Transportation Research Part C 186 (2026) 105619

Contents lists available at ScienceDirect

Transportation Research Part C

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

Investigating the impact of inferring trip purposes in a daily trip generation model

Azam Ali [a,*], Ellen H. Flaata [b], Trude Tørset [b], Stephane Hess [a], Charisma F. Choudhury [a]

[a] *Choice Modelling Centre, Institute for Transport Studies, University of Leeds, Leeds, UK*
[b] *Department of Civil and Environmental Engineering, Norwegian University of Science and Technology, Trondheim, Norway*

ARTICLE INFO

Keywords:
Unvalidated/untagged data
Smartphone-based travel surveys
Daily trip generation model
Inference
Trip purposes
Aggregate demand

ABSTRACT

Smartphones are increasingly being used to record travel behaviour data semi-passively, but low engagement rates at the verification stage leads to large amounts of untagged trip diaries (i.e. trip purposes and modes). Researchers either discard the untagged observations in the modelling stage or assume the labels assigned by the inference algorithms are error-free, i.e. have a deterministic outcome rather than a probabilistic one. In this study, we check the impact of inferring trip purposes probabilistically vs deterministically in a daily trip generation model, and if it is beneficial to utilise *inferred* untagged datasets as opposed to working with tagged datasets only. We use travel diaries collected in Trondheim, Norway, where a third of the trip purposes are untagged. We observe a significant loss in the predictive performance of the daily trip generation model when the trip purposes are inferred deterministically rather than probabilistically. Therefore, it is recommended that researchers working with passive data sources consider the *uncertainty* in the inference process. We also find that the daily trip generation model developed using both tagged and inferred untagged datasets is more efficient but has a slightly lower predictive performance than the model that uses the tagged dataset only, indicating some potential benefits of utilising *inferred* untagged datasets. However, we conclude that data quality is far more important than the number of observations.

1. Introduction

In recent years, smartphone applications have become a popular tool for carrying out surveys to study travel behaviour (Calastri et al., 2020; Cottrill et al., 2013; Harding et al., 2021; Molloy et al., 2023; Winkler et al., 2024). Smartphone-based travel survey applications have a backend that collects data from smartphone sensors, primarily the GPS traces, and utilises algorithms to detect and infer trip details such as trip segments, modes, and purposes. Typically, travel modes are inferred with high accuracy levels (around 90%); however, trip purposes have lower accuracy, ranging from 40 to 96% (Servizi et al., 2021). To ensure that the detected trip details are similar to the ground truth, most travel survey applications[1] require *validation* or *verification* from the respondents. Nevertheless, there is a substantial amount of unverified data due to non-response by participants. For example, nearly 50% of the

* Corresponding author.
E-mail addresses: ts19aa@leeds.ac.uk (A. Ali), ellen.h.flaata@ntnu.no (E.H. Flaata), trude.torset@ntnu.no (T. Tørset), S.Hess@leeds.ac.uk (S. Hess), C.F.Choudhury@leeds.ac.uk (C.F. Choudhury).

[1] There are also some cases where the respondents must specify the modes and purposes from a drop-down menu for all trips, i.e. there is no automatic inferring.

https://doi.org/10.1016/j.trc.2026.105619
Received 5 November 2024; Received in revised form 30 January 2026; Accepted 1 March 2026
Available online 12 March 2026

users who downloaded the application for a time-use survey did not fully carry out verification (Winkler et al., 2024), half of the users did not fully tag their trips in a survey in Canada (Imani et al., 2020), one-third of individuals did not completely verify their trips in a study in Singapore (Cottrill et al., 2013), and nearly a quarter of the total trips recorded were left untagged in a survey in the UK (Calastri et al., 2020). Further, there are developed travel survey applications (Marra et al., 2019) and empirical cases (Heinonen et al., 2024; Meister et al., 2025; Molloy et al., 2023) where researchers have gathered long-duration panel datasets where verification is encouraged but kept optional to reduce response burden, implying that some (or most) trip details are inferred passively without any user verification. Researchers either discard the untagged[2] observations in the modelling stage (which leads to a loss in the total number of observations), or assume that the labels assigned by the inference algorithms are fully accurate. This leads to some critical issues and research gaps.

Firstly, there has been limited research that explores the benefits of utilising *inferred* untagged datasets collected from smartphone-based travel surveys to develop travel demand models as opposed to working with the tagged datasets only. There have been some studies which assess the viability (or unviability) of using inferred untagged datasets by examining the accuracy levels of trip purpose and mode inference algorithms on a testing dataset (Harding et al., 2021; Marra et al., 2019; Yazdizadeh et al., 2019). However, it can be acknowledged that reporting stand-alone prediction accuracies of variables of interest is of limited use and interest to transport modellers and practitioners. Practitioners need to know if it is beneficial to utilise inferred untagged datasets to develop travel demand models. With a near-perfect inference model, inferred trip details will be similar to the ground truth. Utilising such untagged datasets in downstream travel demand models would increase the sample size that should translate to estimates being more accurate (closer to the true or population level parameter values and resulting in better predictive performance) and having higher efficiency (lower standard errors). Vij and Shankari (2015) is a notable exception as they explore the impact of utilising untagged modes to develop a mode choice model, by simulating the case where the modes are missing, even though it was observed in the data. This is different from a real-world scenario where some trip details are observed and some are left untagged. Therefore, in this study, we explore the impact of utilising *inferred* untagged trip purposes to develop travel demand models, specifically daily trip generation models. We focus on trip purposes as they depend on the user's context and typically have lower prediction accuracy compared to mode choices that are more easily inferred using speed and acceleration profiles. The importance of having reliable trip purpose information cannot be overstated as they form a crucial component in travel demand models including trip generation, mode choice, and destination choice models.

Secondly, nearly all of the studies that develop methods to model and infer untagged trip purposes (including Cui et al., 2018; Gao et al., 2021; Liu et al., 2023; Xiao et al., 2016; Yazdizadeh et al., 2019), consider the output of their imputation algorithms as deterministic (and error-free) which removes the stochastic or probabilistic nature of inference algorithms. This may lead to imputation errors and biases in parameter estimates when such inferred variables or datasets are used to develop travel demand models (Vij and Shankari, 2015). Vij and Shankari (2015) highlights this issue in the case of inferring modes in a mode choice model and finds that by inferring modes deterministically the value of travel time is predicted with an error of 25% compared to a 13% error when inferred probabilistically at a mode inference accuracy of 95%; however, the study does not check the impact on the estimates (coefficients and t-ratios). Andersson et al. (2022) also explores the difference between inferring modes probabilistically and deterministically in a mode choice model when inferred from call detail record data; however, they prefer the model where the modes are inferred deterministically as the estimates are more precise (have lower standard errors). Neither study checks the predictive performance on an unseen testing dataset, which prevents insights into potential over-fitting issues. All these gaps highlight the need to rigorously assess the impact of inferring variables of interest probabilistically vs deterministically and check if it is actually beneficial to infer untagged data for later use in the development of a travel demand model.

Utilising untagged datasets to estimate travel demand models implies that models are estimated where variables are inferred or missing. The methods to estimate such models depend on which variables are inferred; for example (Andersson et al., 2022) and (Vij and Shankari, 2015) use a two-staged approach to model mode choice models where modes are inferred, implying a discrete dependent variable is missing. First, the probability of observing each possible value (discrete category) for the dependent variable is inferred. Second, the probability of observing each value for the dependent variable is assigned as weights in a weighted likelihood/log-likelihood estimation (Andersson et al., 2022; Vij and Shankari, 2015). There are also some cases in the literature where the inferred variable is a continuous independent variable, such as missing income (Sanko et al., 2014) or uncertain travel time in a mode choice model (Biswas et al., 2024; Díaz et al., 2015). However, there have been no previous cases in the literature where the inferred variable forms a part of the overall dependent variable. For instance, in this study, we are interested in developing a daily trip generation model where the dependent variable is the number of trips for different trip purposes in a day (i.e. a multi-variate ordinal dependent variable), out of which some (or all) of the trip purposes are inferred. It can be acknowledged that the impact of inferring trip purposes will not just be on the individual trips but rather at the daily trip generation level. In simpler words, uncertainty in the purpose of individual trips leads to uncertainty in the resulting combinations of trip purposes at the day level. Therefore, the study also aims to develop a framework for modelling daily trips using both tagged and untagged datasets while accounting for the probabilistic nature of untagged datasets or inference algorithms.

The research objectives for this paper can be summarised as follows:

1. To assess the impact of inferring trip purposes probabilistically vs deterministically on parameter estimates and predictive performance of daily trip generation models.

[2] We prefer using the term untagged rather than unvalidated as the term validated implies that the tagged trip labels are error-free. But it is possible for people to erroneously tag their trips leading to potential discrepancies with the ground truth.

2. To investigate if inferred untagged datasets offer any improvement in developing a daily trip generation model, in terms of predictive performance and efficiency in parameter estimates, compared to using tagged datasets only.

To investigate the research questions, we use smartphone-based travel survey data collected in Trondheim, Norway where a third of the trips are labelled as “unknown” and considered as untagged. We follow a two-stage approach. We first infer the purpose of untagged trips by using socio-demographic and spatial data in a machine learning (ML) algorithm. We then use an extension of the multiple discrete continuous (MDC) model (Palma and Hess, 2022) to model the daily trips along with their purposes. We compare estimates and predictive performance of the daily trip generation model in scenarios where we use the user-tagged dataset as-is (observed), infer the user-tagged dataset probabilistically and deterministically, and use both user-tagged and probabilistically or deterministically inferred untagged dataset. The rest of the paper is structured as follows. A brief literature review is presented in Section 2 followed by the background theory and methodology described in Section 3. The data is then described in Section 4. The model development and results are discussed in Section 5 followed by the conclusion in Section 6.

2. Literature review

In this section, a brief overview of some of the studies that focus on modelling trip purposes and the number of daily trips is presented.

2.1. *Trip purposes*

A large body of research has been carried out to infer trip details including trip segments, mode choice, and trip purposes using smartphone-based travel surveys in the transport literature; for a recent and detailed literature review, readers are referred to Servizi et al. (2021). In this section, we focus only on studies that model trip purposes.

The most prevalent framework used to model trip purposes are machine learning algorithms such as decision trees, random forest, gradient boosting trees (Cui et al., 2018; Deng and Ji, 2010; Gao et al., 2021; Gong et al., 2018; Montini et al., 2014; Oliveira et al., 2014; Xiao et al., 2016; Yazdizadeh et al., 2019) and neural networks (Cui et al., 2018; Xiao et al., 2016). A few studies utilise econometric models such as multinomial logit, nested logit, and mixed multinomial logit models (Ali et al., 2024; Ermagun et al., 2017; Hossain and Habib, 2021; Liu et al., 2023). Some studies also use rule-based algorithms such as if-else statements to model trip purposes (Bohte and Maat, 2009; Shen and Stopher, 2013).

Since most researchers treat the output of the algorithm as deterministic, the metrics used to compare goodness of fit are based on deterministic predictions such as accuracy, precision, recall, and f-1 score. One issue which arises from treating outputs as deterministic and using metrics such as accuracy is that they give a distorted view of the model’s performance and are potentially affected by issues such as class imbalance. When an ML model is trained using imbalanced data, i.e. a dataset that contains choices which are underrepresented compared to other choices, the ML model does not predict the choice that is underrepresented. This can be crucial when ML models are used to predict individual instances, for example whether a person will default on a loan or be diagnosed with a disease (Chawla et al., 2004). However, in most transport-related applications, the outputs of interest are probabilities and aggregate demand (Train, 2009). There are some studies which mention the class imbalance issue in modelling trip purposes (Gao et al., 2021; Liu et al., 2023); however, the researchers do not recognise that the class imbalance issue is due to adopting deterministic goodness of fit measures. Nevertheless, in this study, we further assess the impact of inferring trip purposes deterministically or probabilistically in a daily trip generation model. Therefore, in this study, we use metrics based on both probabilities, i.e. log-likelihood and aggregate market shares, and deterministic outcomes such as accuracy and confusion matrix.

Every research study models different classes of trip purposes based on their objectives. Some studies focus on three primary categories of trip purposes such as home, work and other; whereas other studies have a large number of trip purposes, such as (Bohte and Maat, 2009) and (Gao et al., 2021), who model 13 and 8 different purposes, respectively. The explanatory variables used in the trip purpose models are divided into three main categories, i.e. socio-demographic features, trip-specific features, and spatial information. Socio-demographic features include age, gender, employment status, etc. The features specific to the trips include the mode choice, start and end time of trips, duration of the trip and activities, and the day when the trip was carried out such as weekday or weekend. Some researchers also take a dynamic approach by predicting future trip purposes based on the previous trip purpose (Cui et al., 2018; Lu et al., 2012). In terms of the spatial information used, some researchers rely on the land use databases published by local governments (Liu et al., 2023), whereas some use online mapping services such as Google Maps (Cui et al., 2018; Ermagun et al., 2017). There have been attempts to utilise social media platforms such as Twitter and Foursquare to infer the land use type as well (Cui et al., 2018; Yazdizadeh et al., 2019). Since it is quite costly to use Google Maps, some studies do not use spatial information (Gao et al., 2021). Additional important spatial variables are the home and work location of the respondent, which makes it easier to infer home and work trips (Ermagun et al., 2017; Xiao et al., 2016).

2.2. *Modelling daily trips*

One of the basic steps of an activity-based model or travel demand model is to estimate the total number of trips generated. In terms of modelling daily trips at a disaggregate level, there is no comprehensive or widely accepted framework. Different studies rely on simulation techniques, rule-based algorithms, econometric models or a combination of different algorithms (Mukherjee and Kadali, 2022). For example, Bhat et al. (2004) developed an econometric micro-simulator based on rule-based decisions and econometric

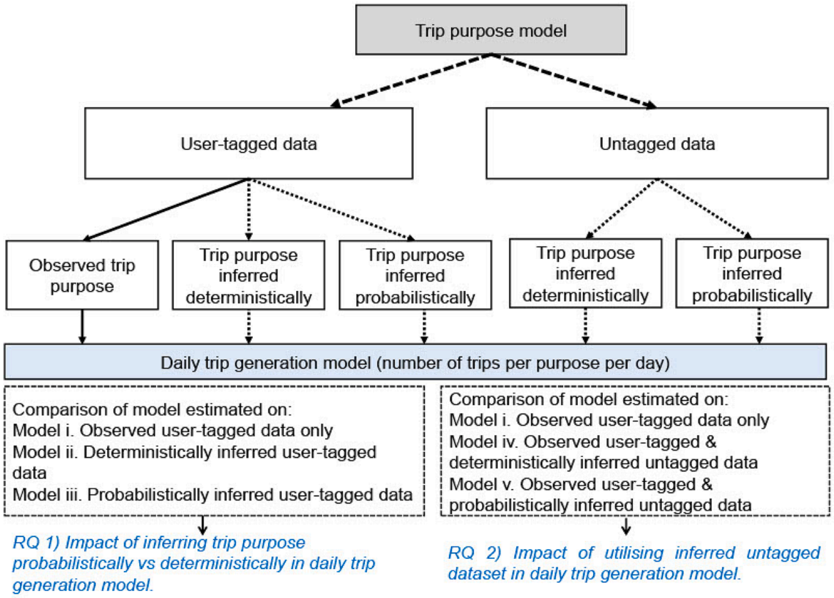

Trip purpose model
User-tagged data
Untagged data
Observed trip purpose
Trip purpose inferred deterministically
Trip purpose inferred probabilistically
Trip purpose inferred deterministically
Trip purpose inferred probabilistically
Daily trip generation model (number of trips per purpose per day)
Comparison of model estimated on:
Model i. Observed user-tagged data only
Model ii. Deterministically inferred user-tagged data
Model iii. Probabilistically inferred user-tagged data
Comparison of model estimated on:
Model i. Observed user-tagged data only
Model iv. Observed user-tagged & deterministically inferred untagged data
Model v. Observed user-tagged & probabilistically inferred untagged data
RQ 1) Impact of inferring trip purpose probabilistically vs deterministically in daily trip generation model.
RQ 2) Impact of utilising inferred untagged dataset in daily trip generation model.

Fig. 1. Flowchart of methodology.

models, to derive daily travel and activity patterns including tours and modes of individuals at a household level. Bowman and Ben-Akiva (2001) use a nested logit framework to model the tours, destinations and modes for activity patterns. Some studies model daily or weekly trips using count models (Bwambale et al., 2021); however, these models typically cater for either one trip purpose or aggregated number of trips.

More recently, Bhaduri et al. (2020), Palma and Hess (2022), Vallejo-Borda et al. (2023), and Wang et al. (2024b) have used a multiple discrete continuous (MDC) framework to model the number of trips along with their purposes or modes at a household and individual level. MDC models present a convenient tool grounded in random utility maximisation theory to model the daily number of trips along with their purposes, if analysts are happy to use a continuous dependent variable treatment for the number of trips which is a discrete outcome. A separate issue is that in their standard form, MDC models require a budget or a constraint to estimate the models, which can be difficult to specify in the context of the number of trips carried out by a person. In our analysis, we therefore make use of the extension of MDC model proposed by Palma and Hess (2022) that does not require a budget.

3. Modelling framework

In this study, our aim is to estimate a daily trip generation model which can be described as the function, $f(Y_{daily\ trip\ generation}|\theta, Z_{daily\ trip\ generation})$. In this model, the dependent variable is a set of the number of trips in a day $(y_1, y_2, .., y_k)$ for trip purposes between $1\ to\ k$, i.e.:

$$Y_{\text{daily trip generation}} = (y_1, y_2, \dots, y_k), \quad y_i \in \mathbb{N}^K \quad \text{where } K = \text{number of trip purposes} \tag{1}$$

The total number of trips in a day for each trip purpose (y_k) can be 0 or more (i.e. $\mathbb{N} = 0, 1, 2, 3, \dots$) as described in Eq. (1). The daily trip generation rate, i.e the number of trips along with their purposes, depends on explanatory variables (Z) such as age, gender, occupation status, the day of the week, etc, and the function f (and correspondingly θ values) is specified using a multiple discrete continuous approach, which is described in Section 3.2.

To simulate cases in the daily trip generation model where the trip purposes are inferred, we also estimate a trip purpose model, which can be described as a function $g(k_{trip\ purpose}|\phi, X_{trip\ purposes})$ that depends on variables related to trip purposes such as the spatial location of the trip, distance from home, the time or day of the trip, etc. In this study, we use neural networks, a machine learning algorithm, to model the trip purposes, which is further described in Section 3.1. We adopt an ML approach as ML algorithms are widely used in the literature to model and predict trip purposes, as highlighted in Section 2.1, and they have better predictive performance when there are a large number of explanatory variables compared to simpler econometric models (Ermagun et al., 2017; Cranenburgh et al., 2022; Wang et al., 2024a). However, any probabilistic inference algorithm, such as the multinomial logit, nested logit, or gradient boosting trees, can be used instead. To utilise the untagged dataset, we use the estimated trip purpose model to predict the probabilities for observing trip purposes k to generate the dependent variables for the daily trip generation model (i.e. $Y_{daily\ trip\ generation}$) either probabilistically or deterministically, which is explained later in in Section 3.2, more specifically Fig. 2.

To carry out the study, five different daily trip generation models are estimated based on:

i) the actual user-tagged dataset only;

ii) the user-tagged dataset inferred deterministically by the developed trip purpose model;
iii) the user-tagged dataset inferred probabilistically by the developed trip purpose model;
iv) the observed user-tagged dataset along with the deterministically inferred untagged dataset; and
v) the observed user-tagged dataset along with the probabilistically inferred untagged dataset.

To understand the impact of inferring trip purpose probabilistically and deterministically on the daily trip generation model, we compare the first three models including the one developed using the observed trip purposes (model i), which serves as the ground truth. To investigate the benefits of utilising the *inferred* untagged dataset, we compare the model estimated using the user-tagged dataset only with models that utilise both the user-tagged and the inferred untagged dataset. The models are compared by evaluating their predictive performance on an unseen testing dataset. Further, the model efficiency is assessed by comparing the t-ratios of parameter estimates. Fig. 1 shows a flowchart summarising our methodology.

3.1. Model for trip purpose

Neural networks are one of the most widely used machine learning algorithms in the transport community. Neural networks are supervised ML algorithms that consist of a series of linear equations which are transformed by non-linear functions. Mathematically, neural networks have a set of interconnected neurons Z with a functional form as indicated in Eq. (2), which contains a nonlinear activation f, weight W, bias b, and explanatory variables x. The probability for $class\ i$ (from a set of k trip purposes) given the explanatory variables x is given in Eq. (3). The weights and biases are estimated by maximising the log-likelihood using a stochastic gradient descent approach (Goodfellow et al., 2016).

$$Z_i = f(W * x + b) \tag{2}$$

$$P(class\ i|x) = \frac{e^{Z_i}}{\sum_i e^{Z_i}} \tag{3}$$

To train the neural network, we specify the hyperparameters. Hyperparameters can be considered as tools that impact the learning mechanism of a function (Goodfellow et al., 2016). Typically, hyperparameters in the case of neural networks, are the architecture of the neural networks (i.e. the number of neurons, layers), regularisation parameters, activation function, and the optimiser used to train the model. To select the hyperparameters, an unseen validation dataset is required to avoid overfitting. Further, it is recommended to carry out numerous trainings under different initialisation settings to achieve less biased results and reduce the risk of model non-identification (Wang et al., 2020).

Machine learning models are considered to be black boxes that do not allow researchers to understand how the ML model works. However, in recent years, there has been an increase in the use of explainable ML techniques including partial dependence plots, feature importance plots, and elasticities in the transport literature (Ali et al., 2023; Wang et al., 2020; Zhao et al., 2020). In this study, a feature importance plot, which describes the relative importance of explanatory variables to the choice probabilities, is used to explain the trained neural network model.

3.2. Model for daily trip generation

As discussed in Section 2.2, a multiple discrete continuous (MDC) approach is used to model the daily trips and purposes, specifically, the budget free version of Palma and Hess (2022), which also caters for the complementarity and substitution patterns among consumption of different alternatives (Palma and Hess, 2022). Some aspects of the model are detailed below; however, readers interested in the complete derivation of the model are referred to Palma and Hess (2022).

In the extended MDC model, a direct utility function is specified where individual n consumes quantity $y_{n,k}$[3] of good k as indicated in Eq. (4).

$$U = U_0(y_{n0}) + \sum_{k=1}^{K} U_k(y_{nk}) + \sum_{k=1}^{K-1}\sum_{l=k+1}^{K} U_{kl}(y_{nk}, y_{nl}) \tag{4}$$

The utility U_0 derived from the consumption of the outside good (y_{n0}) is specified as a linear equation as indicated in Eq. (5); whereas the utility U_k derived from the consumption of good k is further parameterised in Eq. (6). ψ is the baseline marginal utility, which is the utility at zero consumption of the good, and γ is the translation or satiation parameter which alters the consumption levels of goods. The baseline marginal utility ψ controls whether the good k is more (or less) likely to be chosen. Whereas the satiation parameter γ indicates how much (or less) of good k is consumed when chosen. ζ (and Eq. (7)) captures complementarity and substitution effects between goods k and l, where if ζ is positive, there is a complementarity effect between the two goods k and l, while 0 indicates that the consumption of the goods is independent of each other, and negative values indicate substitution between the goods.

$$U_0(y_{n0}) = \psi_{n0} y_{n0} \tag{5}$$

[3] Typically, MDC models use the notation x_{nk} to indicate the quantity $x_{n,k}$ for good k; however, we use $y_{n,k}$ as previously defined.

$$U_k(y_{nk}) = \psi_{nk}\gamma_{nk}\log\left(\frac{y_{nk}}{\gamma_k}+1\right) \tag{6}$$

$$U_{kl}(y_{nk}, y_{nl}) = \zeta_{kl}(1-e^{-y_{nk}})(1-e^{-y_{nl}}) \tag{7}$$

$$\psi_{nk} = e^{\delta_k + \beta_1 * z_1 + \ldots + \epsilon_k} \tag{8}$$

Since the baseline marginal utility should always be positive, the functional form shown in Eq. (8) is assumed where δ is the intercept for the baseline marginal utility, β are the coefficients associated with the explanatory attributes z, and ϵ is a normally distributed error term with mean of zero and a standard deviation σ. To estimate the model, the Lagrangian of the equation is written and the Kuhn Krush Tucker conditions are applied to obtain the likelihood in Eq. (9), where J is the Jacobian matrix with J_{ii} and J_{ij} elements, and $I_{y,k>0}$ and $I_{y,k=0}$ are binary variables of value 1 if y_k is greater than 0 or if y_k is 0, respectively. Since in our case, we observe daily trip generation data from individuals across multiple days, a panel approach is considered as indicated in Eq. (10) where a sum across an individual's daily trips (annotated by t) in the log-likelihood is considered.

$$L(y_k) = |J|\prod_{k=1}^{K} f(-W_k)^{I_{y,k>0}} f(-W_k)^{I_{y,k=0}} \tag{9}$$

$$W_k = \delta_k + \beta_k z_k - \log\left(\frac{y_{nk}}{\gamma_k}+1\right) - \log\left(\psi_0 - e^{-y_k}\sum_{l\neq k}\zeta_{k,l}(1-e^{-y_l})\right)$$

$$J_{ii} = \frac{1}{y_i+\gamma_i} + \frac{E_i}{\psi_0 - E_i}$$

$$J_{ij} = \frac{-\zeta_{i,j}e^{-y_i}e^{-y_j}}{\psi_0 - E_i}$$

$$E_i = e^{-y_i}\sum_{l\neq i}\zeta_{i,l}(1-e^{-y_l})$$

$$LL(y_k) = \sum_{n=1}^{n}\sum_{t=1}^{t}\log(L(y_{n,t})) \tag{10}$$

In the scenario where there are uncertain trip purposes in a day, it is possible to estimate a scale parameter associated with the proportion of trips left untagged in a day. Hence, the standard deviation σ of the error term ϵ (which is mentioned in Eq. (8) in the baseline marginal utility) is parameterised as indicated in Eq. (11), where κ is the scale parameter for the tagged dataset and τ indicates a shift in the error term associated with the proportion of untagged trips in a day. Thus, a heteroskedastic MDC model is developed.

$$\sigma = e^{\kappa + \tau * proportion \quad of \quad unknown \quad trips} \tag{11}$$

There are two possible methods to generate the daily trip generation dependent variable when the trip purposes are inferred, i.e. either probabilistically or deterministically.

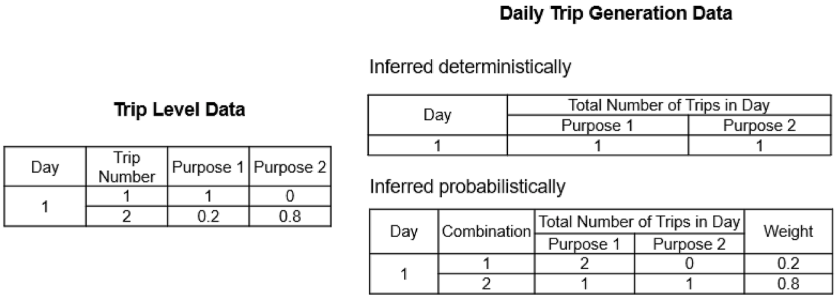
If the daily trip generation dataset is inferred deterministically, the trip purpose that has the highest probability is simply assigned, and no modification to the log-likelihood function from Eq. (10) is needed.

However, if the daily trip generation model is inferred probabilistically, the specification of the log-likelihood function changes. Indeed, for any days with inferred data, different possible combinations c of daily trip generation patterns now exist. These are generated first and then the probability of observing each combination is assigned as a weight ($weight_c$) in a weighted log-likelihood estimation as shown in Eq. (12).

$$LL(y_k) = \sum_{n=1}^{n}\sum_{t=1}^{t}\sum_{c=1}^{c} weight_c * \log(L(y_{n,t,c})) \tag{12}$$

This is more clearly illustrated through the example in Fig. 2. Imagine a case with two trips in a day, where the first is tagged *Purpose 1*, and the second is untagged with inferred probabilities for *Purpose 1* and *Purpose 2* of 0.2 and 0.8, respectively. Then there are two combinations, where one combination has two trips of *Purpose 1*, with a weight of 0.2, and the other combination has one trip of each *Purpose*, with a weight of 0.8.

The total number of possible combinations is exponentially linked by the number of uncertain trips in a day with the total number of purposes. This becomes problematic in the case when there are many untagged trips in a day. For example, if an individual carries out 10 uncertain trips in a day and there are 4 trip purposes there would be a total of 1,048,576 (4^{10}) combinations. Combinations would also be generated for instances when there is only a low level of uncertainty, i.e. a near deterministic outcome. For example, if the developed trip purpose model predicts a probability of 0.99 for one purpose and a negligible probability for other purposes, all of these purposes would still lead to combinations. Therefore, there is a need to reduce or censor the number of combinations being generated to reduce the computational burden for estimating the daily trip generation models in the case of probabilistically inferred untagged dataset.



Trip Level Data

Day	Trip Number	Purpose 1	Purpose 2
1	1	1	0
	2	0.2	0.8

Daily Trip Generation Data

Inferred deterministically

Day	Total Number of Trips in Day	
	Purpose 1	Purpose 2
1	1	1

Inferred probabilistically

Day	Combination	Total Number of Trips in Day		Weight
		Purpose 1	Purpose 2	
1	1	2	0	0.2
	2	1	1	0.8

Fig. 2. Assignment of uncertain trip purposes to day level trips and purposes.

Table 1
Aggregated trip purposes in dataset.

Trip purposes	Count	Percentage %
Home	9,058	24.58
Work/business/education	5,283	14.34
Shopping/services/caring/other	7,585	20.58
Leisure (Entertainment, visiting friends, hobbies)	2,534	6.88
Unknown	12,393	33.63

4. Data

In this study, data from a survey conducted in Trondheim, Norway is used. The survey was carried out using TravelVu, which is a smartphone application that has been used previously in research surveys in Sweden, Denmark, and Germany (Trivector Traffic , 2024). The survey was carried out from mid-October to mid-November in 2019 and completed by 870 individuals for an average of 9.3 days. The total number of trips in the dataset is 40,397; however, after removing inconsistent data and trips that are not in the Trøndelag region, 36,853 trips and 7,491 days of trip remain. For more information regarding the collection of the survey and comparison with the Norwegian national travel survey data, readers are referred to Tørset and Svaboe (2020) and Flaata (2025).

The smartphone application required users to edit and verify the trips, including modes and purposes. After verification from users, the application did *learn* the trip purposes associated with specific locations over time; however, in many cases, the respondent had to actively change from the inferred or the default trip purpose, which for many cases was "unknown", to the actual trip purpose. A major issue in the dataset is that nearly a third of the trip purposes have been assigned as "unknown" as indicated in Table 1. Therefore, the "unknown" trips are caused by respondents who did not edit or specify their trip purposes when the smartphone application prompted the users for verification. All relevant options were made available, with a comprehensive and detailed set of fifteen different trip purposes including home, work, business, shopping, entertainment, escort trips, etc., alongside an "other" trip purpose option available in cases where the standard purpose categories did not apply. Unknown trip purposes, in our study, could possibly be linked with some respondents being careless in the tagging process or due to fatigue. Nonetheless, several checks were made to ensure there were no systematic biases for individuals reporting unknown trips. For instance, it was observed that older people do not have more unknown purpose trips than younger people even though it was hypothesised that older people might have difficulties in navigating the survey application. Similarly, it was considered that shorter trips might be more likely to be reported as unknown as respondents might not recall the purposes of shorter trips; however, there was no evidence to support this. Further, it was hypothesised that the number of unknown trips increases due to survey fatigue; however, this was also not supported by the data (Tørset and Svaboe, 2020). Hence, the unknown or untagged trips are considered as missing at random.

For ease of analysis, the trip purposes are aggregated into four main categories, i.e. home, work (which also includes business and education trips), leisure, and shopping (which also includes trips to access services, trips for caring responsibilities such as pick-up/drop off, errands, and "other" purposes). We combine business trips with work category as they are quite limited in number (229 business trips). Further, education trips are combined with work trips as both are similar in nature, i.e. they are more regular and consistent in their timings and locations. We have a separate category for leisure purposes as leisure trips are discretionary in nature and also vary in terms of frequency when compared to shopping trips.

To better model and predict the trip purposes, the data was augmented with spatial data, which includes the distance to home and the 20 closest points of interest. Readers interested in more details regarding the data enrichment are referred to Appendix Section A.1. Table 2 shows the descriptive statistics for the trip purpose data. It can be observed that work trips (which also include education trips) are carried out mostly by employed people and students. The day of the week has a significant impact on trip purpose as it can be observed that 96% of work trips are carried out on weekdays. Furthermore, the time of day could also be a strong indicator as 52% of work trips are carried out in the morning between 6:00 and 9:00, home trips are mostly carried out between 15:00 and 21:00 h, and leisure and shopping trips mostly from 12:00 to 21:00. In terms of the land use, it can be observed that home trips have a higher proportion of trips to residential areas, work trips have a higher number of offices, universities, schools, and shopping areas at the destination, and shopping trips have a higher proportion of shopping areas in the 20 closest points of interest. Leisure trips have a

Table 2
Descriptive statistics of trip purpose dataset.

Proportion of trips in explanatory variables across trip purposes				
Explanatory variables	Home	Work	Shop	Leisure
Gender				
Male	0.50	0.53	0.50	0.51
Female	0.50	0.47	0.50	0.49
Employment status				
Fully employed	0.71	0.74	0.70	0.66
Partially employed	0.07	0.06	0.06	0.06
Student	0.15	0.19	0.12	0.17
Other work	0.02	0.00	0.02	0.03
Not working	0.06	0.01	0.09	0.08
Type of day				
Weekday	0.75	0.96	0.72	0.57
Weekend	0.25	0.04	0.28	0.43
Starting time of trip				
00:00 to 03:00	0.02	0.00	0.01	0.01
03:00 to 06:00	0.00	0.01	0.00	0.01
06:00 to 09:00	0.03	0.52	0.09	0.02
09:00 to 12:00	0.05	0.21	0.14	0.12
12:00 to 15:00	0.15	0.17	0.23	0.19
15:00 to 18:00	0.40	0.06	0.34	0.33
18:00 to 21:00	0.22	0.02	0.16	0.29
21:00 to 24:00	0.11	0.01	0.03	0.03
Land use categories				
Administration services	0.02	0.38	0.22	0.15
Agricultural	0.13	0.05	0.12	0.28
Conference auditorium	0.00	0.00	0.00	0.00
Entertainment	0.04	0.29	0.23	0.37
Healthcare	0.02	0.51	0.18	0.22
Industrial	0.03	0.37	0.23	0.23
Leisure residences	0.13	0.03	0.10	0.31
Logistics and warehouse	0.15	1.36	0.83	0.82
Offices	0.06	1.73	0.75	0.93
Park	0.01	0.01	0.01	0.01
Religious	0.02	0.11	0.09	0.16
Residential areas	18.41	7.52	11.12	11.34
Restaurant	0.06	0.18	0.39	0.19
School	0.13	1.14	0.52	0.47
Shopping	0.21	1.48	2.80	1.05
Sports	0.01	0.17	0.10	0.36
Temporary accommodation	0.01	0.18	0.24	0.15
Transportation	0.04	0.31	0.22	0.12
University	0.01	2.06	0.16	0.12
Unknown land-use	0.23	0.67	0.47	0.38
Average values of variables for each trip purpose				
In recreational/green area (in %)	3	6	8	18
Activity duration (hours)	10.62	4.47	0.67	2.00
Distance to home (KM)	1.91	7.16	9.04	8.96

similar land-use proportion to shopping trips; however, leisure trips have a higher proportion of sports, entertainment, and leisure residences. Leisure trips also have a higher percentage of trips that are terminated in a green space. Distance to home also acts as a proxy for home trips as home trips have a short distance to home. The average activity duration also differs significantly across the trip purposes as it can be observed that work trips on average are carried out for activities of nearly 5 h, shopping trips for less than 1 h, and leisure activities for 2 h.

Fig. 3 shows the distribution of the daily trip generation rates in the dataset. It can be observed that there is a large proportion of daily trips, i.e. 2,646 out of the 7,491 days of data, where some trip purposes are untagged. Table 3 shows the average daily trip generation patterns for a selection of socio-demographics in the user-tagged dataset. It can be observed that males have a higher number of daily trips, in particular work trips, as compared to females. On weekends, the total number of shopping and leisure trips is higher; whereas on weekdays, the total number of work trips is higher. People who do not work have fewer trips in a day compared to employed individuals and students. Household size also impacts the average number of trips as individuals in larger households carry out more shopping trips.

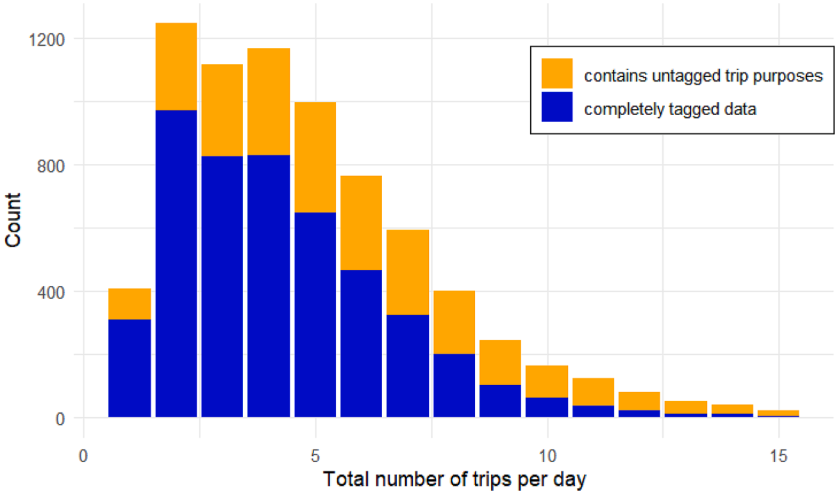


Fig. 3. Histogram of daily trip generation data.

Table 3
Descriptive statistics of daily trips in user-tagged dataset.

Explanatory variables	Average daily trips				
	Home	Work	Shopping	Leisure	Total
Female	1.52	0.86	1.32	0.43	4.13
Male	1.60	1.01	1.39	0.47	4.47
Weekday	1.56	1.21	1.30	0.34	4.42
Weekend	1.54	0.14	1.50	0.76	3.94
Employed	1.58	0.97	1.36	0.43	4.34
Student	1.53	1.21	1.10	0.51	4.35
Not working	1.43	0.16	1.78	0.52	3.90
Single person household	1.39	1.01	1.17	0.45	4.03
Two person household	1.46	0.88	1.29	0.43	4.06
Three person household	1.69	0.93	1.29	0.49	4.40
Four or more person household	1.72	0.94	1.62	0.44	4.72
Household without children	1.50	0.96	1.26	0.46	4.18
Household with children	1.71	0.87	1.60	0.43	4.60
Household without old age individuals	1.56	0.95	1.35	0.45	4.30
Household with old age individuals	1.30	0.17	1.70	0.80	3.97

5. Model development and results

To carry out the study, the user-tagged dataset was randomly split at the day level into three parts[4] with 20% of the data set aside for modelling trip purposes (4,063 trips carried out in 969 days), 60% of data (12,571 trips reported in 2,907 days) for estimating the daily trip generation model (this dataset was also used as a testing dataset for the trip purpose model), and 20% for testing the daily trip generation model (4,196 trips in 969 days). The daily trip generation model also utilises the untagged dataset that consists of 2,646 days of data or a total of 16,023 trips out of which 12,393 trip purposes are left untagged and are inferred from the trip purpose model. We refrain from using the same estimation dataset for the trip purpose and the daily trip generation model to prevent potential bias and endogeneity. Bias can potentially arise in a two-stage modelling approach when the same data used to estimate the first-stage trip purpose model is also used to generate forecasts that subsequently serve as dependent variables for the second-stage daily trip generation model. In this case, information used to build the first-stage model is reused to estimate the second-stage model, preventing independent estimation of the second-stage model and potentially introducing bias. Put simply, the errors associated with and introduced in the trip purpose model would propagate to the daily trip generation model. By employing disjoint datasets, we ensure that model estimates are derived from independent observations, which in turn can reduce potential bias. Nevertheless, some

[4] We tested different proportions (10, 20, 30, and 40%) of the user-tagged dataset for modelling trip purposes. The results were as follows: 10% of data yielded 81.01% accuracy and average LL value of −0.514 on the testing dataset, 20% resulted in 84.37% (LL −0.479), 30% gave 85.20% (LL −0.437), and 40% gave 86.11% (LL −0.412). We further observed that using a trip purpose model estimated with 40% of the tagged dataset did not alter the overall findings. Based on these results, 20% of the user-tagged dataset was set aside for modelling trip purposes, and 60% for modelling the daily trip generation model. We recommend that for future applications, researchers examine alternative data splits and inference algorithms to identify first-stage models that perform well on a testing dataset and ensure sufficient data is available for the second-stage model.

Table 4
Hyperparameter grid space for the trip purpose model.

Description	Grid space	Optimal
Depth	[1, 2, 3, 4, 5]	3
Width	[10, 20, 30, 40, 50, 60, 70]	70
L2 regularisation	[0.0001, 0.001, 0.001, 0.01, 0.1]	0.1

Table 5
Aggregate market shares of observed and predicted trip purposes.

Dataset (total number of trips)	Market shares	Home	Work	Shopping	Leisure
Trip purpose model - training dataset (4,063)	Observed	1,483	869	1,281	430
	Predicted	1,477.89	869.47	1,290.4	425.24
Trip purpose model - testing dataset (12,571)	Observed	4,561	2,753	3,930	1,327
	Predicted	4,539.05	2,702.84	4,083.62	1,245.49
Complete user-tagged dataset (24,460)	Observed	9,058	5,283	7,585	2,534
	Percentage (%)	37.03	21.6	31.01	10.36
Inferred untagged dataset (12,393)	Predicted	2,949.13	2,176.15	5,340.33	1,927.39
	Percentage (%)	23.80	17.56	43.09	15.55

bias may persist due to potential model misspecification or residual correlation between regressors and error terms, as both datasets are drawn from the same underlying population.

The following discussion is divided into two sections where the first section describes the results from the trip purpose model and the second section focuses on the daily trip generation model.

5.1. Trip purpose model

The trip purpose model was estimated using feed-forward neural networks[5] or multilayer perceptrons, where there were more than 30 input parameters, namely all of the variables mentioned in Table 2, and the outputs are the probability for four trip purposes, namely home, work, shop and leisure trips. To find the optimal hyperparameters, different numbers of neurons, layers, and regularisation parameters were tested as mentioned in Table 4. The activation function was set as a rectified linear unit (ReLU). The optimal hyperparameter setting was found to be a neural network with three layers each having 70 neurons, and L2 regularisation of 0.1 based on the performance on a 5-fold cross-validation sample, which was repeated ten times. Further, early stopping was used to limit the number of iterations carried out to reduce the risk of over-fitting on the training dataset. The metric used to find the optimal hyperparameters was the categorical cross-entropy or log-likelihood[6]. The selected neural network model was further trained a hundred times under different initialisation settings to reduce the impact of irregularities and improve the reliability of the results (Wang et al., 2020). The neural network model was estimated using the Sci-kit learn library in Python (Pedregosa et al., 2011).

Table 5 shows the observed and predicted market shares, which are calculated by aggregating the probabilities for the trip purposes in the different datasets. It can be observed that the developed trip purpose model has a good fit on the market shares in training and testing datasets with a mean absolute percentage error of 0.6% and 3.1%, respectively. The trip purpose model slightly under-predicts the market shares of home, work and leisure trips, and over-predicts the market shares of shopping trips in the testing dataset. In terms of other probabilistic goodness of fit metrics, the log-likelihood scores are -1,364.37 and -6,018.35 for the training and testing datasets, respectively. The predicted market share for the untagged dataset is also presented in Table 5. It can be observed that the model predicts that the untagged trips have a higher percentage of shopping trips and a lower percentage of home trips when compared to the complete user-tagged dataset. This seems sensible as the travel survey application would have inferred home and work purposes more accurately and would have not prompted the user to change the purpose of such trips from “unknown”.

Fig. 4 shows the confusion matrix of the developed trip purpose model that compares the observed and predicted trip purposes in the testing dataset. It can be observed that the developed model is able to predict home, work, and shopping purposes with a high accuracy of 94%, 83% and 84%, respectively. However, leisure trips have a low accuracy of 51%. Leisure trips are likely to be misclassified as shopping trips as it can be observed that 33% of the leisure trips are incorrectly labelled as shopping trips. Poor accuracy for the leisure trips could be due to a small number of observations in the training dataset (430 observations) and greater uncertainty in the data-generating process of leisure trips.

Nevertheless, the model has a good overall accuracy of 88.91% and 84.37% for the training and testing datasets, respectively. The model achieves a high goodness of fit comparable to other studies: for instance (Ermagun et al., 2017; Gao et al., 2021; Liu et al.,

[5] We also tested a gradient boosting tree (GDBT) algorithm to model trip purposes, which yielded a similar predictive performance on the validation dataset compared to the neural network (NN) model. However, the GDBT model was prone to overfitting which was the reason why the NN model was preferred. Additionally, we estimated a multinomial logit (MNL) model as a further benchmark and found that the NN model provided higher predictive accuracy on the testing dataset.

[6] The same optimal hyperparameters were observed when the metric was f1-score and accuracy.

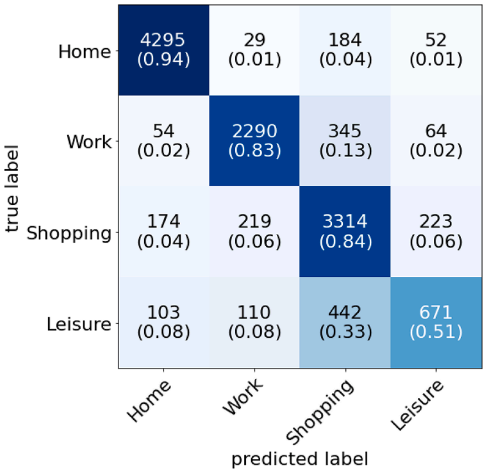


Fig. 4. Confusion matrix of observed and predicted trip purposes in testing dataset (n = 12,571).

2023; Montini et al., 2014; Yazdizadeh et al., 2019) have accuracies of 64.17%, 86.7%, 79%, 79.8%, 71%, respectively. There are some studies which achieve better goodness of fit; for instance, Gong et al. (2018) and Xiao et al. (2016) achieve accuracy over 96%. However, these studies either include more information as (Xiao et al., 2016) includes the distance to home, work, education, and shopping destination locations as explanatory variables, thus increasing the data requirements, or have less heterogeneity present in the dataset, as (Gong et al., 2018) models and predicts observations from 20 individuals only over a long time. Therefore, the developed trip purpose model is considered appropriate to predict unknown trip purposes or the untagged dataset.

To be certain that the developed trip purpose model is intuitive and can be relied to predict the untagged dataset, the top explanatory variables are plotted in Fig. 5. It can be observed that the most important explanatory variable is the duration of activities variable which is intuitive as longer duration of trips tend to happen at home, followed by work, leisure, and then shopping. Further, distance to home is quite an important variable as it is a direct indicator for home trips. Other important features include land-use variables such as shopping, residential areas, universities, offices, healthcare, schools, transportation hubs, and administrative services. The trip's starting time, specifically between 06:00 and 09:00 and between 15:00 and 21:00, is also an important feature in the trained model along with the day of the week.

5.2. Daily trip generation models

To answer our research questions, five different models were developed using the:

i actual user-tagged dataset only;
ii the user-tagged dataset inferred deterministically by the developed trip purpose model
iii the user-tagged dataset inferred probabilistically by the developed trip purpose model;
iv the observed user-tagged dataset along with the deterministically inferred untagged dataset; and
v the observed user-tagged dataset along with the probabilistically inferred untagged dataset.

One of the issues expected in daily trip generation models when the inferred trip purposes are assigned probabilistically is that a large number of combinations are generated when there are a high number of untagged trips in a day, which greatly increases the computation time for estimating the models. Therefore, to address this, we apply two forms of censoring. First, we only consider observations which have fewer than 9 trips in a day, as there were only some observations where individuals carried out 9 or more trips, as observed in Fig. 3. Filtering data in this way is part of standard data cleaning procedures when working with smartphone-based data surveys. This led to a reduction in the total number of observations as the model developed using the user-tagged dataset now consists of 2,740 days of data (reduced from 2,907 days), the testing dataset of 915 days (previously 969 days), and the untagged dataset of 2,126 days of data (from 2,646 days). Second, we do not generate combinations if the probability of inferring a particular trip purpose is lower than 0.01. The threshold represents the lowest value for which model estimation remained computationally feasible, requiring roughly 18 h on a high-performance computing facility. Increasing the threshold further reduces computational time but increases prediction error, thereby illustrating a trade-off between prediction accuracy and computational efficiency. This trade-off is demonstrated using a smaller subset of the data in Appendix A.2.

The explanatory variables used to model the daily trip generation include the type of day, gender, employment status, the number of cars and the presence of children in the household. The satiation and baseline marginal utilities are estimated for the four trip purposes, i.e. home, work, leisure, and shopping. The outside good is set as 0, which implies no trips being carried out. The substitution and complementarity effects between different trip purposes were not estimated (set to 0, implying trips for different purposes are independent of each other) as we observed them to be unintuitive and statistically insignificant for some cases. All of the estimation is carried out using Apollo in R (Hess and Palma, 2019).

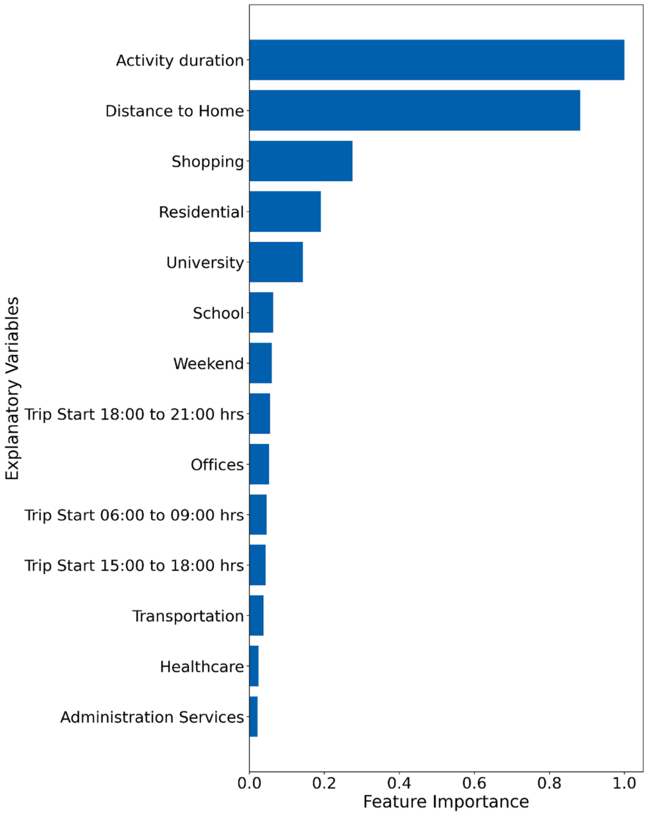


Fig. 5. Feature importance plot for the top explanatory variables in the trip purpose model.

Tables 6 and 7 compare the coefficients and the predictive performance of the estimated models on the testing dataset. First, we focus on the initial three models which use only the user-tagged dataset. It can be observed that the log-likelihood (LL) scores for the model where the *actual* user-tagged dataset has been used is −13,349.55; however, once the dataset is inferred deterministically and probabilistically from the developed trip purpose model, the log-likelihood scores are −13,022.95 and −13,523.37. Since the dependent variable, i.e. the daily trips and purposes, are different in these models, it is not possible to compare the LL scores or model fit; therefore we rely on out-of-sample predictive performance. However, it is intuitive that the model where the dataset is inferred deterministically has a better log-likelihood compared to the model where the dataset is inferred probabilistically, as this ignores uncertainty or stochasticity in the dependent variables leading to better model fit. We observe that the best-performing model is the model developed on the user-tagged dataset with a mean absolute percentage error (MAPE) of 2.99% in the aggregate demand. As expected, the predictive performance decreases when the user-tagged dataset is inferred, with a MAPE of 7.96% with deterministic inference and 4.31% with probabilistic inference. We clearly see that the model where the daily trip generation dataset is inferred deterministically has a significantly poorer predictive performance compared to the model where the dataset is inferred probabilistically.

It can be observed in Table 7 that the model where the dataset is inferred deterministically (model ii) underpredicts the total number of leisure trips in the testing dataset by 28%. This is because leisure trips are underrepresented in the trip purpose dataset leading to a class imbalance issue and also having a lower prediction accuracy as observable in the confusion matrix in Fig. 4. However, in the model where the daily trip generation dataset is inferred probabilistically (model iii), this issue is not that severe as there is an underprediction of leisure trips by only 12%. The primary reason for underprediction in model ii is that by deterministically inferring, we only consider the outcome with the highest probability rather than the complete probability distribution across the different outcomes. This demonstrates that if the inferred details, such as modes or purposes, are deterministically inferred, there is a risk of not capturing the actual data-generating process leading to a biased model.

The magnitude of the coefficients (including the γ, δ, and β parameters) vary across the three models based on the user-tagged dataset. The magnitude of the coefficients is similar for the models developed using the user-tagged dataset and when the dataset is inferred deterministically; however, the difference is quite large when the user-tagged data is inferred probabilistically. This is primarily because of the difference in the σ or the standard deviation of the error term in the models, as the model with probabilistic inference has a κ coefficient of 2.590, whereas the user-tagged model and the model with deterministic inference have a coefficient of 2.118 and 2.123, respectively. Therefore, we cannot directly compare the coefficients of the models; however, we can compare the statistical significance or t-ratios. In terms of the statistical significance of the socio-demographic features (β coeffi-

Table 6
Estimates for daily trip generation models.

Coefficients	Model i User-tagged data only		Model ii User-tagged data inferred deterministically		Model iii User-tagged data inferred probabilistically		Model iv User-tagged & deterministically inferred untagged data		Model v User-tagged & probabilistically inferred untagged data	
	Estimate	Rob.t-ratio	Estimate	Rob.t-ratio	Estimate	Rob.t-ratio	Estimate	Rob.t-ratio	Estimate	Rob.t-ratio
$\gamma\ Work$	7.905	7.29	7.442	7.18	13.021	5.97	12.415	6.80	15.240	5.94
$\gamma\ Home$	4.900	6.42	5.208	6.36	9.773	5.44	8.538	6.27	10.621	5.5
$\gamma\ Shop$	13.347	7.91	13.454	7.62	21.859	6.18	19.730	7.03	23.271	6.07
$\gamma\ Leisure$	13.340	8.66	13.441	8.18	21.327	6.49	21.683	7.08	23.236	6.35
$\delta\ Work$	−0.071	−2.57	−0.078	−3.37	−0.038	−3.31	−0.047	−2.87	−0.025	−2.26
$\delta\ Home$	0.227	6.91	0.218	6.99	0.121	5.78	0.126	6.24	0.104	5.55
$\delta\ Shop$	0.073	5.07	0.089	5.66	0.053	5.12	0.065	5.57	0.058	5.39
$\delta\ Leisure$	−0.048	−3.05	−0.099	−5.47	−0.046	−5.17	−0.032	−3.24	−0.022	−2.93
κ	−2.118	−18.60	−2.123	−18.02	−2.593	−17.08	−2.515	−18.89	−2.691	−17.29
τ	NA	NA	NA	NA	NA	NA	0.234	6.74	0.174	7.1
$\beta\ Work\ female$	−0.018	−2.64	−0.020	−2.86	−0.013	−3.14	−0.013	−3.00	−0.008	−2.52
$\beta\ Home\ female$	−0.005	−0.76	−0.006	−1.02	−0.002	−0.63	−0.007	−1.85	−0.005	−1.65
$\beta\ Shop\ female$	−0.002	−0.35	−0.004	−0.65	−0.003	−0.84	0.000	0.04	0.001	0.18
$\beta\ Leisure\ female$	−0.010	−1.30	0.002	0.25	0.001	0.34	0.002	0.32	−0.001	−0.3
$\beta\ Work\ employed$	0.206	5.63	0.218	6.21	0.120	5.56	0.135	5.58	0.097	5.05
$\beta\ Home\ employed$	0.012	1.06	0.016	1.58	0.007	1.24	0.016	2.14	0.012	2.11
$\beta\ Shop\ employed$	−0.048	−4.02	−0.051	−4.31	−0.032	−4.17	−0.029	−3.62	−0.026	−3.94
$\beta\ Leisure\ employed$	−0.022	−1.75	−0.025	−1.93	−0.012	−2.39	−0.022	−2.58	−0.016	−2.7
$\beta\ Work\ student$	0.230	5.85	0.233	6.17	0.128	5.50	0.140	5.57	0.103	5.07
$\beta\ Home\ student$	0.021	1.65	0.025	2.01	0.012	1.65	0.020	2.34	0.016	2.42
$\beta\ Shop\ student$	−0.056	−3.82	−0.056	−3.89	−0.033	−3.80	−0.038	−3.86	−0.032	−3.94
$\beta\ Leisure\ student$	−0.005	−0.34	−0.006	−0.38	−0.003	−0.45	−0.003	−0.32	−0.003	−0.44
$\beta\ Work\ weekend$	−0.246	−7.57	−0.262	−7.66	−0.137	−6.21	−0.170	−6.96	−0.127	−5.99
$\beta\ Home\ weekend$	−0.003	−0.56	−0.005	−0.81	−0.003	−0.89	0.000	−0.13	0.000	−0.01
$\beta\ Shop\ weekend$	0.011	1.80	0.005	0.82	0.004	1.07	0.007	2.08	0.006	2.08
$\beta\ Leisure\ weekend$	0.058	6.03	0.088	7.02	0.044	6.08	0.048	6.28	0.039	5.7
$\beta\ Work\ cars$	0.002	0.35	0.005	1.16	0.004	1.44	0.000	0.08	0.000	0.05
$\beta\ Home\ cars$	0.010	2.52	0.007	1.66	0.004	1.66	0.010	3.36	0.008	3.26
$\beta\ Shop\ cars$	0.012	2.64	0.009	1.92	0.006	2.37	0.005	1.70	0.003	1.63
$\beta\ Leisure\ cars$	0.001	0.30	0.013	2.81	0.006	2.93	0.000	0.02	0.000	0.19
$\beta\ Work\ children$	−0.013	−2.86	−0.009	−1.98	−0.004	−1.54	−0.008	−3.02	−0.006	−2.52
$\beta\ Home\ children$	0.020	3.80	0.020	4.26	0.014	4.66	0.012	3.88	0.010	3.84
$\beta\ Shop\ children$	0.025	4.86	0.023	4.69	0.012	4.16	0.013	4.03	0.010	3.88
$\beta\ Leisure\ children$	0.000	0.02	−0.003	−0.48	−0.001	−0.35	0.000	−0.06	0.000	−0.13
LL(final)	−13,349.55		−13,022.95		−13,523.37		−25,352.26		−25,439.68	
Number of individuals	521		521		521		798		798	
Number of observations	2,740		2,740		2,740		4,866		4,866	
Estimated parameters	33		33		33		34		34	

Table 7
Aggregate demand (number of trips for each trip purpose) of the daily trip generation models in the testing dataset.

Description	Observed demand	Predicted demand				
		Model i User-tagged data only	Model ii User-tagged data inferred deterministically	Model iii User-tagged data inferred probabilistically	Model iv User-tagged & deterministically inferred untagged dataset	Model v User-tagged & probabilistically inferred untagged dataset
Work	803	768.64	764.37	776.39	736.90	767.85
Home	1,360	1,353.50	1,376.89	1,351.88	1,338.33	1,347.48
Shop	1,104	1,066.21	1,140.16	1,078.56	1,222.78	1,211.85
Leisure	364	348.75	261.82	318.80	380.18	381.63
Total trips	3,631	3,537.10	3,543.24	3,525.63	3,678.19	3,708.81
MAPE (%)	–	2.99	7.96	4.31	5.27	4.41

cients), it can be observed that most of the variables have similar significance levels across these models. There are three coefficients ($\beta\ leisure\ employment$, $\beta\ home\ student$, $\beta\ leisure\ car$) that are not statistically significant at 95% confidence level in model i, developed using the actual user-tagged dataset, but are statistically significant in the model where the dataset is inferred deterministically, i.e. model ii, indicating a type-I error or a false positive. Similarly, two variables ($\beta\ leisure\ employment$, $\beta\ leisure\ car$) have a false positive error when the dataset has been inferred probabilistically (model iii). There are two variables in each of the models developed using deterministic inference ($\beta\ home\ cars$, $\beta\ leisure\ car$) and probabilistic inference ($\beta\ home\ cars$, $\beta\ work\ children$), which have a false negative error or type-II error, i.e. the coefficients are not statistically significant at 95% confidence level in the inferred model even though these coefficients are statistically significant at 95% confidence level in the model developed using actual dataset only (model i). This indicates a risk of imputation bias when the dataset is inferred; however, the risk is slightly higher when the dataset is inferred deterministically rather than probabilistically as one additional parameter has a false positive error. Andersson et al. (2022) also compares the standard errors and finds that when the dataset is inferred deterministically, the standard errors are lower and hence are more precise compared to the model where the dataset is inferred probabilistically. This however could possibly be an indication of imputation bias. Nevertheless, there is a potential bias-variance trade-off as when the dataset is inferred probabilistically the coefficients are more accurate (i.e. closer to the actual data-generating values) but have a higher variance and in the case when the dataset is inferred deterministically, the coefficients are less accurate but have a lower variance (Vij and Shankari, 2015).

Table 6 also shows the comparison between the model that is developed using the user-tagged dataset only with the model that contains the user-tagged along with the probabilistically inferred untagged dataset, i.e. model i and v, respectively. We avoid comparison of model i with model iv, which infers the untagged data deterministically, as deterministic inference does not fully capture the data generating process as explained earlier. There is a significant increase in the total number of observations once the untagged dataset is included, as the total number of observations present in the dataset increases by 80% from 2,740 days to 4,866 days. However, it can be observed that the predictive performance of the developed model (model v) is not better as the MAPE in predicting aggregate demand is 4.41% for model v and 2.99% for model i. Therefore, we do not find any improvement in the predictive performance on a testing dataset by inferring the untagged dataset and having more data, which is somewhat surprising. To further investigate this, we implement cross-validation to examine whether the same findings hold across different data folds and estimate additional models with varying ratios of tagged and untagged datasets, which have been listed in Appendix Section A.3. We find that the results remain consistent across these specifications.

We conclude there are two possible explanations. Firstly, it could be that the inferred untagged dataset is of poor quality, i.e. the inferred trip purposes are not similar to the ground truth leading to no improvement in the predictive performance. This issue is particularly pronounced for leisure trips, which tend to be over-predicted, perhaps due to the trip purpose model's lower predictive accuracy for these trips. Secondly, it can be argued that the parameter estimates for the daily trip generation models differ between the tagged and untagged datasets, reflecting behavioural differences. Although we tested for systematic differences between the two groups, such as whether older respondents were associated with a higher number of untagged trips or whether shorter distance trips were more likely to be untagged, no evidence supporting these hypotheses was found, as discussed in Section 4. Nevertheless, travel behaviour may still differ due to other unobserved factors, such as respondents having insufficient time to tag trips on *busier* days, leading to genuine behavioural differences between tagged and untagged observations. Since the testing dataset is drawn from the same subset as the user-tagged dataset, it does not necessarily represent a true out-of-sample testing dataset, collected at another time period or geographical area. It is therefore intuitive that the model developed on the user-tagged dataset only leads to a better fit on the defined testing dataset. However, this makes it difficult to determine which model actually converges to the true or population-level parameter values.

Another benefit of using more data to estimate models is to have more precise estimates or lower standard errors. There are several coefficients (such as $\beta\ home\ employed$, $\beta\ leisure\ employed$, $\beta\ home\ student$, $\beta\ shop\ weekend$) that are not statistically significant at 95% confidence level in model i, developed using the user-tagged dataset only, while they are statistically significant at 95% confidence level in the model that has the additional probabilistically inferred untagged dataset, i.e. model v, indicating an increase in efficiency. On the other hand, there is one coefficient ($\beta\ shop\ car$) that is statistically significant in the user-tagged dataset (model i) at 95% confidence level but is not statistically significant in model v. However, it is difficult to be certain that the increase in efficiency is from the additional information or because of bias, i.e. a type I error or false positive.

Overall, the estimates for the daily trip generation models as observed in Table 6 are intuitive. In all cases, the γ (satiation parameters) of shopping and leisure trips are higher for work and home purposes, indicating that individuals want to carry out more of these trips, when chosen. It can be observed that the marginal baseline utility for home and shopping trips is larger, indicating that all else equal, these trips are more likely to be carried out. This is understandable as most individuals would make trips for shopping purposes and return home, whereas leisure and work trips are more likely to be impacted by employment status, whether it is the weekend, etc. The τ value in models iv and v is positive and statistically significant at the 95% confidence level, which indicates that with an increase in the proportion of untagged trip purposes, there is more variance and uncertainty in modelling and predicting the daily trips. The impact of explanatory variables across the models is similar. Females are less likely to carry out work trips compared to males, indicating gender differences. This can also be observed in Table 3 as on average females carry out 0.86 daily work trips whereas males carry out 1.01 daily work trips. As expected, employed individuals and students are more likely to carry out work trips, and less likely to carry out shopping trips compared to non-working people (including retired people). With respect to weekdays, people are less likely to carry out work trips and more likely to carry out leisure trips on weekends. Further, an increase in the number of cars leads to an increase in the number of shopping and home trips. With an increase in the number of children in a household, people are more likely to carry out shopping and home trips.

6. Conclusion

One of the most prevalent issues in data collected using smartphone-based travel surveys is that a large share of people fail to tag or verify their trip details, such as modes and purposes, which leads to a large amount of missing or untagged data. In this study, we address two research gaps. First, we assess the impact of inferring missing trip purpose deterministically rather than probabilistically in a daily trip generation model. Second, we determine whether it is beneficial to utilise *inferred* untagged datasets to develop a daily trip generation model as opposed to working with tagged datasets only. We utilise data collected from a smartphone-based travel survey in Norway where one-third of the trip purposes are left untagged, even though the survey app required users to verify or correct all trip details.

To carry out the study, we develop a framework to model daily trips along with their purposes, utilising both tagged and untagged datasets. The modelling framework accounts for the probabilistic nature of untagged datasets by adapting the likelihood of a multiple discrete continuous model. The developed modelling framework will be useful for other researchers who deal with cases where the untagged (or missing) data is a dependent variable that is discrete and multiple in nature, such as uncertain or probabilistic activities in a time-use survey.

Firstly, we find that it is better to consider uncertainty in the inference process when utilising *inferred* untagged datasets to develop a daily trip generation model. We observe a 4.31% mean absolute percentage error (MAPE) in predicting the aggregate demand of the daily trip generation model when the trip purposes are inferred probabilistically (model iii), compared to a higher error of 7.96% when the trip purposes are inferred deterministically (model ii). The daily trip generation model, where the trip purposes are inferred deterministically, has a poorer fit on choices that are less represented in the trip purpose data, i.e. leisure trips. This issue is not as prevalent in the case when the trip purposes are inferred probabilistically as this approach accounts for the complete probability distribution of the trip purposes rather than only considering the trip purpose which has the highest probability. Our comparison therefore indicates that probabilistically inferring trip details better predicts (and implies a better fit on) the underlying data-generating process compared to inferring deterministically. Hence, we provide more evidence that it is better to infer untagged or missing trip details probabilistically, which is supported by Vij and Shankari (2015), compared to Andersson et al. (2022) who prefer the model where missing data is inferred deterministically. Moreover, studies that infer trip purposes and modes should highlight the probabilistic nature of their algorithms and we recommend that they also report probabilistic goodness of fit metrics such as log-likelihood, aggregate market shares, average probability of chosen alternatives, rather than metrics based on deterministic predictions.

Secondly, in our study, we observe that there are limited benefits in utilising *inferred* untagged datasets to develop a daily trip generation model as opposed to working with tagged datasets only. Generally, with more data, the relative model fit improves and the standard errors of the estimates are lower. However, we observe that the daily trip generation model estimated using the probabilistically inferred untagged data in addition to the tagged dataset, which results in a dataset of 4,866 days, has a slightly poorer predictive performance on the testing data with a MAPE of 4.41% compared to a MAPE of 2.99% when we utilise the user-tagged dataset only which comprises of 2,740 days of data. We do observe that utilising the untagged datasets leads to lower standard errors for some variables indicating an increase in efficiency; however, the underlying reasons are unclear as there can potentially be an imputation bias or a type-I/false positive error. Hence, we cannot fully assert that there are benefits of utilising *inferred* untagged datasets. It should also be highlighted that we do not have access to *true* model coefficients or an independent out-of-sample validation dataset that could serve as external benchmarks, which limits the generalisation of our conclusions. Such limitations are common in studies that rely on passively collected or inferred datasets to estimate travel demand models, where appropriate benchmarks are unavailable (Chen et al., 2016; Lee et al., 2016). While benchmarks such as values of time from the literature or average hourly wage rates can sometimes be used for validation in mode or route choice models, comparable benchmarks are not available for daily trip generation models. We thereby recommend that future researchers evaluate the benefits of utilising untagged datasets on a case-by-case basis by assessing the credibility of the parameter estimates. While the lack of improvement in our case may be seen as a negative finding in terms of the benefits of utilising untagged datasets, a counter-argument is that the loss in predictive performance once untagged data is used allows us to highlight that the quality of the data matters more than the quantity of data. Even though the use of smartphones in travel surveys greatly reduces the response burden and allows researchers to collect large amounts of data from respondents compared to traditional paper or call-assisted travel surveys, researchers should be cautious of the quality of the data collected, particularly with regards to tagging. Untagged datasets are inherently prone to inference errors that can introduce bias in downstream travel demand models, and our study suggests that models estimated on smaller, higher-quality tagged datasets may outperform those estimated using larger untagged datasets, highlighting that quantity does not always compensate for quality issues.

It must be highlighted that these findings depend on the prediction accuracy of the trip purpose model and whether untagged trip purposes are missing at random. If the untagged trip purposes are indeed missing at random, improvements in the trip purpose model would lead to greater benefits from utilising untagged datasets in downstream models. However, if untagged trip purposes are missing not at random, meaning untagged observations differ systematically from tagged ones, then, without access to an external validation dataset, it is not possible to assess the benefits of utilising untagged datasets. Future research should therefore examine other untagged datasets or variables and consider different travel demand models, such as mode choice models, to determine whether similar results hold in other contexts. Trip purpose inference can further be improved by incorporating previous travel histories of individuals. Furthermore, adopting semi-supervised machine learning approaches for predicting trip purposes may offer a promising avenue to better leverage untagged datasets. Future studies could more rigorously evaluate the impact of using different proportions of the tagged dataset for estimating the trip purpose and daily trip generation models, and examine the extent to which the use of non-disjoint training datasets may introduce bias in the daily trip generation model. There are also some limitations of the work as indicated before, a multiple discrete continuous model has been used to model the daily number of trips and purpose even though the number of trips is on an ordinal scale rather than a continuous scale. The daily trip generation model can also be improved by assuming a more flexible modelling framework such as a latent class framework.

CRediT authorship contribution statement

Azam Ali: Writing – original draft, Visualization, Software, Methodology, Formal analysis, Conceptualization; **Ellen H. Flaata:** Writing – review & editing, Visualization, Software, Methodology, Formal analysis, Conceptualization; **Trude Tørset:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization; **Stephane Hess:** Writing – review & editing, Supervision, Software, Project administration, Funding acquisition, Conceptualization; **Charisma F. Choudhury:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Data availability

The authors do not have permission to share data.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the UK Economic Social Research Council AQM Grant Reference No ES/P000746/1. Charisma Choudhury's time was supported by the UKRI Future Leaders Fellowship Next Generation Travel Behaviour Models' [Grant Number MR/T020423/1] and Stephane Hess acknowledges support from the European Research Council through advanced Grant 101020940-SYNERGY.

The authors thank Dr. Thijs Dekker for his comments on an earlier version of the study.

Appendix A. Appendix

A.1. Data enrichment for trip purpose data

To better model and predict trip purposes, we added spatial information in the form of distance to home and the twenty closest points of interest variables. We use a spatial clustering algorithm named Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to find the home locations. DBSCAN is a widely used clustering algorithm that has been used in previous studies to model residential locations in the transport literature (Amaya et al., 2018; Lizana et al., 2023; Winkler et al., 2024). DBSCAN is useful to identify home location (or cluster) from origin or destination points detected in GPS-based surveys, as the algorithm detects clusters while ignoring outliers (or noise) normally present in GPS-based data such as cold starts or GPS drifts. The clustering algorithm was used to find the home location for individuals who had tagged their home trips. If there was only one cluster, then the centroid of that cluster was assumed to be the home location. However, if more than one cluster was found, the cluster which had the highest number of points tagged as home trips was chosen as the *primary* home location. For respondents who did not tag any trips as starting or ending in their home, the origin location of their first trip for each day was checked. If only one cluster was formed, the centroid of the cluster was considered the home location, otherwise, observations from that individual were discarded. The minimum number of points to specify a cluster was two and the distance threshold to define a cluster was set to 40 metres as these parameters produced the most consistent and plausible home locations for individuals across the sample.

Points of interest data were added based on the open access resources published by Geonorge (2023), which is the centre of spatial data for Norway. The location and land-use details for all buildings or addresses in the country are available in a cadastre dataset named *Matrikkelen Adresse Leilighetsnivå*. The land-use categories in this dataset were quite detailed and contained a total of 119 land-use types across 437,160 data points in the Trøndelag region in Norway; however, based on our judgement the land-use types were aggregated into 19 types as indicated in Table 2. Further, to identify if the trip was terminated at an open or green space, a study on the value of green spaces (*friluftslivsområder*), which contains geo-referenced green, open, and recreational areas in the country, was used.

A.2. Impact of censoring on daily trip generation model

To assess the impact of censoring in the daily trip generation model using probabilistically inferred untagged datasets, we estimate and compare models without censoring and with censoring below probability thresholds of 0.005, 0.01, 0.025, 0.05, and 0.075. To reduce computational burden (and reduce the number of combinations generated), we restrict the untagged dataset to days with at least one tagged trip in a day, reducing the total number of observations from 2,126 to 1,039 days. This makes estimation feasible as the model without censoring requires about 4.75 h (280.3 min) to estimate on a high-quality workstation, as highlighted in Table A.8.

Since the dependent variables differ across models, we rely on out-of-sample predictive performance to assess the impact of censoring. It can be observed in Table A.8 that censoring at probability values less than 0.005 has virtually no effect on model outputs with no changes in the predicted demand and MAPE; however, the estimation time decreases. Slight reductions in predictive performance appear when censoring is carried out at 0.01, though estimation time is halved. With higher censoring thresholds, the

Table A.8
Impact of censoring on daily trip generation models across alternative censoring thresholds.

Description	Tagged data with deterministically inferred untagged data	Tagged data with probabilistically inferred untagged data with					
		no censoring	censoring < 0.005	censoring < 0.01	censoring < 0.025	censoring < 0.05	censoring < 0.075
Number of individuals	593	593	593	593	593	593	593
Number of observations	3,509	3,509	3,509	3,509	3,509	3,509	3,509
Number of combinations	–	301,328	238,234	119,470	75,297	42,331	20,894
Estimation time (minutes)	2.69	280.30	205.88	132.23	91.25	45.28	21.08
LL (final)	−17,687.61	−17,751.58	−17,751.45	−17,749.92	−17,748.10	−17,740.20	−17,725.05
MAPE (%) of coefficients w.r.t no censoring	–	–	0.01	0.16	0.23	0.97	1.70
Predicted demand on testing dataset							
Work (803)	774.47	780.47	780.47	780.39	780.26	779.73	778.78
Home (1,360)	1,378.61	1,384.73	1,384.72	1,384.62	1,384.41	1,384.13	1,383.58
Shop (1,104)	1,212.11	1,184.84	1,184.86	1,185.30	1,185.96	1,188.15	1,192.27
Leisure (364)	359.73	368.53	368.54	368.55	368.49	368.04	367.23
Total trips (3,631)	3,724.92	3,718.57	3,718.59	3,718.86	3,719.12	3,720.05	3,721.86
MAPE (%)	3.69	3.12	3.12	3.13	3.14	3.17	3.23

predictive performance further decreases as censoring at probability values less than 0.05 and 0.075, increases the MAPE on the testing dataset from 3.12% to 3.17% and 3.23%, respectively. We also compare the estimated coefficients across models with respect to the model estimated with no censoring and find that censoring has a negligible impact. For censoring at a probability threshold below 0.075, the coefficients differed on average by 1.70% (MAPE) relative to the model estimated with no censoring. The comparison indicates that censoring reduces estimation time, offering computational benefits, but results in decreased predictive performance and consequently a poorer model fit. Nevertheless, even at higher censoring levels (< 0.075), predictive performance exceeds that of inferring deterministically, indicating that censoring can reduce computation burden without significantly sacrificing model fit. However, if higher levels of censoring cause predictive performance to fall below that of deterministic inference, it is preferable to use additional computational resources to obtain a better-performing model using probabilistic inference. We therefore recommend that researchers select the lowest censoring value that maintains computational feasibility.

A.3. Sensitivity analysis - daily trip generation model

To assess whether the incorporation of untagged data yields consistent effects across different folds of the dataset, we implement a four-fold cross-validation procedure. The trip purpose model (estimated using 20% of the user-tagged dataset) is held fixed, while the remaining 80% of tagged data are divided into four folds with 20% of the tagged data. In each iteration, one fold is used for testing, and the remaining three folds are used to estimate the daily trip generation models.

Table A.9 presents the results for the three remaining folds, as one fold's results are discussed in detail in Section 5.2. Across all folds, prediction error in daily trip generation is higher when tagged data is used alongside the probabilistically inferred untagged data than when the model is estimated using tagged data only. Specifically, in Fold 2, the mean absolute percentage error (MAPE) increases to 7.20% when the untagged dataset is incorporated, compared to 1.67% for the model estimated using tagged data only. In Fold 3, incorporating the untagged dataset increases the MAPE to 3.56% compared to 3.32%, while in Fold 4, the MAPE rises to 3.63% from 3.14% relative to the model estimated using tagged data only. Although the predictive performance varies across folds due to sampling differences, the main findings remain consistent.

In this section, we also investigate the potential benefit of utilising untagged data in situations where an analyst only has access to a limited amount of tagged data. In order to do this, we compare the predictive performance of the daily trip generation models estimated using varying proportions of the user-tagged dataset along with the entire probabilistically inferred untagged dataset. We only vary the proportions of the user-tagged dataset (going from 10% to eventually 100%) but retain the untagged dataset as it-is, which contains 2,126 days of data. As the share of the tagged dataset increases over these steps, we can then investigate whether the benefit of including untagged data differs depending on the relative size of the tagged sample out of the entire dataset.

It can be observed in Table A.10 that the addition of the inferred untagged dataset does not lead to an improvement in the predictive performance on the testing dataset, in fact, the predictive performance decreases substantially. For example, when 10%

Table A.9
Cross-validation results for the daily trip generation model.

Observed demand (number of trips)	Predicted demand on testing dataset		
	Tagged data only	Tagged data with deterministically inferred untagged data	Tagged data with probabilistically inferred untagged data
Fold 2 - (917 observations in testing dataset)			
Number of individuals	525	801	801
Number of observations	2,738	4,864	4,864
Work (784)	782.95	743.56	775.26
Home (1,348)	1,355.47	1,335.11	1,345.12
Shop (1,066)	1,077.12	1,228.07	1,219.87
Leisure (343)	362.11	394.00	393.65
Total (3,541)	3,577.65	3,700.74	3,733.9
MAPE (%)	1.67	8.14	7.20
Fold 3 - (911 observations in testing dataset)			
Number of individuals	529	801	801
Number of observations	2,744	4,870	4,870
Work (806)	789.67	752.75	783.2
Home (1,337)	1,337.01	1,327.62	1,348.79
Shop (1,091)	1,060.48	1,214.99	1,206.02
Leisure (374)	338.52	370.04	371.26
Total (3,608)	3,525.68	3,665.40	3,709.27
MAPE (%)	3.32	4.26	3.56
Fold 4 - (919 observations in testing dataset)			
Number of individuals	528	796	796
Number of observations	2,736	4,862	4,862
Work (798)	779.86	744.27	774.27
Home (1,385)	1,352.61	1,338.68	1,348.14
Shop (1,106)	1,070.77	1,221.53	1,211.81
Leisure (388)	368.45	380.31	380.56
Total (3,677)	3,571.69	3,684.79	3,714.78
MAPE (%)	3.14	4.54	3.63

of the user-tagged dataset is used, which has 274 observations, the mean absolute percentage error (MAPE) is 6.82%. However, the predictive performance deteriorates to 16.23% when we also make use of the inferred untagged dataset in addition to the 10% user-tagged dataset, which comprises of 2400 observations. We find similar cases with other proportions of tagged datasets, i.e. the predictive performance of the models developed using the tagged dataset only is better than that of the models that also make use of the *inferred* untagged datasets. One thing to note in the models developed using the user-tagged dataset only is that with an increase in the number of observations, the predictive performance either improves or does not worsen substantially, as expected.

We also assessed the benefits of utilising subsets of the untagged datasets that are more likely to be of a higher quality or better prediction accuracy. We use two metrics to select higher quality untagged dataset by subsetting i) days with a low proportion of untagged trips (less than 20% and 40% of the total trips in day), and ii) days with low normalised entropy (less than 0.20 and 0.40). Normalised entropy quantifies the uncertainty in inferred observations that range from 0 to 1, where values close to 0 indicate lower uncertainty and vice versa. To calculate the normalised entropy, we use the expression, $\sum_i^n \frac{p_i \cdot \log(p_i)}{\log(n)}$, where n is the total number of combinations in the probabilistically inferred observation, and p_i is the probability or weight for observing combination i. We assess the benefits of utilising these higher quality subsets of untagged datasets by adding them to a base model that uses 10% of user-tagged data, which has a MAPE of 6.82% and has 274 observations or days of data. Adding low-entropy untagged data (entropy < 0.2 and < 0.4) increases the sample size from 274 days to 376 and 770 days, but the MAPE decreases from 6.82% to 11.29% and 18.70%, respectively. Similarly, including days of data with a lower proportion of untagged trips ($<20\%$ and $<40\%$ in a day) increases the sample size to 383 and 626 days; however, the predictive performance decreases to 18.22% and 21.57%, respectively. Overall, incorporating even higher-quality untagged data reduces the predictive performance compared to the model estimated on tagged data alone.

The results primarily suggest that the parameter estimates for the daily trip generation models are different for the tagged and untagged datasets. It can be argued that the inferred untagged dataset is of poor quality, i.e. not necessarily similar to the ground truth, which does not lead to any improvement in the predictive performance of the model (implying an addition of noise). Similarly, it can be argued that that there are behavioural differences between tagged and untagged datasets. As the testing dataset is from the same subset as the user-tagged dataset. It is therefore intuitive that the model developed using the user-tagged dataset only leads to a better fit on our defined testing dataset. However, this makes it difficult to determine which model actually converges to the true or population-level parameter values.

We also compared the coefficients of these models. In the case of the models developed using 10% and 20% user-tagged data only, there were some coefficients whose signs were not intuitive; but utilising the untagged datasets changed the signs to the correct ones (i.e., more intuitive and similar to the models estimated on the complete user-tagged dataset). This indicates that the behavioural findings of the model developed using the additional untagged datasets are more intuitive compared to the models developed using the tagged dataset only. For other cases, i.e. the models estimated on 40%, 60%, and 80% of the user-tagged dataset, we observed that adding the untagged dataset leads to an increase in the efficiency for some variables. However, we cannot fully assert whether the increase in efficiency is from more observations or imputation errors (such as type-I error) as previously highlighted. Therefore, there could be some potential benefits of utilising untagged datasets which must be evaluated on a case-by-case basis. Nonetheless, our study aims to make researchers aware of the issues in utilising *inferred* untagged datasets.

Table A.10
Predictive performance of the daily trip generation models under varying proportions of tagged dataset.

Description	Number of individuals	Number of observations	Predicted demand on testing dataset (number of trips)					MAPE(%)
			Work (803[a])	Home (1,360[a])	Shop (1,104[a])	Leisure (364[a])	Total (3,631[a])	
10% of user-tagged data only	47	274	661.08	1,412.64	1,179.47	384.19	3,637.38	6.82
10% of user-tagged & probabilistically inferred untagged data	547	2,400	774.56	1,344.42	1,474.59	476.18	4,069.75	16.23
20% of user-tagged data only	102	548	727.85	1,409.54	1,058.67	342.37	3,538.43	5.12
20% of of user-tagged & probabilistically inferred untagged data	576	2,674	763.95	1,352.38	1,401.31	445.19	3,962.83	12.76
40% of user-tagged data only	210	1,096	746.59	1,356.16	1,087.30	349.42	3,539.47	3.07
40% of user-tagged & probabilistically inferred untagged data	635	3,222	759.08	1,343.55	1,327.69	416.53	3,846.85	9.46
60% of user-tagged data only	315	1,644	774.47	1,368.34	1,057.67	358.88	3,559.36	2.35
60% of user-tagged & probabilistically inferred untagged data	693	3,770	771.83	1,352.51	1,267.91	406.74	3,798.99	7.13
80% of user-tagged data only	409	2,192	753.18	1,359.46	1,058.45	351.99	3,523.08	3.33
80% of user-tagged & probabilistically inferred untagged data	745	4,318	759.27	1,349.11	1,234.07	390.32	3,732.77	5.61
100% of user-tagged data only	521	2,740	768.64	1,353.50	1,066.21	348.75	3,537.10	2.99
100% of user-tagged & probabilistically inferred untagged data	798	4,866	767.85	1,347.48	1211.85	381.63	3,708.81	4.41

[a] observed demand in testing dataset.

References

Ali, A., Hess, S., Dekker, T., Choudhury, C.F., 2024. Using posterior analysis to predict missing information in passively collected data sources. presented at the 12th Symposium of the European Association for Research in Transportation.

Ali, A., Kalatian, A., Choudhury, C.F., 2023. Comparing and contrasting choice model and machine learning techniques in the context of vehicle ownership decisions. Transp. Res. Part A: Plicy Pract. 35, 103727.

Amaya, M., Cruzat, R., Munizaga, M.A., 2018. Estimating the residence zone of frequent public transport users to make travel pattern and time use analysis. J. Transp. Geogr. 66, 330–339.

Andersson, A., Engelson, L., Börjesson, M., Daly, A., Kristoffersson, I., 2022. Long-distance mode choice model estimation using mobile phone network data. J. Choice Model. 42, 100337.

Díaz, F.D., Cantillo, V., Arellana, J., Ortúzar, J. D.D., 2015. Accounting for stochastic variables in discrete choice models. Transp. Res. Part B Methodol. 78, 222–237.

Bhaduri, E., Manoj, B., Wadud, Z., Goswami, A.K., Choudhury, C.F., 2020. Modelling the effects of covid-19 on travel mode choice behaviour in india. Transp. Res. Interdiscip. Perspect. 8, 100273.

Bhat, C.R., Guo, J.Y., Srinivasan, S., Sivakumar, A., 2004. Comprehensive econometric microsimulator for daily activity-travel patterns. Transp. Res. Record J. Transp. Research Board 1894 (1), 57–66.

Biswas, M., Bhat, C.R., Ghosh, S., Pinjari, A.R., 2024. Choice models with stochastic variables and random coefficients. J. Choice Model. 51, 100488.

Bohte, W., Maat, K., 2009. Deriving and validating trip purposes and travel modes for multi-day gps-based travel surveys: A large-scale application in the netherlands. Transp. Res. Part C Emerging Technol. 17 (3), 285–297.

Bowman, J., Ben-Akiva, M., 2001. Activity-based disaggregate travel demand model system with activity schedules. Transp. Res. Part A: Plicy Pract. 35 (1), 1–28.

Bwambale, A., Choudhury, C.F., Hess, S., Iqbal, M.S., 2021. Getting the best of both worlds: a framework for combining disaggregate travel survey data and aggregate mobile phone data for trip generation modelling. Transportation 48 (5), 2287–2314.

Calastri, C., Sourd, R. C.D., Hess, S., 2020. We want it all: experiences from a survey seeking to capture social network structures, lifetime events and short-term travel and activity planning. Transportation 47 (1), 175–201.

Chawla, N.V., Japkowicz, N., Kotcz, A., 2004. Editorial: special issue on learning from imbalanced data sets. ACM SIGKDD Explorations Newsl. 6 (1), 1–6.

Chen, C., Ma, J., Susilo, Y., Liu, Y., Wang, M., 2016. The Promises of Big Data and Small Data for Travel Behavior (aka Human Mobility) Analysis. Vol. 68. Transportation research part C: emerging technologies.

Cottrill, C.D., Pereira, F.C., Zhao, F., Dias, I.F., Lim, H.B., Ben-Akiva, M.E., Zegras, P.C., 2013. Future mobility survey: Experience in developing a smartphone-based travel survey in singapore. Transp. Res. Rec. J. Transp. Res. Board 2354 (1), 59–67.

Cranenburgh, S.V., Wang, S., Vij, A., Pereira, F., Walker, J., 2022. Choice modelling in the age of machine learning - discussion paper. J. Choice Model. 42, 100340.

Cui, Y., Meng, C., He, Q., Gao, J., 2018. Forecasting current and next trip purpose with social media data and google places. Transp. Res. Part C Emerging Technol. 97, 159–174.

Deng, Z., Ji, M., 2010. Deriving rules for trip purpose identification from gps travel survey data and land use data: a machine learning approach. In: Traffic and Transportation Studies 2010. American Society of Civil Engineers, pp. 768–777.

Ermagun, A., Fan, Y., Wolfson, J., Adomavicius, G., Das, K., 2017. Real-time trip purpose prediction using online location-based search and discovery services. Transp. Res. Part C Emerging Technol. 77, 96–112.

Flaata, E.H., 2025. Tracking National Travel Surveys: A Thesis on the Implications of Transitioning to App-Based Methods. Technical Report. Norwegian University of Science and Technology.

Gao, Q., Molloy, J., Axhausen, K.W., 2021. Trip purpose imputation using gps trajectories with machine learning. ISPRS Int. J. Geo-Inf. 10 (11), 775.

Geonorge, 2023. Geonorge - The Map Catalogue. https://www.geonorge.no/.

Gong, L., Kanamori, R., Yamamoto, T., 2018. Data selection in machine learning for identifying trip purposes and travel modes from longitudinal gps data collection lasting for seasons. Travel Behav. Soc. 11, 131–140.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press.

Harding, C., Imani, A.F., Srikukenthiran, S., et al. 2021. Are we there yet? Assessing smartphone apps as full-fledged tools for activity-travel surveys. Transportation 48, 2433–2460. https://doi.org/10.1007/s11116-020-10135-7

Heinonen, S., Freitas, L. M.D., Meister, A., 2024. The e-biking in switzerland (ebis) study: methods and dataset. Transportation, 48, 1–24. https://doi.org/10.1007/s11116-024-10552-y

Hess, S., Palma, D., 2019. Apollo: a flexible, powerful and customisable freeware package for choice model estimation and application. J. Choice Model. 32, 100170.

Hossain, S., Habib, K.N., 2021. Inferring the purposes of using ride-hailing services through data fusion of trip trajectories, secondary travel surveys, and land use data. Transp. Res. Rec. J. Transp. Res. Board 2675 (9), 558–573.

Imani, A.F., Harding, C., Srikukenthiran, S., Miller, E.J., Habib, N., K, 2020. Lessons from a large-scale experiment on the use of smartphone apps to collect travel diary data: The “city logger” for the greater golden horseshoe area. Transp. Res. Rec. J. Transp. Res. Board 2674 (7), 299–311.

Lee, R.J., Sener, I.N., Mullins, I., J, A., 2016. An evaluation of emerging data collection technologies for travel demand modeling: from research to practice. Transp. Lett. 8 (4), 181–193.

Liu, Y., Miller, E.J., Habib, K.N., 2023. Inferring trip destination purposes for trip records collected through smartphone apps. J. Transp. Eng. Part A Syst. 149 (2), 17.

Lizana, M., Choudhury, C., Watling, D., 2023. Using smart card data to model public transport user profiles in light of the covid-19 pandemic. Travel Behav. Soc. 33, 100620.

Lu, Y., Zhu, S., Zhang, L., 2012. A machine learning approach to trip purpose imputation in GPS-based travel surveys. In: 4th Conference on Innovations in Travel Modeling.

Marra, A.D., Becker, H., Axhausen, K.W., Corman, F., 2019. Developing a passive gps tracking system to study long-term travel behavior. Transp. Res. Part C Emerging Technol. 104, 348–368.

Meister, A., Bashan, N.F., Basu, R., 2025. The bostonwalks study: a longitudinal travel survey using smartphone tracking. Transportation 52, 2249–2279. https://doi.org/10.1007/s11116-025-10637-2

Molloy, J., Castro, A., Götschi, T., Schoeman, B., Tchervenkov, C., Tomic, U., Hintermann, B., Axhausen, K.W., 2023. The mobis dataset: a large gps dataset of mobility behaviour in switzerland. Transportation 50 (5), 1983–2007.

Montini, L., Rieser-Schüssler, N., Horni, A., Axhausen, K.W., 2014. Trip purpose identification from GPS tracks. Transp. Res. Rec. J. Transp. Res. Board 2405 (1), 16–23.

Mukherjee, J., Kadali, B.R., 2022. A comprehensive review of trip generation models based on land use characteristics. Transp. Res. Part D Transp. Environ. 109, 103340.

Oliveira, M. G.S., Vovsha, P., Wolf, J., Mitchell, M., 2014. Evaluation of two methods for identifying trip purpose in gps-based household travel surveys. Transp. Res. Rec. J. Transp. Res. Board 2405 1, 33–41.

Palma, D., Hess, S., 2022. Extending the multiple discrete continuous (mdc) modelling framework to consider complementarity, substitution, and an unobserved budget. Transp. Res. Part B Methodol. 161, 13–35.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.

Tørset, T.T., Svaboe, G. B.A., 2020. Show case - using travelvu and travelviewer in trondheim. https://www.ntnu.no/documents/1283830479/1283895465/Show+case+Trondheim.pdf/6107fca9-55b5-ef02-9fb4-2f2dbe1ed994?t=1663068717725.

Sanko, N., Hess, S., Dumont, J., Daly, A., 2014. Contrasting imputation with a latent variable approach to dealing with missing income in choice models. J. Choice Model. 12, 47–57.

Servizi, V., Pereira, F.C., Anderson, M.K., Nielsen, O.A., 2021. Mining user behaviour from smartphone data: a literature review. Eur. Transp. Res. Rev. 13 (1), 57.

Shen, L., Stopher, P.R., 2013. A process for trip purpose imputation from global positioning system data. Transp. Res. Part C Emerging Technol. 36, 261–267.

Train, K.E., 2009. Discrete Choice Methods with Simulation. Cambridge University Press. 2nd edition edition.

Trivector Traffic, 2024. TravelVu - a New Way to Perform Travel Surv. https://en.trivectortraffic.se/software/travelvu.

Vallejo-Borda, J.A., Bhaduri, E., Ortiz-Ramirez, H.A., Arellana, J., Choudhury, C.F., Rodriguez-Valencia, A., Wadud, Z., Goswami, A.K., 2023. Modeling the covid-19 travel choices in colombia and india: A hybrid multiple discrete-continuous nested extreme value approach. Transp. Res. Rec. J. Transp. Res. Board 2677 (4), 778–801.

Vij, A., Shankari, K., 2015. When is big data big enough? implications of using gps-based surveys for travel demand analysis. Transp. Res. Part C Emerging Technol. 56, 446–462.

Wang, S., Mo, B., Zheng, Y., Hess, S., Zhao, J., 2024a. Comparing hundreds of machine learning and discrete choice models for travel demand modeling: an empirical benchmark. Transp. Res. Part B Methodol. 190, 103061.

Wang, S., Wang, Q., Zhao, J., 2020. Deep neural networks for choice analysis: extracting complete economic information for interpretation. Transp. Res. Part C Emerging Technol. 118, 102701.

Wang, Y., Choudhury, C., Hancock, T.O., Wang, Y., Ortúzar, J. D.D., 2024b. Influence of perceived risk on travel mode choice during covid-19. Transp. Policy 148, 181–191.

Winkler, C., Meister, A., Axhausen, K.W., 2024. The timeuse data set: 4 weeks of time use and expenditure data based on GPS tracks. Transportation, pp. 1–27. https://doi.org/10.1007/s11116-024-10517-1

Xiao, G., Juan, Z., Zhang, C., 2016. Detecting trip purposes from smartphone-based travel surveys with artificial neural networks and particle swarm optimization. Transp. Res. Part C Emerging Technol. 71, 447–463.

Yazdizadeh, A., Patterson, Z., Farooq, B., 2019. An automated approach from GPS traces to complete trip information. Int. J. Transp. Sci. Technol. 8 (1), 82–100.

Zhao, X., Yan, X., Yu, A., Hentenryck, P.V., 2020. Prediction and behavioral analysis of travel mode choice: a comparison of machine learning and logit models. Travel Behav. Soc. 20, 22–35.