−0.15	−37.53	−0.18	−77.08	−0.15	−37.87	−0.13	−36.33
Intersection type																
Road hierarchy downgraded	−0.19	−9.54	−0.23	−10.08	−0.19	−28.87	−0.22	−56.33	−0.18	−37.53	−0.22	−111.99	−0.18	−38.59	−0.17	−37.50
Road hierarchy upgraded	−0.32	−15.65	−0.42	−17.94	−0.33	−49.19	−0.38	−93.10	−0.34	−72.22	−0.38	−184.80	−0.33	−69.97	−0.30	−66.77
Roundabouts	−0.14	−2.12	−0.11	−1.20	−0.08	−3.87	−0.11	−7.33	−0.05	−3.47	−0.11	−14.32	−0.02	−1.31	−0.04	−2.83
Traffic lights	−0.05	−3.62	−0.02	−0.82	−0.03	−7.02	−0.03	−7.56	−0.01	−3.80	−0.03	−14.45	−0.02	−6.36	−0.03	−8.69
Infrastructure																
No. of stair segments	−0.10	−1.97	−1.23	−1.86	−0.96	−5.36	−1.45	−11.25	−1.04	−7.68	−1.47	−22.49	−1.09	−8.17	−1.02	−8.28
Parameters in VoD space																
Infrastructure																
Medium roads w/ protected b. tracks	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Medium roads w/ painted b. lanes	0.08	−4.85	0.04	−2.00	0.07	−13.82	0.06	−14.88	0.06	−17.10	0.06	−29.25	0.06	−16.86	0.07	−15.62
Medium roads w/o b. infrastructure	0.11	−10.19	0.10	−7.50	0.10	−31.32	0.10	−45.17	0.11	−45.39	0.10	−88.09	0.12	−46.24	0.12	−37.77
Large roads w/ protected b. tracks	−0.01	1.51	0.00	0.00	−0.02	9.36	0.00	1.67	−0.02	14.85	0.00	3.03	−0.02	14.82	−0.02	10.45
Large roads w/ painted b. lanes	0.25	−6.68	0.24	−5.24	0.26	−22.81	0.26	−30.14	0.21	−27.06	0.26	−59.48	0.22	−27.01	0.25	−23.91
Large roads w/o b. infrastructure	0.23	−15.93	0.22	−11.32	0.21	−46.72	0.22	−60.52	0.21	−67.77	0.22	−118.35	0.22	−67.11	0.23	−53.04
Res. roads w/ protected b. tracks	0.08	−5.12	0.06	−3.29	0.08	−17.27	0.06	−19.59	0.10	−29.22	0.06	−38.55	0.10	−27.76	0.09	−23.16
Res. roads w/ painted b. lanes	−0.04	0.75	−0.01	0.20	−0.05	3.22	−0.04	4.92	−0.06	5.24	−0.04	9.81	−0.07	5.85	−0.10	8.03
Res. roads w/o b. infrastructure	0.20	−24.74	0.17	−18.67	0.19	−77.59	0.18	−105.36	0.18	−110.71	0.18	−206.77	0.18	−102.65	0.18	−86.11
Cycleways	−0.03	4.62	−0.04	3.95	−0.04	18.20	−0.04	23.20	−0.04	28.53	−0.04	45.53	−0.04	23.68	−0.03	17.65
Footways	0.45	−31.92	0.40	−27.24	0.42	−96.70	0.41	−151.97	0.41	−134.60	0.42	−300.59	0.43	−131.88	0.49	−112.98
Living streets	0.01	−0.14	−0.10	2.15	0.05	−3.16	−0.08	9.98	0.06	−5.13	−0.09	23.09	0.04	−3.19	0.00	−0.30
Shared paths	0.18	−17.08	0.17	−14.45	0.18	−56.03	0.18	−81.37	0.17	−77.44	0.18	−159.43	0.17	−70.05	0.16	−57.12
Pedestrian zones	0.28	−6.07	0.30	−6.37	0.30	−20.14	0.29	−33.06	0.30	−26.81	0.29	−65.42	0.29	−25.83	0.33	−22.38
Stairs	1.82	−0.22	4.63	−0.55	6.29	−2.21	−0.48	0.27	5.89	−2.60	−0.82	0.90	5.16	−2.28	2.83	−1.13

*Time for weighted models includes both the algorithm to find the optimal weights and the model estimation.

Table 8
Bicycle route choice models (cont.).

	PMXL T1		PMXL T1 (W)		PMXL T20		PMXL T20 (W)		PMXL		PMXL (W)		MXL		MNL	
	coeff.	z-value	coeff.	z-value	coeff.	z-value	coeff.	z-value	coeff.	z-value	coeff.	z-value	coeff.	z-value	coeff.	z-value
Parameters in VoD space (cont.)																
Surface type																
Asphalt	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Cobblestones	0.25	−9.00	0.25	−9.08	0.24	−27.41	0.25	−50.48	0.25	−37.61	0.25	−100.71	0.25	−37.58	0.27	−32.42
Gravel	0.13	−8.17	0.10	−5.64	0.12	−25.16	0.10	−34.48	0.13	−39.17	0.10	−67.23	0.15	−41.23	0.15	−29.73
Cycle superhighways																
No classification	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Existing	−0.03	6.30	−0.03	4.54	−0.04	24.22	−0.04	26.30	−0.04	31.81	−0.04	51.05	−0.03	28.60	−0.02	17.23
Proposed	−0.03	5.95	−0.03	3.88	−0.03	19.22	−0.03	21.41	−0.03	26.20	−0.03	41.67	−0.03	28.37	−0.04	30.87
Wrong way	0.29	−21.62	0.27	−19.15	0.27	−71.06	0.26	−105.79	0.27	−97.68	0.26	−209.24	0.28	−95.90	0.30	−75.34
Land-use (right-hand side)																
High-rise urban areas	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Green areas	−0.03	2.33	−0.02	0.92	−0.01	0.54	−0.02	6.67	−0.01	2.62	−0.02	14.41	−0.02	8.37	−0.07	23.49
Areas near water	−0.12	4.42	−0.12	4.16	−0.08	11.02	−0.11	20.25	−0.12	26.03	−0.11	42.57	−0.12	20.57	−0.18	38.90
Industrial areas	0.03	−1.78	0.01	−0.75	0.04	−8.30	0.03	−8.78	0.02	−7.05	0.03	−18.07	0.01	−3.30	−0.03	8.43
Low-rise urban areas	−0.02	1.99	−0.02	1.40	0.01	−3.07	−0.01	4.73	0.02	−8.89	−0.01	9.84	0.00	0.31	−0.03	13.67
Open landscape	0.01	−0.46	−0.01	0.11	0.02	−3.67	0.02	−3.22	0.04	−10.44	0.02	−7.38	−0.01	2.22	−0.06	13.55
Green areas (sd)	1.66	11.97	2.27	9.18	2.36	69.27	2.33	62.47	2.64	119.51	2.33	121.85	1.68	54.62		
Areas near water (sd)	2.15	9.12	2.46	6.66	3.04	40.07	3.04	50.40	2.12	48.07	2.98	98.52	1.43	23.37		
Industrial areas (sd)	2.20	13.03	2.72	9.28	2.89	65.23	2.88	68.32	2.78	105.79	2.88	133.29	2.12	59.52		
Low-rise urban areas (sd)	1.39	9.83	1.82	7.23	2.26	67.08	2.17	60.65	2.59	120.73	2.15	116.93	1.65	58.56		
Open landscape (sd)	2.65	10.49	3.36	7.29	3.61	57.26	3.67	51.94	3.67	93.27	3.67	100.58	2.41	45.92		
Evaluation																
Number of observations	8,555		8,555		80,035		80,035		159,451		159,451		159,451		159,451	
Number of individuals	8,555		8,555		8,555		8,555		8,555		8,555		8,555		8,555	
Number of parameters	38		38		38		38		38		38		38		33	
Final log likelihood	−13,996.5		−14,146.8		−125,459.3		−125,896.6		−244,695.4		−245,693.5		−257,966.5		−261,831.2	
Final log likelihood (W)	−11,897.4		−11,788.5		−118,336.2		−118,114.1		−235,612.0		−234,899.1		−239,904.5		−243,425.5	
McFadden's pseudo-R2	0.312		0.305		0.347		0.344		0.364		0.361		0.329		0.319	
McFadden's pseudo-R2 (W)	0.333		0.339		0.335		0.336		0.335		0.338		0.324		0.314	
D-error	1.11×10^{-2}		2.57×10^{-2}		1.16×10^{-3}		6.78×10^{-4}		5.47×10^{-4}		1.74×10^{-4}		6.07×10^{-4}		5.01×10^{-4}	
Evaluation time (GPU; in s)*	150.9		312.8		254.7		760.0		799.2		1617.8		153.4		259.3	

*Time for weighted models includes both the algorithm to find the optimal weights and the model estimation.

References

Alizadeh, H., Farooq, B., Morency, C., Saunier, N., 2019. Frequent versus occasional drivers: A hybrid route choice model. Transp. Res. Part F: Traffic Psychol. Behav. 64, 171–183. http://dx.doi.org/10.1016/j.trf.2019.05.009, URL https://www.sciencedirect.com/science/article/pii/S1369847818300780.

Arteaga, C., Park, J., Beeramoole, P.B., Paz, A., 2022. Xlogit: An open-source Python package for GPU-accelerated estimation of mixed logit models. J. Choice Model. 42, 100339.

Ben-Akiva, M., Bierlaire, M., 1999. Discrete choice methods and their applications to short term travel decisions. In: Handbook of Transportation Science. Springer, pp. 5–33.

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

Bierlaire, M., 2020. A Short Introduction to PandasBiogeme. Technical Report TRANSP-OR 200605, Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne.

Bliemer, M.C., Rose, J.M., 2010. Construction of experimental designs for mixed logit models allowing for correlation across choice observations. Transp. Res. B 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. Transp. Res. Part A: Policy Pract. 96, 146–153.
Christiansen, H., Baescu, O., 2022. The Danish national travel survey: Annual statistical report for Denmark for 2021. http://dx.doi.org/10.11581/dtu:00000034.
Gundersen, G., 2020. The log-sum-exp trick. https://www.gregorygundersen.com/blog/2020/02/09/log-sum-exp/.
Hess, S., Train, K.E., 2011. Recovery of inter-and intra-personal heterogeneity using mixed logit models. Transp. Res. B 45 (7), 973–990.
Hood, J., Sall, E., Charlton, B., 2011. A GPS-based bicycle route choice model for San Francisco, California. Transp. Lett. 3 (1), 63–75.
Kessels, R., Goos, P., Vandebroek, M., 2006. A comparison of criteria to design efficient choice experiments. J. Mar. Res. 43 (3), 409–419.
Krueger, R., Bierlaire, M., Daziano, R.A., Rashidi, T.H., Bansal, P., 2021. Evaluating the predictive abilities of mixed logit models with unobserved inter-and intra-individual heterogeneity. J. Choice Model. 41, 100323.
Lee, K., Sener, I.N., 2021. Strava metro data for bicycle monitoring: A literature review. Transp. Rev. 41 (1), 27–47.
Łukawska, M., Paulsen, M., Rasmussen, T.K., Jensen, A.F., Nielsen, O.A., 2023. A joint bicycle route choice model for various cycling frequencies and trip distances based on a large crowdsourced GPS dataset. Transp. Res. Part A: Policy Pract. 176, 103834.
Manski, C.F., Lerman, S.R., 1977. The estimation of choice probabilities from choice based samples. Econometrica 1977–1988.
Manski, C.F., McFadden, D., 1981. Alternative estimators and sample designs for discrete choice analysis. Struct. Anal. Discrete Data Econometr. Appl. 2, 2–50.
McFadden, D., Train, K., 2000. Mixed MNL models for discrete response. J. Appl. Econometrics 15 (5), 447–470.
McFadden, D., et al., 1973. Conditional Logit Analysis of Qualitative Choice Behavior. Institute of Urban and Regional Development, University of California ....
Molloy, J., Becker, F., Schmid, B., Axhausen, K.W., 2021. Mixl: An open-source R package for estimating complex choice models on large datasets. J. Choice Model. 39, 100284.
Nelson, T., Ferster, C., Laberee, K., Fuller, D., Winters, M., 2021. Crowdsourced data for bicycling research and practice. Transp. Rev. 41 (1), 97–114.
Ortelli, N., de Lapparent, M., Bierlaire, M., 2022. Faster estimation of discrete choice models via dataset reduction.
Prato, C.G., Halldórsdóttir, K., Nielsen, O.A., 2018. Evaluation of land-use and transport network effects on cyclists' route choices in the Copenhagen Region in value-of-distance space. Int. J. Sustain. Transp. 12 (10), 770–781. http://dx.doi.org/10.1080/15568318.2018.1437236.
Revelt, D., Train, K., 1998. Mixed logit with repeated choices: Households' choices of appliance efficiency level. Rev. Econ. Stat. 80 (4), 647–657.
Robbins, H., Monro, S., 1951. A stochastic approximation method. Ann. Math. Stat. 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. Transp. Res. Rec. (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. J. Choice Model. 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. Netw. Spat. Econ. 11 (3), 393–418. http://dx.doi.org/10.1007/s11067-010-9143-6.