Capturing Heterogeneity in Mode Choice Decisions: Comparing and Combining Machine Learning and Discrete Choice Models

Panagiotis Tsoleridis
Institute for Transport Studies
University of Leeds, Leeds, United Kingdom, LS2 9JT
Email: ts17pt@leeds.ac.uk

Charisma F. Choudhury
Institute for Transport Studies
University of Leeds, Leeds, United Kingdom, LS2 9JT
Email: C.F.Choudhury@leeds.ac.uk

Stephane Hess
Institute for Transport Studies
University of Leeds, Leeds, United Kingdom, LS2 9JT
Email: S.Hess@leeds.ac.uk

ABSTRACT

In the era of big data, machine learning (ML) has emerged as a strong competitor of econometric modelling, discrete choice models (DCMs) in particular. However, a key limitation of the purely data driven models is the lack of valuation measures – for example the difficulty in calculation of value of time to feed into the cost-benefit analyses. The current study focuses on combining ML-based segmentation approaches with DCM and tests their performances against latent class choice models (LCCM). LCCM allows the simultaneous allocation of individuals to a predefined number of classes alongside the estimation of their mode choice behavior. For the combined ML and DCM, this is done as a two-stage approach, where individuals are first allocated into clusters using a deterministic K-modes/K-means algorithm and then a mode choice model is estimated per cluster. In both cases, the underlying mode choice component is specified as a multinomial logit model. The dataset utilized in the current study includes the trips of 540 unique individuals, 12524 in total, captured through GPS traces collected using a smartphone app. The results suggest that LCCM performs significantly better than both variants of the ML counterpart in terms of model fit. It also has clearer insights about the segment compositions. The latter, however, results in a much faster estimation and produces reasonable Value of Time estimates and demand elasticities. The findings are expected to provide guidance to researchers and practitioners in choosing the most efficient method to capture the taste heterogeneity among the segments of the travelers.

Keywords: Unsupervised learning, Clustering, K-modes algorithm, Latent classes, mode choice modelling

1. INTRODUCTION

During the last decade the abundance of data has provided not only the potential for further research advancements, but also posed challenges as to how to derive value out of those emerging data sources. In the field of transport, specifically, different forms of data, such as GPS traces, Call Detail Records, geotagged tweets etc. require significant pre-processing and knowledge that transcends into different fields of research (e.g. computer science) (Antoniou et al., 2019). The massive size of the data has also led to increase in the popularity of Machine Learning (ML) techniques over traditional econometric models (Discrete Choice Modelling (DCM) techniques for instance). This has led to the necessity of comparing and contrasting ML with traditional DCM.[1]

Originating from the field of Computer Science, ML algorithms are generally characterised as non-parametric methods (with some exceptions) aiming to minimise the errors between actual and predicted outcomes without relying on any behavioural assumptions of the underlying model. ML encompasses a large array of algorithms, which can be broadly categorised into supervised and unsupervised learning. The majority of comparative studies in the literature between ML and DCM focus on comparing supervised ML algorithms, e.g. Artificial Neural Networks and Random Forests, with DCM frameworks, such as MNL and Nested Logit, usually in the context of mode choice (Hensher and Ton, 2000; Xie et al., 2003; Cantarella and de Luca, 2005; Zhang and Xie, 2008; Sekhar et al., 2016; Hagenauer and Helbich, 2017). Their findings suggest that ML algorithms have the potential to be used as an alternative method for behavioural modelling due to their superior predictive performance, although Hensher and Ton (2000) also highlight the limitations associated with the lack of interpretable results compared to a DCM framework. Furthermore, a wide range of studies has implemented clustering algorithms to analyse individual behaviour and uncover mobility patterns (see Anda et al. 2017 for details). Though such studies provide good insights about the state of the network, they have limited applications in the context of predictions and/or valuation (e.g. calculation of value of time to feed into the cost-benefit analyses).

A key advantage of an ML-based approach is the efficiency in capturing the patterns in the data which can then be used in segmenting the travelers and capturing the taste heterogeneity in the sample. Several studies have used clustering techniques for market/sample segmentation (Salomon and Ben-Akiva, 1983; Lanzendorf, 2002; Krizek and Waddell, 2003) and reported that different lifestyle clusters (identified empirically) have different choice elasticities. However, these studies have three key limitations. *Firstly,* the previous studies relied on "*traditional*" samples with regard to their data collection methods (e.g. single RP choice scenarios, short trip diaries, etc.) and it is worth investigating the performance of similar approaches with passively collected larger samples (more participants and/or longer panels). *Secondly,* these studies did not compare the model performances with the more advanced discrete choice models that account for heterogeneity among groups of decision makers - Latent Class Choice Modelling (LCCM) for example. *Thirdly,* the studies compared the goodness of fit and/or prediction capabilities of ML and DCM as opposed to in-depth efforts to formulate models that has the best of both worlds – computational advantages on ML and the behavioral interpretation of DCM that can be used for valuation.

The current research aims to address this research gap by comparing the performance of different sample segmentation techniques with a discrete choice model as the base for the actual choice component. The aim is to find the best method to capture the heterogeneity among the travelers and simultaneously have outputs that can be used for valuation. The following variants of LCCM and hybrid models (ML combined with DCM) have been tested in this regard:

Model 1: LCCM where the probabilistic sample segmentation (class membership) and mode choices are modelled simultaneously.

Model 2: ML based segmentation using socio-demographic data only combined with DCM for the mode choice component

Model 3: ML based segmentation using socio-demographic and choice data combined with DCM for the mode choice component

For all models, the *a priori* assumption is that there is a discrete number of segments (classes/clusters); the taste heterogeneity is constant within the same segment but varies among different segments. While Model 1 provides a behavioural model specification, where both the individual sociodemographic characteristics and the observed mobility choices are taken into consideration, Models 2 and 3 are expected to offer much faster and potentially more efficient segmentation.

The dataset used in the current study is a combination of sociodemographic features and a trip diary captured through GPS tracking using a smartphone app (56693 trips from 721 unique individuals in the raw dataset). Therefore, the dataset provides a rich level of semantic information in addition to capturing a wider spectrum of trips and activities from a set of individuals. Though such passive data collection methods have the potential to lead to more accurate estimated parameters of travel behaviour, there are challenges associated with deriving the full set of variables that are important from a behavioural modelling perspective (e.g. accurate travel times, costs etc.) and defining the choice sets, specifically determining the availabilities for the unchosen alternatives. In addition to the contribution regarding the comparison of ML- and DCM-based segmentation, the current study also demonstrates how these data limitations can be overcome.

The remainder of this paper is structured as follows. In Sections 2 and 3, the methodological framework and the data used for the study's practical implementation are described, respectively. Section 4 focuses on the results with focus on the comparison between the different approaches. The main conclusions and the direction of future research are summarized in the concluding section.

2. METHODOLOGY

The methodological framework developed for this study involves the implementation of probabilistic and deterministic sample segmentation with the use of traditional Discrete Choice Modelling (DCM) techniques and Machine Learning (ML) algorithms, respectively. For the former (Model 1), Latent Class Choice Modelling (LCCM) is used to simultaneously allocate individuals probabilistically into a discrete number of classes, based on selected sociodemographic features, and estimate a choice model regarding their mode choice behaviour based on level of service attributes. For the latter, a two-stage approach is performed by first segmenting the sample deterministically using K-modes/K-means clustering and then applying a DCM framework to each cluster separately. The inputs in the K-modes clustering vary between Models 2 and 3 with the clustering using only the sociodemographic data and the combination of sociodemographic and choice data, respectively.

2.1 Latent Class Choice Modelling (Model 1)

Discrete Choice Modelling (DCM) has been used extensively in the field of transport, since the seminal study of McFadden (1973), inspired by the previous work of Luce (1959) and Marschak (1960) linking behavioural choice theory with microeconomics. As described in McFadden (2000), the main concept of Random Utility Theory and its basic Multinomial Logit model (MNL) had been extended significantly in the following decades leading to the formulation of more behaviourally accurate model representations of individual behaviour.

DCM is based on the underlying theory of Random Utility Maximisation (RUM) suggesting that a rational individual n is more likely to choose among a set of possible alternatives J the one resulting in the maximisation of utility U. The utility $U_{n,j}$ is comprised by a deterministic and a stochastic part. An additive form of the Utility function is shown in the following equation (**Equation 1**) (Train, 2009):

$$U_{n,j}=V_{n,j}+\varepsilon_{n,j}\forall j\in J \tag{1}$$

where $V_{n,j}$ and $\varepsilon_{n,j}$ are the deterministic and stochastic utilities, respectively, for individual n and alternative j. The *deterministic* or *systematic part* of the utility $V_{n,j}$ is usually formed as a linear-in-the-parameters function (**Equation 2**) of a vector $X_{n,j}$ containing the observed alternative's j attributes and the individual's n characteristics, and their respective taste coefficients $\beta_{n,j}$ including an inherent preference towards alternative j as $\beta_{0,j}$, also known as the Alternative Specific Constant (ASC). The *stochastic part* or the *error term* $\varepsilon_{n,j}$ is considered to include all the uncaptured features that could influence the utility of alternative j for individual n, but are currently unknown to the researcher

$$V_{n,j}=\beta_{0,j}+\beta_{n,j}X_{n,j} \tag{2}$$

LCCM is an extension of DCM which uses an endogenous sample segmentation technique to capture the heterogeneity in the sample (Kamakura and Russell, 1989). The basic premise of LCCM is that the individuals in the sample can be allocated into a discrete number of classes based on their sociodemographic features and their observed behaviour. The LCCM framework includes the estimation of a class allocation model, where individuals are allocated into a pre-specified number of classes, and a choice model aiming to explain the behaviour of individuals given their class (Kamakura and Russell, 1989; Bhat, 1997; Hess, 2014). Under that framework, the likelihood of individual n is calculated as:

$$L_n=\sum_{s=1}^{S}\pi_{n,s}¿¿ \tag{3}$$

where $\pi_{n,s}$ is the probability that individual n belongs to class s from a total number of S classes and $P_{j_{n,t}}\left(\beta_s\right)$ is the probability that individual n chooses alternative j in choice task t given that he/she belongs to class s. The underlying choice model is usually specified as an MNL model, but extensions to more advanced modelling specifications are also possible. Under that framework, the taste coefficients β_s are the same for each individual in a specific class.

2.2 ML-based sample segmentation techniques (Models 2 & 3)

ML unsupervised learning contains a range of algorithms capable of effectively segmenting the sample into a discrete number of clusters based on patterns in the dataset. Clustering algorithms can be categorised into:

- Centroid-based, such as the "hard" and "soft" K-means/modes algorithms
- Connectivity-based, such as Hierarchical clustering

- Density-based, such as the DBSCAN and OPTICS algorithms
- Distribution-based, such as the Gaussian Mixture Models

K-means (MacQueen, 1967) and K-modes (Aranganayagi and Thangavel, 2009) are two of the most widely-known clustering algorithms that can be applied to continuous and discrete data, respectively. As most clustering algorithms, they are based on a distance or (dis-)similarity measure between the data points in a dataset. Points with high similarity between them are allocated in the same cluster. The points within the same cluster should also be distinguishable from points in other clusters. Traditional K-means and K-modes perform a form of *"hard"* clustering as each point is assigned to a unique cluster. Contrary to that the more generalized *"soft"* or *"fuzzy"* K-modes algorithm (Kim et al., 2004) assigns a weight or probability to each point and for each cluster. Therefore, each point has a non-zero probability to belong to any cluster, similar to LCCM. In the current study, *"hard"* K-means and K-modes algorithms are implemented and compared with LCCM.

As a starting point, the K-means/K-modes algorithms choose a predetermined number of random points in the multidimensional data space as the initial cluster centroids and the similarities from all data points to those centroids are calculated. The points are allocated to their closest centroid and then a new cluster centroid is calculated as the mean/mode of all the data points in the cluster. At the next step, new similarities are calculated again from all the data points to all the centroids and the algorithm continuous iteratively until no more changes occur in the cluster centroids. The underlying purpose of the algorithm is to minimise the within-cluster-sum-of-differences (WCSD) and at the same time maximise the between-cluster-sum-of- differences (BCSD). That means that points within the same cluster would have the greatest degree of similarity, while also being as dissimilar as possible from points in different clusters.

3. DATA

The data used for the practical implementation of the current study was collected as part of the research project DECISIONS carried out by the Choice Modelling Centre at ITS Leeds, between October 2016-March 2017. Several submodules are included in the dataset as part of the survey capturing various aspects of the participants' mobility behaviour, in-home and out-of-home activities, their daily energy usage and their social network using a name generator. In addition, the project included a household survey module capturing the participants' most important socio-demographic information. The dataset and its submodules is thoroughly detailed in Calastri et al. (2018a). In regard to mobility behaviour, the individuals' daily trips were captured through GPS tracking using a smartphone app, while the participants also had to tag their completed trips with further information regarding their trip purpose and mode of transport. As a result, the "DECISIONS" dataset provides a combination of emerging and traditional data sources.

3.1 Data cleaning

Two separate datasets are used for the "DECISIONS" project, one corresponding to the daily trip diary captured through GPS tracking and the other for the rest of the project's submodules including the socio-demographic component. The former was acquired in the form of an SQL-database file, while the latter as an excel sheet. Initially, 56693 trips were included in the raw database file performed by 721 unique individuals. These trips had to be combined with the sociodemographic file using a unique trip identifier number. The pre-processing analysis involved the identification of trip errors in the GPS file.

More specifically, the trips' timestamps were checked to detect trips starting at the same time or before the end of the previous trip resulting in zero or negative activity durations.

At the next stage of data cleaning, trip purposes were examined. The trip purposes, reported by the participants, among others included the purpose of "*Change of travel mode*". It was decided to merge those journeys together with the following trips, since the subsequent mode choice analysis would focus solely on the main mode and not on the access mode. Furthermore, trips with no additional information on mode and purpose (untagged trips) were removed and the geographical scope was limited to the region of Yorkshire. This led to a cleaned dataset with 38624 observations from 721 respondents. The mode alternatives selected for the mode choice analysis were car, bus, rail, taxi, cycling and walking.

3.2 Data enrichment

Due to the nature of the data at hand, certain important variables from a behavioural modelling perspective were missing (travel cost and travel time for non-chosen alternatives). For that reason, a data enrichment process was performed with different components summarised in the following paragraphs.

3.2.1 Travel time/distance estimation

In the current study, real network travel times and distances for each origin-destination pair and for each mode were derived using the Google "*Directions*" API (https://developers.google.com/maps/documentation/directions/intro). The API was implemented to estimate the travel times/distances both for the chosen and the non-chosen alternatives, since using the stated time/distance for the chosen mode and the network travel times/distances for the non-chosen ones would have the risk of producing biased estimates (Calastri et al., 2018b). The API provided travel times based on the stated time of day, hence the state of the network was taken into consideration. Furthermore, travel times per link segment were obtained, instead of a total travel time for the whole trip, with the purpose of more accurately estimating fuel consumption.

3.2.2 Cost estimation

For car trips, travel cost was segmented into fuel/operating costs and parking costs. The first two were calculated using WebTAG's specifications (Department for Transport, 2014). Taking advantage of the per segment travel times and distances derived from the API, car travel cost was initially calculated for each trip segment and further aggregated per trip. Location-specific parking costs were defined using average parking prices per hour for specific places in the region of Yorkshire, such as CBDs, rail stations and the airport. For all other locations, it was assumed that there was no parking cost involved.

For taxi, the cost was calculated as a combination of the initial charge (fixed price), the average kilometre cost (pounds per kilometre) and the average time cost (pounds per minute) taking into account the different tariffs during day, night and weekend. Relevant information on taxi costs for Yorkshire was found only for the cities of Leeds and Sheffield. Separate taxi travel costs (distance and time costs) were calculated using the average prices for Leeds and Sheffield for each trip. The final taxi travel cost was calculated as the average of the distance and time cost per trip.

For bus and rail, two different cost calculations were performed. The first one took into account the discounted trip cost for season ticket holders, while for the remaining individuals a fixed fare price was applied. For the second calculation, only a fixed fare was applied to each participant regardless of their possession of a season ticket. The results of both cost calculations were used at the initial model specification stage (MNL estimation) and the respective model results were compared with values suggested by WEBTag (Value of Time, bus fare elasticity) for validation purposes. The first approach (separate costs for season and non-season ticket holders) was the one that yielded results closer to WEBTag's values and was selected for the subsequent more advanced model specifications.

3.2.3 Choice set generation

In the current study a tour-based approach is followed in contrast to a trip-based one, in which a tour is defined as a sequence of trips starting and finishing at the same location (e.g. the individual's home location) (as used by Miller et al., 2005; Hasnine and Habib, 2019). Although the modelling framework in the current study does not take into consideration the dynamics inside a tour (activity schedule, next location etc.), the implemented tour-based approach helps to specify more behaviourally accurate mode availability assumptions. Whether or not a mode is included in the choice set hence depends on the general availability of the mode (person-specific), consideration of the mode (trip-specific) (Calastri et al., 2018b) and feasibility of using the mode (restrictions imposed by earlier choices made in the same tour) .

For person-specific mode availability, the participants' stated availability for car and cycling was taken into consideration, while the remaining alternatives (taxi, bus, rail and walking) were considered to be available for everyone.

In regard to trip-specific mode consideration, different feasibility assumptions per mode were considered taking advantage of the information derived from the Google API. Walking and cycling are considered for trips less than 3 km and 20km respectively (the maximum distance of trips where each have been chosen in the full dataset). Car is considered for trips of more than 50m distance to avoid the inclusion of short insignificant trips with close to 0 min travel time. Taxi is considered for trips of more than 50m and less than 81km (the maximum distance where taxi is the chosen mode in the full dataset). Bus is not considered for very short trips (where Google API suggests only walking segments) and only considered for trips with less than 3 transfers and with distance more than the minimum and less than the maximum distance of trips where bus is the chosen mode. Rail is not considered for very short trips (where Google API suggests walking only) and only considered for trips with less than 2 transfers and with distance more than the minimum distance of trips where rail is the chosen mode. Rail trips were allowed to include bus segments, since bus can be considered as an access/egress mode for rail, but that was not the case for the opposite scenario, i.e. using rail as an access/egress mode for bus .

In terms of restrictions imposed by the earlier modes used in the tour, if car is the chosen mode for the outbound trip, then the driver has no choice but to return the car to the starting location (home) during the last trip of the tour making the rest of the alternatives unavailable. For that purpose, when car is chosen for the first trip of a tour, the remaining alternatives are available for the remaining trips of the tour except from the returning trip to home (**Figure 2**). Due to this constraint, those returning trips were not included for mode choice modelling, since there is not any actual choice involved. On the contrary, when bus, rail, taxi, cycling or walking is chosen for the first trip of the tour, car is available only for the first trip, but not for the remaining trips of the tour (**Figure 2**). Finally, after considering all the availability/consideration assumptions, the trips with only one available alternative were removed from the following mode choice modelling.

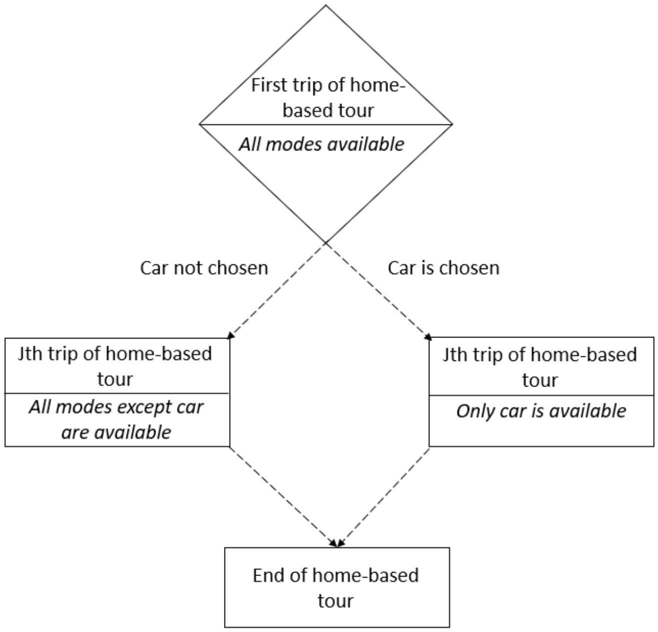

First trip of home-based tour
All modes available
Car not chosen
Car is chosen
Jth trip of home-based tour
All modes except car are available
Jth trip of home-based tour
Only car is available
End of home-based tour

Figure :2 Tour-based mode availability

3.3 Descriptive statistics

The final dataset consists of 12524 trips made by 540 unique individuals. In **Figure 3**, the mode share and the OD pairs are depicted both for the trips starting and ending inside the region of Yorkshire and those inside the Local Authority of Leeds. In general, the majority of trips start and finish inside Leeds and most of them are between Leeds and the neighbouring city of Bradford. In **Table 1** the descriptive statistics of the sociodemographic features included in the dataset and the level of service attributes per available/considered mode, derived from the Google API, are presented. The sample includes more females than males and more individuals of higher education (from undergraduate level and above). Regarding the level of service attributes, the travel times, costs and distances are consistent with the expected values indicating an accurate estimation of those variables. Rail has the highest travel time and distance as it is considered a longer-distance mode compared to the rest and has the highest cost, since rail tickets are more expensive. Cycling and walking have the lowest average distance due to the require human effort. Bus travel distance is the lowest compared to the rest of mechanised modes, an indication that travelling longer distances by bus produces generally more discomfort than in the remaining alternatives. Taxi trips despite using the same network and routes as car trips have a lower travel time and distance, since the cost is significantly more compared to car.

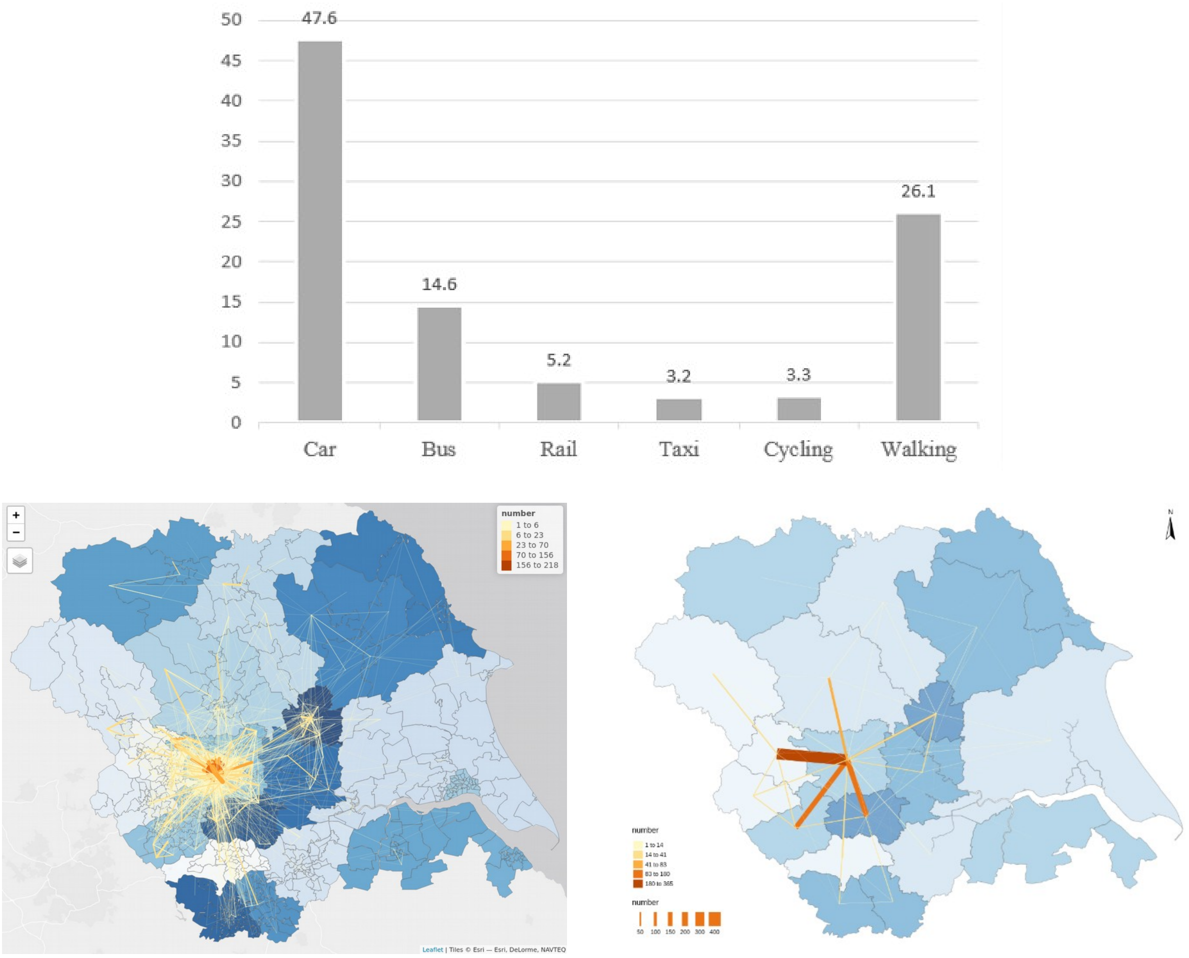


Figure 3: Mode share and OD pairs for trips inside the region of Yorkshire and the local authority of Leeds

TABLE 1: Descriptive statistics of sociodemographic features and Level of Service attributes

Attributes (variable name)	**Attribute levels**	**Number of observations**	**Percentage (%)**
Sociodemographic features			
Gender	Male	230	42.6
	Female	310	57.4
Age	18-24	98	18.1
	25-39	200	37.1
	40-59	201	37.2
	Above 60	41	7.6
Occupation	Employed	359	66.5
	Student	69	12.8
	Other (unemployed, retired, other occupation)	8	20.7
Household income	Below 50k	288	53.3
	Above 50k	198	36.7

	Non reporters	54	10.0
Education	Lower (O-level, A-level, vocational)	156	28.9
	Higher (undergraduate, MSc, PhD)	384	71.1
Marital status	Single	165	30.6
	Married	238	44.1
	Other (divorced, widowed, cohabiting)	137	25.4
Number of cars owned	mean	0.9	-
Number of bicycles owned	mean	0.8	-
Household size	mean	1.9	-
Transit season ticket holder	Yes	144	26.7
	No	396	73.3
Level of Service attributes			
Car travel time (min)	mean	17.1	-
Bus travel time (min)	mean	35.9	-
Rail travel time (min)	mean	60.5	-
Taxi travel time (min)	mean	15.8	-
Cycling travel time (min)	mean	22.8	-
Walking travel time (min)	mean	18.6	-
Car travel cost (£)	mean	1.3	-
Bus travel cost (£)	mean	2.6	-
Rail travel cost (£)	mean	10.4	-
Taxi travel cost (£)	mean	8.7	-
Car travel distance (km)	mean	10.7	-
Bus travel distance (km)	mean	8.3	-
Rail travel distance (km)	mean	21.2	-
Taxi travel distance (km)	mean	9.0	-
Cycling travel distance (km)	mean	6.3	-
Walking travel distance (km)	mean	1.5	-

4. RESULTS

4.1 LCCM development (Model 1)

Regarding the LCCM development, the basic form for the mode choice model component was developed first. Different model specifications were tested including the available level of service variables, namely *travel time, travel cost*, *in-vehicle travel time* and *access-egress time,* with the last two being specified only for bus and rail alternatives.

For the latent class membership component, 2, 3, 4 and 5 classes have been tested with the aim of providing interpretable results and distinguishable individual behaviour among the classes. A model with 6 latent classes could not be estimated due to numerical issues and no further attempt was made to try models with additional classes. For the covariates included in the class allocation model, the requirement was to include variables that were statistically significant for at least one class and be able to provide a meaningful behavioural interpretation. Following that approach, the sociodemographic features of gender and age alongside the long-term mobility choices of number of cars in the household and the ownership of a transit season ticket for either bus or rail were included in the class allocation model. After the allocation of individuals into classes, the class-specific mode choice model is estimated using the ASCs, travel time and travel cost as independent variables. The final LCCM framework developed is illustrated in **Figure 4**.

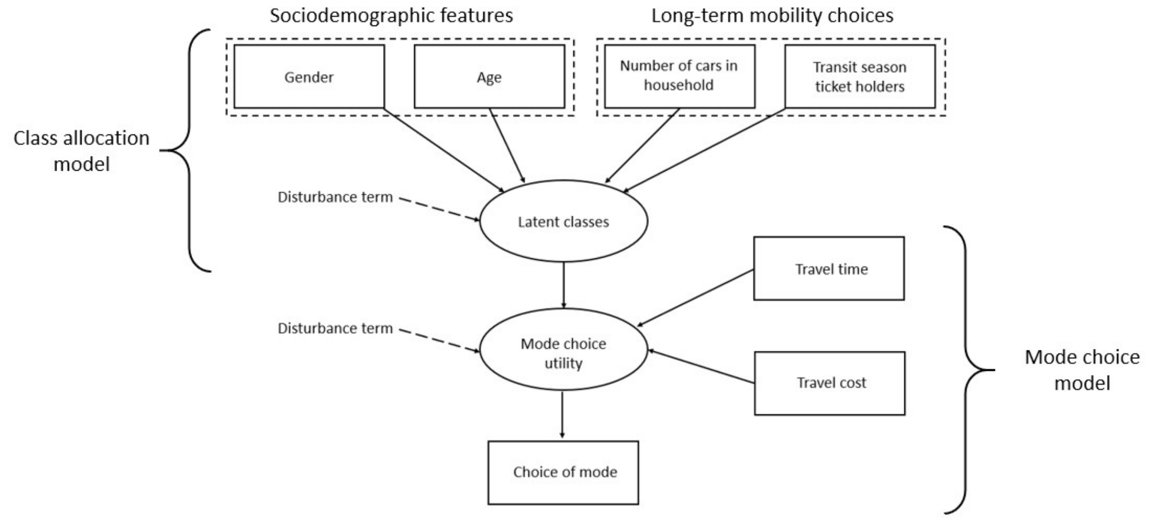


Figure 4 LCCM modelling framework

In order to select the preferred number of classes, a comparison was made among the estimated models with 2, 3, 4 and 5 classes in terms of LL, AIC, BIC and adjusted Rho-square Models consistently performed better with the addition of more classes as shown by the continuous improvement of all model fit statistics. Nonetheless, after the 3-class model the LL improved less with additional classes indicating diminishing returns in terms of model fit when choosing a more complex model specification. As a result, the 3-class model was selected as the most optimal segmentation of the sample population.

In **Table 2,** the modelling outputs of the class-specific and class allocation model for the 3-class LCCM are depicted and compared with the estimates of the unsegmented model. Further, as a measure of validation, the total VoT was calculated as the weighted average of the class-specific VoTs, excluding Class 1, due to a travel cost coefficient non-significantly different than 0. Moreover, the bus fare elasticity was estimated as the demand change for 1% ticket price increase. Those values were compared to the values suggested by the Transport Appraisal Guidance in the UK. The estimated VoT (9.9 £/hr) is very close to the suggested value (13.87 £/hr) (Department for Transport, 2014) although a little lower, and the estimated bus fare elasticity is -0.18 being at the lower bound of the suggested range of values (from-0.16 to -0.65) (Dunkerley et al., 2018).

Table 2: Estimated parameters for the class allocation, the class-specific (3-class model) and the unsegmented mode choice models

Estimated parameters	Class 1	Class 2	Class 3	Unsegmented model
		Class allocation model		
Sample size percentage	51%`	28%	21%	-
Class-specific constants	-0.2070 *(-0.35)*	-0.7228 *(-1.59)*	-	-
		Individual characteristics		-
Male (base)	-	-	-	-
Female	0.9559*** *(2.73)*	0.8117** *(2.20)*	-	-
Age 18-24 (base)	-	-	-	-
Age 25-39	0.2000 *(0.40)*	0.8678* *(1.89)*	-	-
Age 40-59	0.2179 *(0.43)*	0.7593 *(1.50)*	-	-
Age above 60	1.6649 *(1.60)*	2.6537** *(2.42)*	-	-
Number of cars	0.5604** *(2.12)*	-0.7986** *(-2.43)*	-	-

Transit season ticket holders	-1.177** *(-2.17)*	1.0875*** *(2.61)*	-	-
	Class-specific mode choice model			**Unsegmented model**
		Alternative-specific constants		
Car	-	-	-	-
Bus	-4.1758*** *(-10.63)*	-0.2998 *(-0.73)*	-2.0197*** *(-4.49)*	-1.7177*** *(-10.82)*
Rail	-1.7068*** *(-4.53)*	-2.0385*** *(-3.56)*	0.8816** *(2.18)*	-1.2837*** *(-5.34)*
Taxi	-4.9392*** *(-13.46)*	-2.1033*** *(-4.45)*	-1.2008*** *(-1.75)*	-2.8984*** *(-12.06)*
Cycling	-6.4372*** *(-11.48)*	-5.6561*** *(-6.95)*	-1.2855*** *(-3.94)*	-3.4101*** *(-15.18)*
Walking	-1.3043*** *(-5.12)*	0.0450 *(0.08)*	0.5140 *(1.40)*	-0.3191* *(-1.89)*
		Level of Service attributes		
Travel time (min)	-0.061*** *(-5.62)*	-0.082** *(-2.35)*	-0.042*** *(-6.17)*	-0.0711*** *(-8.02)*
Travel cost (£)	-0.138 *(-0.42)*	-0.335*** *(-3.52)*	-0.285*** *(-4.99)*	-0.2650*** *(-7.51)*
		Validation measures		
VoT (£/hr)	-[1]	10.8	8.9	16.1
Total VoT (£/hr)		9.9		16.1
Bus fare elasticity		-0.18		-0.27
		Fit statistics		
LL (final)		-4605.82		-5911.14
AIC		9281.64		11836.27
BIC		9541.88		11888.32
Adjusted R-square		0.6901		0.6048
Number of individuals		540		540
Number of observations		12524		12524

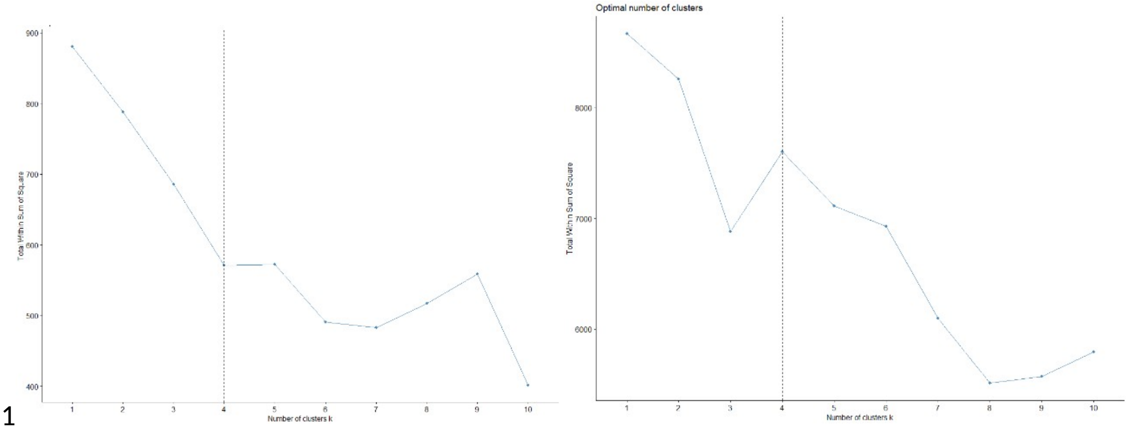
*(***, **, * Significant at the 99% (2.575), 95% (1.96) and 90% (1.645) confidence level, respectively)*

[1] *VoT for Class 1 could not be computed due to the non-significant travel cost parameter*

4.2 K-modes/K-means clustering approach (Models 2 & 3)

In order to select the optimum number of clusters for K-means/K-modes, the "*Elbow*" method (Thorndike, 1953) was chosen to identify the cut-off point where no significant improvement on within-the-cluster-sum-of-differences occurs anymore with the addition of more clusters. For the implementation of the *"Elbow"* method, the K-means/K-modes algorithm has been applied iteratively in the dataset for each case by specifying a different number of clusters each time, ranging from 1 to 10.

In Model 2 only sociodemographic data is included for clustering, namely the same covariates as in the LCCM, and the second stage a DCM is estimated per cluster. Since the included variables were all discrete, the K-modes algorithm was utilised. In Model 3, sociodemographic data and mobility-related information are included for clustering. The mobility data refers to the number of times each mode is chosen per individual. The K-means algorithm was implemented in that case, since there was a combination of discrete and continuous variables. At the second stage, a DCM is estimated per cluster.



(a) Model 2 **(b) Model 3**

Figure 5 The "Elbow" method results

4.2.1 Model 2

The K-modes clustering algorithm was applied to the dataset and the sample was segmented into 4 clusters covering 53.4%, 31.6%, 10.8% and 4.2% of the sample, respectively. Comparing the sociodemographic percentages within each cluster with the unsegmented sample percentages, it was seen that members of Cluster 1, which covers more than half of the sample, have a high inherent preference for car. Cluster 2 has a distinct preference for car over the rest of the alternatives except for walking where there is a non-significant ASC. Furthermore, it is worth noting that this cluster has the highest VoT and bus fare elasticity. Cluster 3 has a non-significant predisposition for car over bus, rail and walking. Cluster 4, covering the smallest sample percentage, has a strong preference for car.

Regarding the model assessment, the combined total loglikelihood from the 4-cluster model provides a significant improvement over the unsegmented model as suggested by the LR-test. The validation measures of total VoT and bus fare elasticity are again within an acceptable range. The total LL (-5712.58), however, is significantly worse than the LCCM (-4605.82) indicating that there is still uncaptured heterogeneity in the two-stage Clustering-based model using only sociodemographic data.

Table 3: Estimated parameters for the cluster-specific mode choice model of Model 2

Estimated parameters	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Alternative-specific constants				
Car	-	-	-	-
Bus	-2.0620*** *(-8.44)*	-1.6257*** *(-6.12)*	-0.3972 *(-1.04)*	-1.7740*** *(-4.28)*
Rail	-1.3368*** *(-3.94)*	-1.4795*** *(-4.21)*	-0.4955 *(-0.93)*	-1.4798*** *(-2.72)*
Taxi	-3.6579*** *(-11.04)*	-2.3727*** *(-5.73)*	-1.9202*** *(-3.06)*	-3.4811*** *(-3.95)*
Cycling	-3.2769*** *(-12.28)*	-4.9209*** *(-9.39)*	-3.2395*** *(-3.89)*	-3.1407** *(-2.57)*
Walking	-0.6367** *(-2.32)*	0.1383 *(0.53)*	0.2974 *(0.71)*	-1.2049** *(-2.45)*
Level of Service attributes				
Travel time (min)	-0.0591*** *(-4.26)*	-0.1027*** *(-7.21)*	-0.0735*** *(-5.13)*	-0.0561*** *(-2.73)*
Travel cost (£)	-0.2319*** *(-5.52)*	-0.3273*** *(-5.07)*	-0.2408** *(-2.24)*	-0.2805** *(-2.47)*
Validation measures				
VoT (£/hr)	15.3	18.8	18.3	12.0

Bus fare elasticity	-0.26	-0.38	-0.12	-0.20
Total VoT (£/hr)		16.6		
Total bus fare elasticity		-0.28		
Fit statistics				
LL (final)	-2983.54	-1659.70	-827.32	-242.04
Total LL		-5712.58		
AIC	5981.08	3333.4	1668.64	498.08
BIC	6028.74	3377.39	1705.09	527.89
Adjusted R-square	0.6351	0.6357	0.4633	0.6156
Sample percentage (%)	53.4	31.6	10.8	4.2
Number of individuals	288	178	50	24
Number of observations	6693	3959	1349	523

4.2.2 Model 3

For the second clustering-based model, the 4-cluster model performed the best. The 4 estimated clusters cover 35.4%, 33.7%, 18.5% and 12.4%. The model outputs and the sociodemographic percentages of the clusters are presented in **Tables 4** and **6**, respectively. The striking feature is the difference in the socio-demographic characteristics of each cluster from Model 2.

Table 4: Estimated parameters for the cluster-specific mode choice model of ML_choice_DCM

Estimated parameters	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
Alternative-specific constants				
Car	-	-	-	-
Bus	-0.5170** *(-2.54)*	-3.3353*** *(-11.70)*	-0.7306** *(-2.19)*	-4.1227*** *(-8.44)*
Rail	-0.5871** *(-2.28)*	-1.9193*** *(-3.43)*	-0.0832 *(-0.15)*	-0.3457 *(-0.48)*
Taxi	-2.3488*** *(-7.13)*	-3.5611*** *(-10.79)*	-1.7874*** *(-3.22)*	-3.2076*** *(-6.47)*
Cycling	-2.1775*** *(-7.33)*	-3.8796*** *(-10.97)*	-2.7978*** *(-4.37)*	-6.9390*** *(-7.25)*
Walking	0.2429 *(1.05)*	-1.4793*** *(-7.15)*	1.7403*** *(4.95)*	-2.1422*** *(-5.55)*
Level of Service attributes				
Travel time (min)	-0.0683*** *(-5.89)*	-0.0652*** *(-6.25)*	-0.0706*** *(-3.08)*	-0.0634*** *(-5.02)*
Travel cost (£)	-0.2056*** *(-4.54)*	-0.2717*** *(-4.80)*	-0.1851** *(-2.56)*	-0.4104*** *(-7.05)*
Validation measures				
VoT (£/hr)	19.9	14.4	22.9	9.3
Bus fare elasticity	-0.16	-0.43	-0.22	-0.72
Total VoT (£/hr)		17.3		
Total bus fare elasticity		-0.33		
Fit statistics				
LL (final)	-3011.97	-1047.97	-1089.19	-205.55
Total LL		-5354.68		

AIC	6037.94	2109.93	2192.39	425.1
BIC	6082.73	2154.36	2232.62	462.54
Adjusted R-square	0.4011	0.8033	0.561	0.8975
Sample percentage (%)	35.4	33.7	18.5	12.4
Number of individuals	305	147	62	26
Number of observations	4435	4219	2316	1554

*(***, **, * Significant at the 99% (2.575), 95% (1.96) and 90% (1.645) confidence level, respectively)*

4.3 COMPARISON

As a comparison of the 3 models estimated it should be highlighted that the Model 1(LCCM) outperformed both clustering-based models in terms of model fit, thus resulting in more accurate estimates. Model 3 performed better than Model 2 which is expected given that it uses more inputs.

In terms of validation with the WEBTag values, all the values for the full models where close to the suggested values and within the suggested limits. The estimated VoT of LCCM was smaller than those of the two clustering-based models. On the other hand, Model 3 showed the highest VoT and bus fare elasticity. The other striking feature is the dissimilarity in the sociodemographic composition in the segments in the three models (**Table 6**) which denote how inherently different the three models are.

Table 5 Model comparison

Model comparison measures	LCCM			ML_socio_DCM				ML_choice_DCM				WEBTag values
	Class 1	Class 2	Class 3	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 1	Cluster 2	Cluster 3	Cluster 4	
VoT (£/hr)	-	10.8	8.9	15.3	18.8	18.3	12.0	19.9	14.4	22.9	9.3	-
Bus fare elasticity	-	-	-	-0.26	-0.38	-0.12	-0.20	-0.16	-0.43	-0.22	-0.72	-
Total VoT (£/hr)	9.9			16.6				17.3				13.87
Total bus fare elasticity	-0.18			-0.28				-0.33				From -016 to -0.65
Total LL	-4605.82			-5712.58				-5354.68				-

Table 6: Sociodemographic percentages and qualitative assessment per model

Sociodemographic characteristics (*sample percentages/mean values*)	LCCM			ML_socio_DCM				ML_choice_DCM			
	Class 1	Class 2	Class 3	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Sample size percentage	0.51	0.28	0.21	0.53	0.32	0.11	0.04	0.35	0.34	0.19	0.12
Included sociodemographic features											
Male *(0.43)*	0.36 (lower)	0.43 (average)	0.58 (higher)	0.71 (higher)	0.00 (lower)	0.00 (lower)	1.00 (higher)	0.45 (average)	0.37 (lower)	0.47 (average)	0.35 (lower)
Female *(0.57)*	0.64 (higher)	0.57 (average)	0.42 (lower)	0.29 (lower)	1.00 (higher)	1.00 (higher)	0.00 (lower)	0.55 (average)	0.63 (higher)	0.53 (average)	0.65 (higher)
Age 18-24 *(0.18)*	0.16 (average)	0.17 (average)	0.25 (higher)	0.13 (lower)	0.23 (higher)	0.38 (higher)	0.00 (lower)	0.18 (average)	0.07 (lower)	0.53 (higher)	0.00 (lower)
Age 25-39 *(0.37)*	0.35 (average)	0.41 (higher)	0.36 (average)	0.27 (lower)	0.69 (higher)	0.00 (lower)	0.00 (lower)	0.40 (average)	0.36 (average)	0.31 (lower)	0.23 (lower)
Age 40-59 *(0.37)*	0.40 (higher)	0.32 (lower)	0.37 (average)	0.60 (higher)	0.00 (lower)	0.58 (higher)	0.00 (lower)	0.35 (average)	0.46 (higher)	0.13 (lower)	0.77 (higher)
Age above 60 *(0.08)*	0.09 (average)	0.09 (average)	0.02 (lower)	0.00 (lower)	0.08 (average)	0.04 (lower)	1.00 (higher)	0.07 (average)	0.11 (higher)	0.03 (lower)	0.00 (lower)
Number of cars *(0.9)*	1.1 (higher)	0.6 (lower)	0.8 (average)	1.0 (average)	0.8 (average)	0.6 (lower)	1.1 (higher)	0.8 (average)	1.3 (higher)	0.04 (lower)	1.5 (higher)
Transit season	0.12	0.53	0.28	0.23	0.12	1.00	0.21	0.39	0.10	0.15	0.08

ticket Yes *(0.27)*	(lower)	(higher)	(average)	(lower)	(lower)	(higher)	(lower)	(higher)	(lower)	(lower)	(lower)
Transit season ticket No *(0.73)*	0.88 (higher)	0.47 (lower)	0.72 (average)	0.77 (higher)	0.88 (higher)	0.00 (lower)	0.79 (higher)	0.61 (lower)	0.90 (lower)	0.85 (higher)	0.92 (higher)
Non included sociodemographic features											
Household income below 50k *(0.53)*	0.49 (lower)	0.63 (higher)	0.52 (average)	0.50 (average)	0.54 (average)	0.58 (higher)	0.75 (higher)	0.57 (higher)	0.48 (lower)	0.61 (higher)	0.31 (lower)
Household income above 50k *(0.37)*	0.44 (higher)	0.25 (lower)	0.34 (average)	0.42 (higher)	0.32 (lower)	0.28 (lower)	0.25 (lower)	0.33 (lower)	0.46 (higher)	0.21 (lower)	0.65 (higher)
Household income not reported *(0.10)*	0.08 (average)	0.12 (average)	0.14 (higher)	0.08 (average)	0.14 (higher)	0.14 (higher)	0.00 (lower)	0.11 (average)	0.06 (lower)	0.18 (higher)	0.04 (lower)
Education lower (A-level, O-level, vocational) *(0.29)*	0.29 (average)	0.33 (higher)	0.24 (lower)	0.33 (average)	0.17 (lower)	0.36 (higher)	0.46 (higher)	0.31 (average)	0.33 (average)	0.13 (lower)	0.19 (lower)
Education higher (undergraduate, MSc, PhD) *(0.71)*	0.71 (average)	0.67 (lower)	0.76 (higher)	0.67 (average)	0.83 (higher)	0.64 (lower)	0.54 (lower)	0.69 (average)	0.67 (average)	0.87 (higher)	0.81 (higher)
Marital status-single *(0.31)*	0.24 (lower)	0.35 (higher)	0.41 (higher)	0.30 (average)	0.33 (average)	0.40 (higher)	0.00 (lower)	0.32 (average)	0.18 (lower)	0.68 (higher)	0.04 (lower)
Marital status-married *(0.44)*	0.51 (higher)	0.35 (lower)	0.40 (average)	0.49 (higher)	0.38 (lower)	0.24 (lower)	0.75 (higher)	0.43 (average)	0.54 (higher)	0.08 (lower)	0.85 (higher)
Marital status-other (divorced, widowed, cohabiting) *(0.25)*	0.26 (average)	0.30 (higher)	0.19 (lower)	0.21 (lower)	0.29 (higher)	0.36 (higher)	0.25 (average)	0.25 (average)	0.29 (average)	0.24 (average)	0.12 (lower)
Household size *(1.9)*	2.0 (average)	1.6 (lower)	2.3 (higher)	2.0 (average)	1.9 (average)	2.1 (higher)	1.0 (lower)	1.9 (average)	1.8 (average)	2.3 (higher)	2.4 (higher)
Occupation working *(0.67)*	0.71 (higher)	0.63 (average)	0.61 (lower)	0.75 (higher)	0.58 (lower)	0.62 (lower)	0.29 (lower)	0.68 (average)	0.73 (higher)	0.39 (lower)	0.85 (higher)
Occupation student *(0.13)*	0.10 (lower)	0.11 (average)	0.23 (higher)	0.10 (average)	0.18 (higher)	0.16 (higher)	0.04 (lower)	0.11 (average)	0.03 (lower)	0.50 (higher)	0.00 (lower)
Occupation other (unemployed, retired, other occupation) *(0.20)*	0.19 (average)	0.27 (higher)	0.17 (average)	0.15 (lower)	0.24 (higher)	0.22 (average)	0.67 (higher)	0.21 (average)	0.24 (average)	0.11 (lower)	0.15 (lower)

1

5. CONCLUSIONS

As a general conclusion, it should be highlighted that both sample segmentation approaches resulted in models with better statistical fit than the unsegmented model. That alone is of great significance from a policy perspective. The general practice of estimating and applying unsegmented behavioural models to real-world projects could have adverse effects in project valuation and in creating a model shift to more sustainable modes, since they are based on biased estimates. Incorporating latent lifestyles into a behavioural model can provide an insight on the individuals' subconscious taste preferences leading to better planned policy initiatives. Regarding the comparison of the models estimated, the LCCM outperformed the clustering-based models by a large margin. From the two clustering-based models, Model 3 (which included both socio-demographics and choices) resulted a better model fit than the one that uses only socio-demographics. Furthermore, significant differences were observed in the composition of the segments in the three methods.

Nonetheless, the current study was also subject to certain limitations. The initial goal during model development was to estimate a mixed LCCM model and mixed MNL models for the clustering-based approach to capture heterogeneity both across classes/clusters and among individuals within the same class/cluster. That was not possible, however, due to numerical issues in mixed LCCM. Specifically, the presence of a non-significant travel cost parameter for Class 1 resulted in errors during the calculation of the covariance matrix, possibly because no further heterogeneity could be captured for an already insignificant parameter. Furthermore, there are more advanced clustering algorithms, such as the *"Soft"* or *"Fuzzy"* K-means/K-modes which performs a probabilistic clustering by allocating a data point to a cluster with a certain probability or weight. As a next step, the performance of those more advanced "*Soft*" *K-means/*K-modes algorithms is going to be assessed, which is believed to provide better results than the traditional algorithms tested in the current study.

Even in their current form, however, the findings are expected to provide guidance to researchers and practitioners in choosing the most efficient method to capture the taste heterogeneity among the segments of the travelers.

ACKNOWLEDGMENTS

The lead author is supported by the ESRC Advanced Quantitative Scholarship. Professor Stephane Hess acknowledge the support of the European Research Council through the consolidator grant 615596-DECISIONS.

AUTHOR CONTRIBUTIONS

The authors confirm contribution to the paper as follows: study conception and design: PT, CC & SH; data collection: SH, PT & CC; analysis and interpretation of results: PT, CC & SH. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

Anda, C., Erath, A. and Fourie, P.J., 2017. Transport modelling in the age of big data. International Journal of Urban Sciences, 21(sup1), pp.19-42.

Antoniou, C., Dimitriou, L. and Pereira, F.C. (eds.). 2019. *Mobility Patterns, Big Data and Transport Analytics: Tools and Applications for Modelling*. London, United Kingdom: Elsevier.

Aranganayagi, S. and Thangavel, K. 2009. Improved K-Modes for Categorical Clustering Using Weighted Dissimilarity Measure. *World Academy of Science, Engineering and Technology*. **3**(3), pp.729–735.

Bhat, C.R. 1997. An Endogenous Segmentation Mode Choice Model with an Application to Intercity Travel. *Transportation Science*. **31**(1), pp.34–48.

Calastri, C., Crastes dit Sourd, R. and Hess, S. 2018a. We want it all: experiences from a survey seeking to capture social network structures, lifetime events and short-term travel and activity planning. *Transportation*.

Calastri, C., Hess, S., Choudhury, C.F., Daly, A. and Gabrielli, L. 2018b. Mode choice with latent availability and consideration: theory and a case study. *Transportation Research Part B: Methodological*.

Cantarella, G.E. and de Luca, S. 2005. Multilayer feedforward networks for transportation mode choice analysis: An analysis and a comparison with random utility models. *Transportation Research Part C*. **13**, pp.121–155.

Department for Transport 2014. *Values of Time and Vehicle Operating Costs TAG Unit 3.5.6*. Department for Transport.

Dunkerley, F., Wardman, M., Rohr, C. and Fearnley, N. 2018. *Bus fare and journey time elasticities and diversion factors for all modes: A rapid evidence assessment*. RAND Europe and SYSTRA.

Hagenauer, J. and Helbich, M. 2017. A comparative study of machine learning classifiers for modelling travel mode choice. *Expert Systems with Applications*.

Hasnine, M.S. and Habib, K.N. 2019. A Dynamic Discrete Choice model for tour-based mode choices *In*: Washington, D.C., January 13-17, 2019.

Hensher, D.A. and Ton, T.T. 2000. A comparison of the predictive potential of artificial neural networks and nested logit models for commuter mode choice. *Transportation Research Part E*. **36**, pp.155–172.

Hess, S. 2014. Latent class structures: taste heterogeneity and beyond *In*: S. Hess and A. Daly, eds. *Handbook of Choice Modelling*. Cheltenham, Uk: Edward Elgar Publishing Limited, pp.311–329.

Jin, X. and Han, J. 2011. K-Medoids Clustering *In*: C. Sammut and G. I. Webb, eds. *Encyclopedia of Machine Learning*. Boston, MA, USA: Springer.

Kamakura, W.A. and Russell, G. 1989. A probabilistic choice model for market segmentation and elasticity structure. *Journal of Marketing Research*. **26**, pp.379–390.

Kim, D.-W., Lee, K.H. and Lee, D. 2004. Fuzzy clustering of categorical data using centroids. *Pattern Recognition letters*. **25**, pp.1263–1271.

Krizek, K. and Waddell, P. 2003. Analysis of lifestyles choices: neighborhood type, travel patterns, and activity participation. *Transportation Research Record*. **1807**, pp.119–128.

Lanzendorf, M. 2002. Mobility Styles and Travel Behavior Application of a Lifestyle Approach to Leisure Travel. *Transportation Research Record*. **1807**, pp.163–173.

Luce, D. 1959. *Individual Choice Behavior*. New York: John Wiley & Sons.

MacQueen, J. 1967. Some Methods for classificaiton and Analysis of Multivariate Observations *In*: *Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, pp.281--297.

Marschak, J. 1960. Binary choice constraints on random utility indications *In*: K. Arrow, ed. *Stanford Symposium on Mathematic Models in the Social Sciences*. Stanford, CA: Stanford University Press, pp.312–329.

McFadden, D. 1973. Conditional logit analysis of qualitative choice behavior *In*: P. Zarembka, ed. *Frontiers in Econometrics* [Online]. New York: Academic Press, pp.105–142. Available from: https://eml.berkeley.edu/reprints/mcfadden/zarembka.pdf.

McFadden, D. 2000. Disaggregate Behavioural Travel Demand’s RUM Side: A 30-Year Retrospective *In*: D. Hensher, ed. *Travel Behaviour Research: The Leading Edge* [Online]. Pergamon Press: Oxford, UK, 2000, pp.17–63. Available from: https://www.researchgate.net/publication/2331526_Disaggregate_Behavioral_Travel_Demand's_RUM_Side_-_A_30-Year_Retrospective.

Salomon, I. and Ben-Akiva, M. 1983. The use of the life-style concept in travel demand models. *Environment and Planning A*,. **15**, pp.623–638.

Sekhar, C.R., Minai and Madhu, E. 2016. Mode Choice Analysis Using Random Forrest Decision Trees. *Transportation Research Procedia*. **17**, pp.644–652.

Thorndike, R.L. 1953. Who belongs in the family? *Psychometrika*. **18**(4), pp.267–276.

Train, K. 2009. *Discrete Choice Methods with Simulation*. Cambridge, Massachusetts: Cambridge University Press.

Xie, C., Lu, J. and Parkany, E. 2003. Work Travel Mode Choice Modeling Using Data Mining: Decision Trees and Neural Networks. *Transportation Research Record: Journal of the Transportation Research Board*. **1854**(1), pp.50–61.

Zhang, Y. and Xie, Y. 2008. Travel Mode Choice Modeling with Support Vector Machines. *Transportation Research Record: Journal of the Transportation Research Board*. **2076**, pp.141–150.