Transportation Research Part C 179 (2025) 105289

Contents lists available at ScienceDirect

Transportation Research Part C

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

Using probabilistic clustering techniques as a specification tool for capturing heterogeneity in choice models

Panagiotis Tsoleridis *, Charisma F. Choudhury, Stephane Hess

University of Leeds, University road, Leeds, LS2 9JT, UK

ARTICLE INFO

Keywords:
Latent class choice models
Individual heterogeneity
Probabilistic clustering
Data mining
Mode choice
Destination choice

ABSTRACT

In the era of big data, data-driven methods have emerged as strong competitors to traditional econometric models for analysing choice behaviour. In particular, data-driven models offer flexible classification methods that are well-suited to capturing the heterogeneity among decision makers and improving model fit. A key limitation of the purely data-driven models, however, is the difficulty in the calculation of welfare measures, such as the value of travel time estimates (VTT) that are essential for cost–benefit analyses. This motivates the current study which focuses on combining data mining based segmentation approaches used in ML with traditional discrete choice models (DCM) to get the best of both - a clustering-based component to capture the heterogeneity among the travellers and a utility-based choice component that is suitable for quantifying policy-relevant measures, such as VTT estimates. In the proposed hybrid framework, travellers are probabilistically allocated into clusters based on their degree of similarity from each cluster and cluster-specific random-utility-based mode choice models are estimated simultaneously. The proposed hybrid framework is tested on 2 RP datasets (a GPS diary and a traditional household survey) and on 3 different choice contexts, providing a range of different sample sizes and data complexity. The performance of the proposed hybrid model (H-LCCM) is compared with that of the traditional latent class choice models (LCCM), where both the class membership and mode choice components are based on utility-based frameworks and two other state-of-the-art ML-assisted LCCM frameworks. Results indicate that H-LCCM outperforms the remaining specifications in the majority of the contexts examined, while offering a more scalable approach for contexts with a large number of observations (which is the case for big data sources) and/or with large choice sets (which is typical in spatial choice contexts). The proposed framework is practically applicable for policy-making as it allows the calculation of VTT estimates, therefore not sacrificing the microeconomic interpretability of traditional DCMs. The results are promising, especially in the current era of big data and are expected to contribute to the emerging literature looking at cross-synergies between traditional econometric approaches and new data-driven methods.

1. Introduction

During the last decade, the abundance of passively generated location data has provided interesting insights into human mobility behaviour. For instance, GPS traces, mobile phone call detail records, and public transport smart card data, to name but a few, not only provide digital footprints of a very large sample of travellers, but often also have repeated observations of the same person

* Corresponding author.
E-mail addresses: P.Tsoleridis@leeds.ac.uk (P. Tsoleridis), C.F.Choudhury@leeds.ac.uk (C.F. Choudhury), S.Hess@leeds.ac.uk (S. Hess).

https://doi.org/10.1016/j.trc.2025.105289
Received 6 September 2024; Received in revised form 2 June 2025; Accepted 21 July 2025
Available online 17 August 2025
0968-090X/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

over a long period of time. The panel nature of the data provides rich insights into the similarity/dissimilarity of the travellers, which can be used to better capture the heterogeneity in their travel decisions.

Deriving 'value' out of these new forms of data, however, typically requires significant pre-processing and the use of methods from distinct fields of research (e.g. Computer Science) (Antoniou et al., 2019). In addition to that, the massive size of the data has started to highlight the limitations of well-established tools and methods for their analysis and the scope for improvements (Milne and Watling, 2019). This has led to an increase in the popularity of Machine Learning (ML) techniques. In particular, data mining techniques have been widely used in capturing patterns in the data (Crawford, 2017; Hasan and Ukkusuri, 2018) and hold the promise to better capture the behavioural heterogeneity among individuals.

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 about the underlying model. ML encompasses a large array of algorithms, which can be broadly categorised into supervised and unsupervised learning. A wide range of studies have implemented clustering algorithms (unsupervised learning) to analyse individual behaviour and uncover mobility patterns (Joh et al., 2001; Hasan and Ukkusuri, 2014, 2015; Anda et al., 2017; Crawford, 2017; Hasan and Ukkusuri, 2018; Song et al., 2021). Though such studies provide good insights into the state of the network, they have limited applications in the context of predictions and/or valuation (e.g. calculation of the value of time estimates to feed into the cost–benefit analyses).

Travel behaviour researchers have arguably shown a larger interest in the use of supervised ML algorithms, such as Artificial Neural Networks and Random Forests, and on their comparison with traditional econometric Discrete Choice Modelling (DCM) frameworks, such as Multinomial logit (MNL) and Nested Logit (NL) models, 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 at large suggest that ML algorithms have the potential to be used as an alternative method for behavioural modelling due to their superior predictive performance, although even early work by Hensher and Ton (2000) already highlighted the limitations associated with the lack of interpretable results compared to a DCM framework.

The majority of studies from that initial stream of literature is subject to three key limitations. Firstly, they relied on data collected using traditional methods (e.g. single RP choice scenarios, short trip diaries, etc.) where the advantages of using ML are likely to be limited. Secondly, these studies did not compare the model performance to that of more advanced discrete choice models that account for heterogeneity among groups of decision-makers. Thirdly, those earlier studies focused on comparing the goodness of fit and/or prediction capabilities of ML and DCM as opposed to a more in-depth effort of formulating models that combine the best of both worlds – the computational advantages of data mining that can more efficiently uncover associations within complex data and the behavioural interpretation of DCM that can produce outputs suitable for valuation and cost–benefit analysis. It is therefore worth investigating the performance on passively collected large samples (with repeated observations), to extend the work to more advanced discrete choice models, and to go beyond model fit in comparisons.

Wang et al. (2021a) aimed to generalise the empirical results of the studies so far by comparing a vast range of ML algorithms and choice models on a range of different datasets concluding that it would be advantageous to use ML algorithms for predicting travel behaviour, while also highlighting the need for DCM to improve their computational efficiency to be more suitable for estimating models on large datasets. These initial studies have also motivated researchers to investigate methodologies to combine the ML and DCM paradigms. More specifically, data mining techniques can provide a more flexible and scalable approach for identifying patterns in recent datasets of increasing complexity, both in terms of size (larger samples) and type (text and images). DCM, with their econometrics grounding and strong theoretical underpinnings of human behaviour (McFadden, 1973, 1978; Ben-Akiva and Lerman, 1985; McFadden, 2000; Train, 2009), can provide behavioural insights that can be used for policy making (e.g. Values of Travel Time), while also providing clear interpretations on the impact of the utilised independent variables and their statistical significance. Hence, cross-fertilisation of ML approaches with DCM is very appealing for policy analyses to get the best out of both worlds (van Cranenburgh et al., 2022). Prominent examples of combining DCM and ML include the studies of Sifringer et al. (2020), Wang et al. (2021b) and Wong and Farooq (2021) in all of which Deep Learning architectures have been integrated with DCM specifications in the context of mode choice, risk and time preference.

In a similar notion, there have been attempts to harness the power of unsupervised learning for uncovering latent segments of the population to aid the estimation of advanced choice models, namely Latent Class Choice Models (LCCM) (Kamakura and Russell, 1989). An LCCM framework is typically used to simultaneously identify latent classes of individuals in the sample (class allocation component) and to understand their observed behaviour (behavioural model) in a joint estimation process. LCCMs can provide rich behavioural insights that can help the analyst to link types of individuals to unique styles of mobility behaviour -originally latent-, which can help to design policy measures more suitable to the needs of the underlying population (Vij et al., 2013). In a recent series of studies, Sfeir et al. (2021) and Sfeir et al. (2022) provided significant research advancements towards that direction by integrating probabilistic ML algorithms, namely Gaussian mixture models and Gaussian processes respectively, into a Latent Class Choice Model (LCCM) framework effectively replacing the random utility-based class allocation component with ML algorithms. In both cases, their proposed specifications were tested on models of mode choice behaviour using traditional Revealed (RP) and Stated Preference (SP) datasets. Overall, the non-parametric Gaussian processes outperformed the Gaussian mixture variants of LCCM and the traditional LCCM in terms of model fit and estimation stability, while also resulting in estimates with behaviourally consistent signs. Nonetheless, the models in those studies were estimated only on traditional RP and SP data and it is not clear if that ML-DCM integration can provide additional benefits when used for modelling travel behaviour in the context of passively generated big data sources where the increased number of observations per traveller could offer a more detailed depiction of the underlying heterogeneity. Furthermore, the Gaussian process LCCM models would require an additional specification test to

find the best-performing kernel function out of a range of available functions and additional time to calibrate it, which according to the authors will *"add further burden to the modeller"* (Sfeir et al., 2022). Finally, both models, Gaussian mixture and Gaussian process LCCMs, would require more estimated parameters than the traditional LCCM specification leading to generally more difficult estimation processes. That could also result in limitations in terms of the sample sizes used in estimation and therefore in the potential heterogeneity to be captured and in terms of use cases, such as choice contexts with large choice sets. This highlights the need of a simpler approach based on similar principles that can highlight the benefits of incorporating data mining techniques with DCM in large-scale datasets, while not sacrificing the interpretability of DCM so that they can be used for policymaking.

The present research aims to contribute to the above stream of literature by proposing an approach that integrates a probabilistic clustering algorithm, namely K-means clustering, in LCCM analysis. Several studies have used clustering techniques for market/sample segmentation (Salomon and Ben-Akiva, 1983; Lanzendorf, 2002; Krizek and Waddell, 2003) reporting that different lifestyle clusters (empirically identified) could have different choice elasticities. Nonetheless, the clustering algorithms in those studies were used to deterministically allocate individuals into clusters, while the clustering process was independent from the choice behaviour itself. More recently, Hafezi et al. (2019) utilised a probabilistic clustering approach, Fuzzy C-Means, to allocate individuals to homogeneous clusters of activity schedules by a probability that is relative to the distance of each data point (person-day activities) from all the cluster centroids. Despite taking advantage of a probabilistic approach leading to richer behavioural outcomes, the clustering framework of Hafezi et al. (2019) was not linked to any behavioural model that would aim to understand the observed activity behaviours.

The goal of the present paper is twofold. The main goal is to propose a probabilistic variant of K-means that will be properly integrated with a behavioural model in a joint fashion, mimicking the properties of an LCCM. The second goal is to illustrate that an integration of a clustering algorithm can provide model fit improvements to a LCCM specification, while the lower computational cost of K-means makes it possible to increase the range of potential case studies and estimate models with large sample sizes and large choice sets, resulting in a step change from the previous research combining DCM and ML. The novelty of the proposed framework is to illustrate how a deterministic clustering algorithm can be transformed effectively into a probabilistic one, enabling a simultaneous estimation of parameters of class membership and choice components, with the former receiving feedback from the latter. Therefore, the aim of the current study is to combine a clustering technique from the data mining literature and a DCM specification (MNL model at the lower level) in a combined LCCM framework, while still being able to produce outputs that can be used for valuation, thus making ML relevant for policy making.

The proposed methodology is tested empirically on 2 RP datasets, a GPS diary and a traditional household survey, and on 3 different choice contexts providing a range of different sample sizes and data complexity. The three case studies utilised for the empirical application of the proposed approach focus on:

1. a mode choice model estimated using a GPS trip diary
2. a shopping destination choice model estimated using a GPS trip diary
3. a mode choice model estimated using a traditional trip diary

The remainder of this paper is structured as follows. In Sections 2 and 3, the methodological framework and the different datasets used for the study's practical applications are described, respectively. Section 4 focuses on the results and the comparison among the different approaches. The main conclusions and a potential direction for future research are summarised in the final section.

2. Methodology

2.1. Latent class choice model

DCM and the MNL model in particular have been the main behavioural framework for analysing individual preferences since the seminal study of McFadden (1973). According to that framework, an individual n facing a specific choice task t will choose the alternative i that provides the largest utility U_{int} among a set of J alternatives. The utility U_{int} is a latent construct consisting of two parts, a deterministic utility V_{int} and a disturbance term ε_{int}. The deterministic part of the utility is a function of individual- and alternative-specific attributes x_{int} and parameters β to be estimated, as shown in Eq. (1).

$$U_{int} = V_{int} + \varepsilon_{int} = f(\beta, x_{int}) + \varepsilon_{int} \tag{1}$$

Different distributional assumptions about the disturbance term will lead to different specifications, with independent and identically distributed (iid) extreme value error terms leading to an MNL model. With MNL, the probability of choosing alternative i can then be calculated using Eq. (2).

$$P_{int}(\beta) = \frac{e^{V_{int}}}{\sum_{j=1}^{J} e^{V_{jnt}}} \tag{2}$$

Heterogeneity in an MNL model can be captured by specifying interactions with socio-demographic characteristics, usually specified as shifts of taste parameters away from their base level. Despite those interactions, however, a significant portion of unobserved heterogeneity can remain uncaptured with an MNL model. LCCMs together with mixed logit models (McFadden and Train, 2000) have established themselves as important behavioural modelling specifications capable of accommodating unobserved individual choice heterogeneity. The former achieves this by probabilistically segmenting the sample into a finite number of latent classes

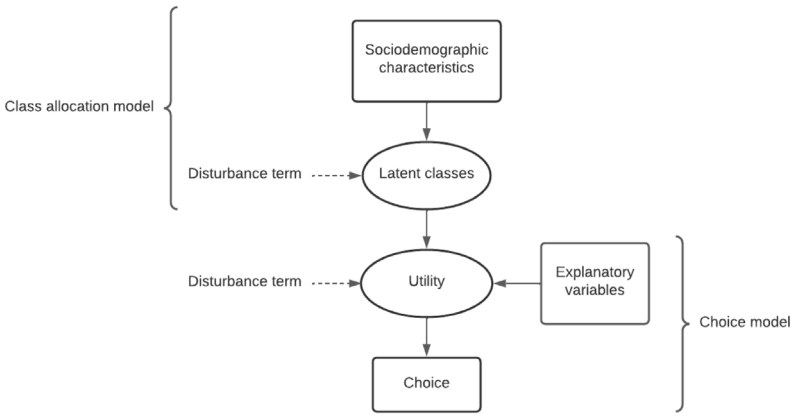


Fig. 1. Schematic diagram of the LCCM framework and its constituent components.

based on individuals' socio-demographic characteristics and their observed choice behaviour. It uses two model components that are jointly estimated, a class allocation model and a choice model conditional on the allocated class. Mixed logit models, on the other hand, require the specification of continuous distributions over the individual taste parameters, thus resulting in non-closed form solutions for the choice probabilities that require numerical approaches, usually simulated estimation procedures (Train, 2009). Besides its closed form solution, the LCCM provides the additional benefit of a more straightforward interpretation of the context of each estimated class, since they can be directly linked with socio-demographic characteristics for each class (when covariates are included in the class allocation component) that could be important for the segmentation of the population during policy formulation.

In a LCCM, it is assumed that the sample can be segmented into a finite number of S heterogeneous classes. The class allocation component of the LCCM, commonly specified as an MNL model, is responsible for probabilistically allocating individuals into the latent classes. Socio-demographic characteristics z_n ($z \in Z$) are included in the class allocation model as covariates, while additional parameters γ_{sz} are estimated per class s and demographic attribute z together with $S-1$ constants, δ_s. The probability π_s of an individual belonging into class s is thus calculated using Eq. (3), with $0 \leq \pi_s \leq 1$ and $\sum_{s=1}^{S} \pi_s = 1$ for each individual n.

$$\pi_s = \frac{e^{\delta_s + g(\xi_s, z_n)}}{\sum_{r=1}^{S} e^{\delta_r + g(\xi_r, z_n)}} \tag{3}$$

For each class in the model, a separate utility function is specified for each alternative, say V_{sint} in class s, where the parameters (and hence utilities) vary across classes. Homogeneity of preferences is usually assumed to hold within each class, although there is also the possibility to capture additional within-class heterogeneity by including covariates in the within class utilities, or by specifying continuous distributions over covariates (Hess, 2014). A choice model at the lower level is then estimated conditional on the class, as depicted in Fig. 1. The choice probabilities for the class-specific model are calculated from Eq. (4). Finally, the unconditional likelihood of observing a sequence of choices for individual n is calculated as Eq. (5) in which class probabilities are used to weight the respective class-specific conditional probabilities for each alternative j. The coefficients of both levels are jointly determined by maximising the logarithm of the likelihood function.

$$P_{sint} = \frac{e^{V_{sint}}}{\sum_{j=1}^{J} e^{V_{sjnt}}} \tag{4}$$

$$L_n^{LCCM}(\beta, \pi) = \sum_{s=1}^{S} \pi_s \prod_{t=1}^{T} P_{sint} \tag{5}$$

2.2. Clustering - latent class choice model

Focusing now on our proposed modelling framework, the main idea is to incorporate a clustering algorithm into an LCCM modelling framework to take the role of the class allocation model. In the current study, we use the K-means clustering algorithm to take that role, mainly for its simplicity, but the same principles can be applied to more advanced algorithms as well. The K-means clustering algorithm (Lloyd, 1982) based on the data mining literature is an unsupervised learning algorithm, which actually predates the popularity of the more recent Machine Learning methods. The traditional K-means is a partition-based clustering algorithm allocating individuals deterministically into a finite K number of clusters based on specific Z socio-demographic characteristics, which are found after a specification search similar to the covariates in a class allocation model. The clustering process itself is an iterative algorithm that tries to minimise a measure of distance among the data points (i.e. individuals) and their respective allocated cluster centroid (within cluster sum of square distance), while at the same time maximising their distance to the centroids

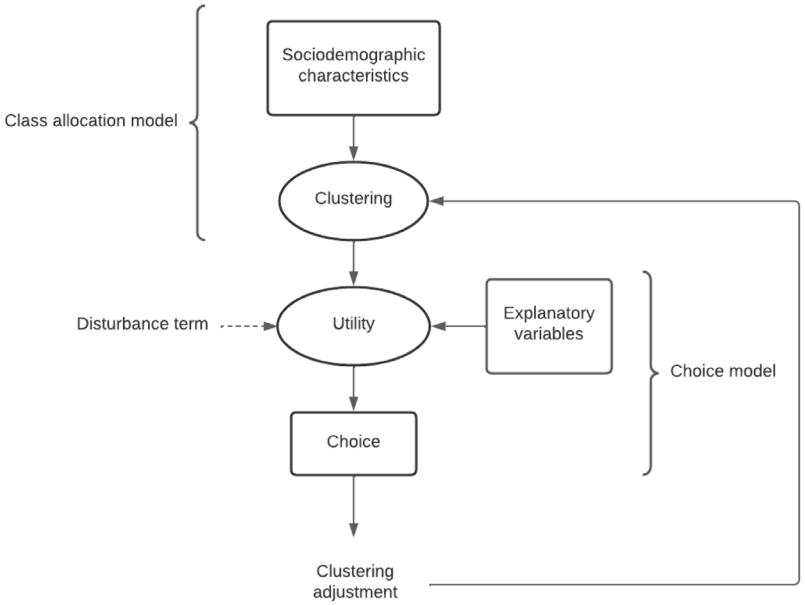


Fig. 2. Schematic diagram of the H-LCCM framework and its constituent components.

of the remaining cluster centroids (between cluster sum of square distance) (Ripley, 2009). Different measures of distance can be used for that purpose, such as the Euclidean, the Manhattan distance (Bishop, 2006; Singh et al., 2013; Bora and Gupta, 2014) or the Mahalanobis distance (Ghorbani, 2019), with the former (cf. Eq. (6)) being implemented in the current paper.[1] In order to avoid the calculated Euclidean distance measure being influenced by the potential scale discrepancies among the different variables used for clustering, it is important to scale the variables prior to the initialisation of the clustering algorithm, either by normalising or by standardising the variables, with the latter approach being utilised here.

$$d_{nk} = \sqrt{\sum_{z=1}^{Z} \left(m_{nz}^{*} - m_{kz}^{*}\right)^2} \tag{6}$$

where d_{nk} is the distance of individual n from cluster centroid k for the socio-demographic attribute z out of a set of Z attributes.

The proposed hybrid methodological framework, *H-LCCM*, developed for this study involves the implementation of a probabilistic transformation of the traditional deterministic K-means clustering algorithm and its efficient integration into an LCCM specification. The probabilistic K-means algorithm is designed to handle the identification of latent segments of travellers based on specific socio-demographic characteristics, while it gets adjusted with the information provided by the choice model with a feedback loop as depicted in Fig. 2.

The class allocation model in a traditional LCCM framework is used to probabilistically allocate individuals into latent classes based on their sociodemographics and their observed behaviour with regard to a specific choice situation. The two important things to note here is first that each individual is allocated with a non-zero probability to every class and second that the class allocation model is getting feedback from the choice model at the lower level. In order to mimic that specification with a K-means algorithm, the first step is to transform it from a deterministic algorithm into a probabilistic one. This is achieved by taking advantage of the fact that each data point (or individual) n is allocated to its closest centroid k, but there still is a non-zero distance $d_{nk} > 0$ with $d_{nk} < d_{nl}$ and $k, l \in K$. Therefore, instead of assuming that an individual n would be allocated entirely into the closest centroid, we re-define her allocation by taking into account her distance from all centroids. Bishop (2006) has highlighted in his book the benefits from moving from a deterministic to a probabilistic K-means algorithm, also known as soft K-means or fuzzy C-means. In our framework the class allocation probability is thus defined as:

$$\pi_{nk} = \frac{e^{\gamma_{dist}^{k} d_{nk}}}{\sum_{l=1}^{K} e^{\gamma_{dist}^{k} d_{nl}}} \tag{7}$$

[1] Following the suggestion by a referee, we also tried using the Mahalanobis distance as an alternative distance measure for our experiments and to account for potential non-isotropic clusters, but for the case studies examined the Euclidean distance measure resulted in a better model fit.

where π_{nk} is the allocation probability of individual n to her closest centroid k, d_{nl} is the distance of individual n from centroid l and finally γ^k_{dist} is a parameter to be estimated controlling the allocation to the closest centroid k relative to the remaining ones. If $\gamma^k_{dist} < 0$, it means that the individual is allocated with a higher probability to the closest centroid k relative to the rest, while the opposite would be true in the case of $\gamma^k_{dist} > 0$, signifying the need for readjusting the allocation of individuals into the clusters. Finally, the use of $\gamma^k_{dist} = 0$ would result into an equal allocation to every cluster. The role of γ^k_{dist} is two-fold. First, it helps to transfer information from the choice model back to the clustering algorithm (class allocation) by adjusting the allocation probabilities and hence the cluster centroids. Second, by specifying a γ^k_{dist} that varies across clusters, it allows us to capture the heterogeneity among clusters on how the points of each cluster are allocated to their own cluster relative to the remaining ones.

The joint model is estimated by maximising the joint likelihood of the class allocation model (Eq. (7)) and the class-specific choice models at the second stage (Eq. (4)) using Eq. (8) with Maximum Likelihood Estimation. It may be noted that the only difference between Eq. (5) and Eq. (8) is the definition of the class allocation models and the way we calculate π. In the case of the traditional LCCM that is estimated from an econometric model (usually an MNL model), while for the proposed H-LCCM that is calculated with probabilistic clustering. That also makes the estimation process between the two approaches fairly similar with the only difference being the iterative calibration of the centroids in H-LCCM.

$$L_n^{H-LCCM}(\beta,\pi) = \sum_{k=1}^{K} \pi_k \prod_{t=1}^{T} P_{sint} \tag{8}$$

Regarding cluster initialisation, three approaches were used to define initial centroids, namely i) random initialisation, ii) initialisation based on a pre-calibrated deterministic K-means and iii) initialisation based on K-means++ algorithm (Arthur and Vassilvitskii, 2007). According to the K-means++ algorithm, an initial data point (i.e. an individual) is randomly selected and assigned as the centroid of the first cluster k_1. The distance d_{nk_1} of all data points n from that initial centroid k_1 are calculated and the second centroid is sampled with a probability equal to $\frac{d^2_{nk_1}}{\Sigma d^2_{mk_1}}$. This implies that data points further away from the initial centroid will have a higher probability of being selected as the second centroid from that process. For the third centroid, the distances of all data points from the two selected centroids are calculated and the next centroid is sampled with a probability based on the square of the minimum distance from the other two centroids. In a similar way, the remaining centroids are sampled until the predetermined number of centroids is reached. Following that, the K-means algorithm can initialise using the previously sampled centroids during the first iteration. In the current study, both the pre-calibrated K-means and the K-means++ approach resulted in superior and more stable results compared to having a random initialisation of centroids. A fourth approach was also implemented that involved the analyst having to define specific centroids manually by trying different sign combinations for the clustering covariates (i.e. same or different sign per cluster etc.), which can be more straightforward in the case of two classes compared to models with multiple classes.

The developed algorithm behind H-LCCM is presented in the flow chart of Fig. 3, which consists of the following steps:

1. Define minimum difference threshold for reaching convergence, $d^{thres}_{k,k'}$ between the centroids of the previous step, k, and the respective ones in the current step k'.
2. Scaling of variables used as covariates in the clustering process.
3. Define starting values for parameters used in the choice model.[2]
4. Initialisation of centroids m_k using the K-means++ algorithm as described above.
5. Calculate distances of data points (individuals) from the initial centroids (Eq. (6)) to define initial cluster allocation probabilities (Eq. (7)).
6. Estimation of choice model for the first iteration using Maximum Likelihood. From that step we get an estimated value for the $\gamma^{k'}_{dist}$.
7. Update cluster allocation π'_{nk} based on the new estimated $\gamma^{k'}_{dist}$ from the choice model of the previous step and the previously defined centroids using Eq. (7).
8. Definition of new centroids $m_{k'}$ for following iteration as the mean of the covariates of the individuals that are being attracted with a higher probability to the same cluster, $m_{k'} = mean(z_{nk'})$, where $k' = argmax(\pi_{nk'})$.
9. Compute the distance between previous and new centroids $d_{k,k'}$ and compare that with the threshold $d^{thres}_{k,k'}$ defined in step 1.
10. If the difference is larger than the threshold, $d_{k,k'} > d^{thres}_{k,k'}$, then estimate a new choice model for next iteration and repeat steps 6–8. If the difference is equal or smaller than the threshold, $d_{k,k'} \leq d^{thres}_{k,k'}$, then convergence is reached.

According to Bishop (2006), the iterative process of K-means can also be described as the Expectation–Maximisation (EM) algorithm (Dempster et al., 1977), where the update of the cluster membership probabilities refers to the expectation step and the adjustment of the cluster centroids refers to the maximisation step.

After reaching a stable point and terminating the iterative estimation process, an additional Fractional Multinomial Logit (FMNL) model (Papke and Wooldridge, 1996) is estimated using the cluster probabilities π_{nk} of cluster k out of a total K clusters for individual

[2] For the initialisation of γ^k_{dist}, an initial value of 1.0 is assigned in the current study assuming a higher allocation probability of each individual to their initially defined closest centroid, but the impact of this should be negligible since the initial centroids are randomly assigned without having any behavioural connection with the decision making process under examination.

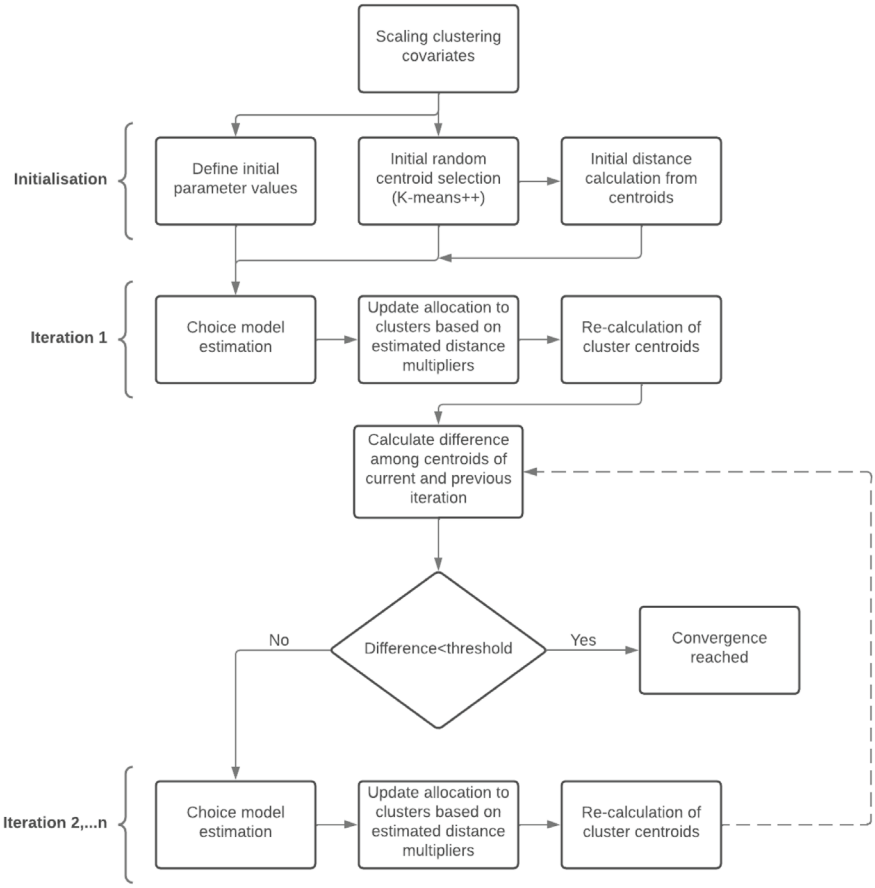


Fig. 3. Flow chart of the H-LCCM algorithm.

n, with $0 \leq \pi_{nk} \leq 1$ and $\sum_{k=1}^{K} \pi_{nk} = 1$, as the dependent variable. The covariates used for clustering z_{nd} for individual n and demographic attribute d take the role of the independent explanatory variables with the estimable parameters ξ_d capturing their impact.

According to the FMNL model, the probability for individual n to choose a share of π_{nk} from alternative k (meaning here the cluster k) is defined as per Eq. (9). It has to be mentioned that the FMNL model is not part of the main estimation process depicted in Fig. 3 and its purpose is solely for post processing and for providing some further interpretability of the clustering process, such as the cluster composition, the relative impact of the covariates used for clustering and their statistical significance .

$$P_{kn}(\xi) = \frac{e^{\sum_{k=1}^{K} \pi_{nk} V_{kn}}}{\sum_{l=1}^{K} e^{V_{ln}}} \tag{9}$$

2.3. Machine learning-based model variants

For the sake of benchmarking, the proposed models are compared against two state-of-the-art models developed by Sfeir et al. (2021) and Sfeir et al. (2022).

2.3.1. Gaussian/Bernoulli mixture - latent class choice model

Sfeir et al. (2021) developed a model structure that involves an integration of Gaussian Mixture models (GMM) and LCCM, where GMM-based components were used to replace the class allocation model of the traditional LCCM. More specifically, Gaussian mixtures are used for continuous covariates and Bernoulli mixtures for discrete/binary ones (Bishop, 2006) to determine the allocation of individuals to a finite number of class, while a class-specific MNL model is estimated at the next level. The gaussian mixtures for the continuous covariates also allow the analysts to estimate the mean and the covariance structure, thus capturing additional correlation among the covariates and the classes. The means μ_{ck} and covariance s_{ck} for continuous mixtures and the

means for the discrete ones μ_{dk} are estimated together with the parameters of the choice model at the second stage by using the EM algorithm as detailed in Sfeir et al. (2021). Gaussian mixtures can be considered as a more generalised approach compared to soft K-means, since they can capture non-spherical clusters. On the other hand, K-means is a much faster approach requiring less parameters for the estimation of the centroids.

2.3.2. Gaussian process - latent class choice model

In this non-parametric variant proposed in Sfeir et al. (2022), Gaussian Processes (GP) were used to replace the gaussian/bernoulli mixtures and perform the probabilistic class allocation to the latent classes. Initially, GP approaches were mostly used in geostatistics, a process known as kriging (MacKay, 1998), but in recent decades, they have become more popular for other ML tasks, as well, due to the recent computational improvements.

According to GP, a set of gaussian distributions equal to the number of observations in the sample can be estimated in order to uncover the data generating process. A GP is characterised by the mean of the gaussian distributions and their covariance defined by a kernel function $k(S_n, S_m)$ measuring the variance between any pair of observations/individuals m, n. It is typical to assume a zero mean function, thus leading to $GP(m, S) = GP(0, k(S_n, S_m))$. Some of the most commonly used kernel functions are the radial basis function and its generalisation the matérn kernel. Many other options exist, such as the dot product, the exponential, the constant, and the radial quadratic, while combinations of those can also be used by adding or multiplying them together. This provides a high level of flexibility to the analyst to use a wide range of approaches for a given sample (see Sfeir et al. (2022) for more details). The parameters of the GP depend on the number of observations, thus making it increasingly more difficult to estimate GPs with larger sample sizes.

3. Data

Our empirical work makes use of three case studies, with data from two separate datasets, which are described in the following. Case studies 1 and 2 use a GPS trip diary (“DECISIONS”), while Case study 3 uses a traditional pen and paper trip diary for London.

3.1. DECISIONS dataset

Our first dataset was collected between October 2016–March 2017 as part of the research project “DECISIONS” conducted at the University of Leeds. The dataset includes several submodules aiming to capture different aspects of everyday individual behaviour, such as indoor/outdoor activity behaviour, energy consumption, and social network formation. More information on the range of the different submodules of the dataset can be found in Calastri et al. (2020). In the current study, we focus on two specific submodules, namely a 2-week GPS-based trip diary captured through a smart-phone application and a household survey capturing important sociodemographic information of the participants. While the survey captured trips across all of the UK, the vast majority of them are in the Yorkshire region and more specifically the city of Leeds.

Two different subsets of the DECISIONS dataset were used, namely one for the mode choice *(Case study 1)* and one for the shopping destination choice model *(Case study 2)*. For both of them, only the trips in Yorkshire were selected. Data enriching steps followed the initial data cleaning stages, during which the dataset was augmented with travel time and travel cost information for all the alternatives. More specifically, travel times and distances were estimated for both chosen and unchosen alternatives (for consistency reasons) using a combination of the “Directions” Google API and Bing maps API. Both APIs allow for a detailed routing plan between an Origin and a Destination for different transport modes and times of day, while also accounting for traffic for car trips and for service timetable for public transport (PT) trips. Travel costs for car trips were calculated using WEBTag’s official specifications for fuel and operating costs, while bus and rail travel costs were calculated based on average distance-based costs of PT services operating in the region. A discount was applied for season ticket holders. The final mode choice dataset used for model estimation included 12,524 trips by 540 individuals, with a choice set of six alternative modes of transport, namely car, bus, rail, taxi, cycling and walking. Out of those trips, 47.6% were by car, 14.6% by bus, 5.2% by rail, 3.2% by taxi, 3.3% by cycling and finally 26.1% by walking.

A further subset of the mode choice dataset was selected for the shopping destination choice model (*Case study 2*), which included only shopping trips (for groceries, clothes and durables) from an initial origin O to a shopping location j and the trips to the following destination D. In total, 1,541 trip pairs (by 270 individuals) were included in this dataset, with 82% of them being for groceries, 12.7% for clothing and 5.3% for durables. The purpose of including the subsequent trip was to study the impact of the next destination D (considered fixed in this study) on the choice of the intermediate shopping location. The choice set in the destination model was defined by clustering the observed elemental locations utilising the Hierarchical Agglomerative Clustering (HAC) algorithm with a 800 m distance threshold. HAC was chosen since it does not require the analyst to make a priori assumptions regarding the number of clusters. The aforementioned procedure resulted in the creation of a choice set of 176 shopping destinations, most of them within the administrative boundaries of the local authority of Leeds. The main shopping mall of Leeds city centre attracted the majority of shopping trips, namely 11.3%. The remaining 5 shopping locations in the city centre attracted 9.7% of trips followed by a further 103 locations spread around the city of Leeds (62.6%) and finally 67 locations in the remaining region of Yorkshire (16.7%). More details regarding the data cleaning/enriching steps and the approach followed to define the availability of mode and destination alternatives in both subsets can be found in the studies of Tsoleridis et al. (2021, 2022a,b), respectively.

3.2. London travel demand survey

The dataset used for *Case study 3* is the openly available London Passenger Mode Choice (LPMC), collected as part of the London travel demand survey, in which travellers' choice of mode of transport among walking, cycling, transit and car was recorded. The dataset was augmented at a later stage by Hillel et al. (2018) with travel cost and travel time information for chosen and unchosen alternatives using the "Directions" Google API, in a similar to the DECISIONS dataset. An additional interesting variable was defined during that data enriching stage measuring the traffic variability for car trips as captured by the different routing procedures of the Google API. More details about the specific dataset can be found in Hillel et al. (2018). For the current application, a subset of only home-based trips by individuals of at least 12 years of age was selected, similarly to the study of Krueger et al. (2020) and Hancock et al. (2021). The resulting dataset contains a total of 58,584 trip observations performed by 26,904 individuals.

In terms of the observed mode choices, 42.8% of trips were made by car, followed by 37.6% of PT trips, 16.6% walking and finally 3.2% cycling trips. With regard to socio-demographic, 53.5% of individuals are female, the mean age is 42 years old, and 69.8% of participants have at least one car in their household. Notwithstanding the richness of individual mobility information, an important limitation of the London dataset is the absence of income information, either personal or household.

4. Results

The proposed hybrid specification, *H-LCCM*, which is an integration of a clustering algorithm used in data mining, K-means, with an econometric MNL model, is compared against the following structures:

***MNL-base*:** a traditional MNL model, where heterogeneity is captured with demographic interactions with the ASCs.

***C-MNL*:** a two-stage clustering model, where K-means is used at the first stage to allocate individuals into latent clusters based solely on sociodemographic characteristics, and then a choice model is estimated per cluster at the second stage, with no feedback loop from the choice model to the clustering algorithm. The final log-likelihood of that model is calculated by adding the log-likelihoods of the K cluster-specific models and the remaining fit statistics are computed relative to that. It should be noted that the same specification can be defined with different clustering algorithms for the first stage and/or different distance measures. For the purposes of the current study, we have tried using K-harmonic means (Zhang et al., 1999), DBSCAN (Ester et al., 1996), as well as K-means with Mahalanobis distance (Ghorbani, 2019) instead of Euclidean distance. The analysis, which also includes a train/test set validation assessment (80%/20% split) is presented in the Appendix (Tables 20–22). Here for the sake of brevity we only report the results of the K-means based C-MNL model.

***LCCM*:** a traditional latent class model with a class allocation model determining the probability that an individual falls within each class and a class-specific MNL model aiming to explain the observed choices.

***GBM-LCCM* and *GP-LCCM*:** the models from Sfeir et al. (2021) and Sfeir et al. (2022). A full covariance matrix was favoured in all cases for *GBM-LCCM*, which resulted in better model fit than the remaining options (Sfeir et al., 2021). For the hyperparameters of *GP-LCCM*, due to the vast range of possible kernel combinations, a non-exhaustive search was performed focussed around the dot product, the radial basis function and the matérn kernel, with the latter leading to a better model fit and more stable results (Sfeir et al., 2022).

In all cases, the same specification was used in terms of covariates in the clustering/class allocation model and explanatory variables in the utility functions to ensure consistency in our evaluation comparison. The number of classes and the specified covariates in the final specification reported in the following for each case refer to the one which resulted in the best model fit for the traditional *LCCM* model. For the clustering methods examined, a common way to determine the optimal number of classes is to use the *"elbow"* method by assessing the Silhouette score across an increasing number of clusters. In the case examined, however, even increasing the number of clusters by one in each case resulted in dissolving clusters for Case study 1 and issues with the covariance matrix in Case study 3. We have not attempted to increase the number of clusters in Case study 2 due to the already high computational cost of estimating that model.

For the fit statistics reported in the following Case studies, the centroids of the H-LCCM models are not included as additional parameters because they are only indirectly estimated through the estimation of the γ parameters. We instead include the estimated γ parameters as such. We did not opt to include both the centroids and the γ as parameters in the model as that would result in double counting of the relevant parameters. Furthermore, for the GP-LCCM models only the models' parameters and the kernel hyperparameters (smoothness υ and length-scale l for the matérn kernel) were considered in the fit statistics – and not the total number of observations – following the approach in Sfeir et al. (2022).

In terms of using AIC and BIC, there are limitations of course — but in absence of better measures, these have been widely used in the literature. Among the two, the key difference lies in how heavily they penalise model complexity, with AIC generally favouring slightly more complex models compared to BIC, especially when dealing with large datasets where BIC becomes more stringent in selecting simpler models; essentially, AIC prioritises prediction accuracy while BIC aims to identify the "true" model generating the data if it is within the candidate models. According to Chakraborty and Ghosh (2011), BIC is more useful in selecting a correct model, while the AIC is more appropriate in finding the best model for predicting future observations. Besides the fit statistics, the model assessment also focused on analysing the quality of the estimated classes/clusters. A common metric being

used for that is the Silhouette score that measures how well separated and homogeneous the clusters are. In this study, we mainly focused instead on the behavioural examination of the estimated classes/clusters and what they mean in terms of the sensitivities of the underlying population. Nonetheless, we do mention the Silhouette scores, as well, and specifically how they improve across the H-LCCM iterations from the very first randomly selected centroids to the final calibrated ones.

Model instability was a common denominator in all models examined that were able to jointly capture latent segments and taste heterogeneity, namely *LCCM*, *GBM-LCCM*, *GP-LCCM* and *H-LCCM*. To counter that, several attempts were performed for each model and case study and the best performing (either in terms of model fit and/or parameter explainability) was kept and reported in the following.

4.1. Case study 1: Yorkshire mode choice

4.1.1. Model specification

The specification of the Yorkshire mode choice model presented in the following contains 5 alternative specific constants (ASCs) with the ASC for car being kept fixed as the base. Mode-specific linear travel time sensitivities were specified in addition to a logarithmic generic specification for travel cost for the purpose of capturing cost damping effects (Daly, 2010).

4.1.2. Model outputs

For Case study 1, we were able to estimate all six specifications, namely *MNL-base*, *C-MNL*, *LCCM*, *GBM-LCCM*, *GP-LCCM* and *H-LCCM*, with their respective fit statistics being presented in Table 1. All models that capture unobserved heterogeneity among individuals are able to outperform the *MNL-base* model, even the *C-MNL*, in which the sample is segmented into clusters solely based on socio-demographic characteristics. As expected, however, more significant heterogeneity can be captured by including the observed choice behaviour in the process of calibrating the class allocation, as illustrated by the remaining four LCCM-based models. Out of those four specifications, the proposed *H-LCCM* is able to outperform the rest in terms of model fit. More specifically, it provided a better model fit compared to the traditional *LCCM* by 6.05 LL units with 19 parameters less. It also outperformed the two ML-based variants of *GBM-LCCM* with a full covariance matrix and *GP-LCCM* with a matérn kernel by 6.55 and 63.46 LL units with 54 less and 3 more parameters, respectively. Those improvements in model fit are more evident by looking at the adjusted rho^2, the AIC and BIC statistics in Table 1.

A closer comparison between the estimated parameters of *LCCM* and *H-LCCM* is shown in Table 2. For the sake of brevity, the parameters of *MNL-base*, *C-base*, *GBM-LCCM* and *GP-LCCM* have been omitted from that table, but are included in the Appendix (Table 14). Furthermore, the estimated covariances for the continuous elements of the class allocation in *GBM-LCCM* are shown in Table 17. The specification search resulted in a model with 5 classes and with gender, age, number of cars, season ticket ownership, and household income in the class allocation model. An equivalent specification was estimated for *H-LCCM* (and for the remaining *C-MNL*, *GBM-LCCM* and *GP-LCCM*). An increase to 6 classes resulted in numerical issues in the covariance matrix for *LCCM*, *GBM-LCCM*, *GP-LCCM* and in dissolving classes for *H-LCCM*, hence no attempt was made to estimate a model with 6 clusters for the remaining models under examination.

Overall, the *H-LCCM* results in more balanced cluster membership probabilities compared to *LCCM*, with cluster 5 representing the largest segment of the sample (23.0%) followed by cluster 2 (21.0%). The Silhouette score for *H-LCCM* increased by 8.5% from 0.2360 to 0.256 between the first randomly selected centroids and the final calibrated ones. Furthermore, by examining the estimated distance multipliers γ of *H-LCCM*, it is evident that individuals of cluster 5 are allocated to their class with a higher probability relative to others (48.2% to cluster 5 on average), while there is a higher degree of uncertainty in the allocation of individuals of cluster 4 (26.0% to cluster 4 on average). On the other hand, *LCCM* leads to more imbalanced class allocation with the majority of the sample (42%) being allocated to class 1. All level-of-service parameters have the expected negative sign in both models. A cost-insensitive class is identified in *LCCM*, class 5, representing the smallest segment of the sample (9%). Contrary to that, all clusters of *H-LCCM* show significant cost sensitivities, which illustrates the discrepancies of the two approaches in the heterogeneity they are able to capture. The estimated parameters of the covariates used in the class allocation of *LCCM* are presented in the same Table 2. The respective parameters for *H-LCCM* are obtained from an FMNL model on the cluster-specific probabilities for each individual as described in Section 2.2 and are reported in Table 3. In both cases, class/cluster 5 was used as the base and the remaining parameters were estimated relative to that.

A closer look at the average probabilities per sociodemographic group and their respective average values allows us to get a better understanding of the profile of each class/cluster (Fig. 4). Regarding the classes resulting from the *LCCM*, class 1 is more likely to contain car dependent (average number of cars=1.1/average season ticket ownership=0.16), middle-aged individuals (average age=42.8) of higher household income (average income=£55,731). Class 2 is more likely to contain individuals who are frequent public transport users (average season ticket ownership=0.6) have a lower than average number of cars in their household (average number of cars=0.76) and a lower household income (average income=£40,760). Class 3 is more likely to have younger (average age=29.2) female (average value for female=0.63) individuals with both a low number of season ticket ownership (average season ticket ownership=0.12) and number of cars (average number of cars=0.66). Class 4 has a higher share of younger (average age=34.5) male (average value for female=0.5) individuals, and finally class 5 has the highest share of female individuals (average value for female=0.72) with the lowest number of cars (average number of cars=0.4) and a high season ticket ownership (average season ticket ownership=0.46).

Moving on to the behavioural profiling of the clusters estimated from *H-LCCM*, cluster 1 can be characterised by mostly higher income individuals (average income=£59,650) with a high number of cars in their household (average number of cars=1.15). Cluster

Table 1
Fit statistics of the Yorkshire mode choice models.

Fit statistics	MNL-base	C-MNL	LCCM	GBM-LCCM	GP-LCCM	H-LCCM
Log-likelihood (0)	−14,974.45					
Log-likelihood (model)	−5,275.42	−4,928.32	−3,946.96	−3,946.55	−4,003.46	−3,940.00
Adjusted ρ^2	0.6469	0.6669	0.7308	0.7285	0.7286	0.7325
AIC	10,574.83	9,976.64	8,061.91	8,131.10	8,130.92	8,010.00
BIC	10,664.05	10,422.76	8,686.48	9,015.91	8,591.91	8,493.31
Number of parameters	12	60	84	119	62	65
Number of individuals	540					
Number of observations	12,524					

2 contains a high share of female individuals (average value for female=0.60) together with the highest share of season ticket holders (average season ticket ownership=0.56), a low number of cars (average number of cars=0.64) and low household income (average income=£43,970). Cluster 3 has a high share of younger individuals (average age=32.1) with a low number of cars (average number of cars=0.64) and of low income (average income=£44,140). They are also the most cost sensitive according to their estimated travel cost parameter. Cluster 4 can be characterised by younger (average age=36.86) female (average value for female=0.60) with the lowest share of season ticket ownership (average season ticket ownership=0.11). Finally, cluster 5 contains the highest share of older (average age=44.91) male (average value for female=0.53) individuals with a higher than average number of cars (average number of cars=0.97) and with a quite low season ticket ownership (average season ticket ownership=0.18).

As a further measure of validation, the estimated Values of Travel Time (VTT) estimates for each class and the weighted ones are presented in Table 4. The VTT estimates from all models are close to the latest official values suggested by the Department for Transport (no official VTT estimates for taxi) (Batley et al., 2019) with the exception of *GP-LCCM* resulting in significantly higher VTTs. The reason for the higher VTTs of *GP-LCCM* is the low and non-statistically significant travel cost parameter for class 2 (see Table 14), which increases the weighted average VTT, as well. Furthermore, in the same model there is a positive and non-significant cost parameter for the travel time parameter for taxi in cluster 5. A similar problem, but to a lesser extent, occurs in the econometric *LCCM* with a small and non-significant cost sensitivity for class 5. That finding can act as a further supporting argument for considering the *H-LCCM* framework for real-world policy making, since it has the ability to lead to more behaviourally accurate valuation measures, at least in the current study.

Mode shares, both class-specific and weighted, were also calculated across models and presented in Table 5. Using the MNL-base mode shares as a guide, since they are guaranteed to match the observed ones in the sample, we can observe that the proposed *H-LCCM* leads to weighted VTTs that are very close to them. On the contrary, the *GP-LCCM* shows the most significant discrepancies with almost a double bus share and lower car and walking shares. The higher bus shares are mostly driven by the high bus share in class 2, namely 82%. On the other hand, the *GBM-LCCM* performs leads to almost equal shares with the proposed *H-LCCM*.

4.2. Case study 2: Yorkshire shopping destination choice

4.2.1. Model specification

The specification of the models presented in the following is based on the size variable specification of Daly (1982) and more specifically of Kristoffersson et al. (2018). According to that, the attraction of a destination j is captured with the addition in the utility function of a composite term A_j inside a logarithmic function, i.e. $log(A_j)$. The composite term A_j includes various variables aiming at capturing the attraction of the target destination j and their respective parameters, as $A_j = a_{1j} + \sum_{r>1} exp(\gamma_r) a_{rj}$, where one attraction variable, a_{1j}, is defined as the base and its parameter is kept fixed to 1.0. For simplification purposes, in the current case study we have used only one attraction variable inside the logarithm pertaining to total retail area per type of shopping activity (clothes shopping, grocery, durables) in a 400 m buffer around the shopping destination.

The final specification was able to uncover 2 latent classes/clusters of heterogeneous decision-makers using annual personal income and the areal measure of Index of Multiple Deprivation (IMD) calculated for a 400 m buffer around home locations. The IMD is a composite measure developed by the Office for National Statistics (ONS) aimed to capture deprivation among a range of different domains, such as crime, environment and housing among others and at a high spatial resolution (Lower Super Output Areas). The IMD is calculated as a weighted measurement of the constituent deprivation domains with a higher number signifying a more deprived area. More details can be found at the IMD technical report in Smith et al. (2015). The IMD indices for the year 2015 were used in the current study.

4.2.2. Model outputs

The entire range of models was estimated for this case study, although we had to simplify the specifications to be able to estimate the *GBM-LCCM* and *GP-LCCM* models due to the increased computational requirements as a result of the large choice set (176 destination alternatives). Because both our class allocation covariates are continuous variables, only the Gaussian Mixture Modelling part of the GBM-LCCM was required.[3] The fit statistics for *Case study 2* are presented in Table 6. Capturing latent

[3] The model is still described as GBM-LCCM to avoid any confusion for the reader.

Table 2
Modelling estimates and t-ratios of LCCM and H-LCCM models for the Yorkshire mode choice context.

Parameter	LCCM	H-LCCM
Alternative-specific constants		
Constant Car (base)	–	–
Constant Bus - class 1	−5.2074 (−11.22)	−3.5083 (−4.09)
Constant Bus - class 2	−1.3102 (−3.73)	−1.5507 (−3.26)
Constant Bus - class 3	0.3470 (0.33)	1.1644 (1.02)
Constant Bus - class 4	−3.7784 (−5.22)	−5.0562 (−4.64)
Constant Bus - class 5	−2.4179 (−3.81)	−3.6437 (−5.16)
Constant Rail - class 1	−2.7486 (−2.95)	−0.1604 (0.21)
Constant Rail - class 2	−0.4778 (−0.49)	−0.8823 (−1.06)
Constant Rail - class 3	−4.7155 (−3.03)	−3.4326 (−2.25)
Constant Rail - class 4	−1.6476 (−1.35)	−3.9105 (−5.95)
Constant Rail - class 5	−8.1766 (−5.02)	−3.7222 (−2.71)
Constant Taxi - class 1	−4.7217 (−6.57)	−3.9144 (−5.59)
Constant Taxi - class 2	−6.5988 (−3.28)	−2.1399 (−3.47)
Constant Taxi - class 3	3.4928 (3.06)	4.8633 (3.75)
Constant Taxi - class 4	−0.6965 (−1.18)	−2.8189 (−3.58)
Constant Taxi - class 5	−3.9916 (−4.02)	−4.6772 (−6.55)
Constant Cycling - class 1	39.8735 (3.84)	−5.0748 (−5.10)
Constant Cycling - class 2	−1.6347 (−1.85)	−4.2522 (−3.45)
Constant Cycling - class 3	−2.0753 (−1.74)	−3.6530 (−2.96)
Constant Cycling - class 4	−2.5533 (−4.09)	−2.0757 (−2.92)
Constant Cycling - class 5	−1.5036 (−1.04)	−2.9096 (−1.81)
Constant Walking - class 1	0.2920 (0.46)	2.4618 (2.16)
Constant Walking - class 2	1.4847 (2.61)	0.8567 (1.30)
Constant Walking - class 3	4.4302 (4.81)	4.1692 (4.72)
Constant Walking - class 4	2.1666 (2.80)	0.8978 (1.02)
Constant Walking - class 5	−0.4120 (−0.54)	0.2404 (0.41)
LOS parameters		
Car travel time (mins) - class 1	−0.1712 (−8.78)	−0.0809 (−3.12)
Car travel time (mins) - class 2	−0.0794 (−1.19)	−0.2629 (−7.09)
Car travel time (mins) - class 3	−0.1312 (−3.08)	−0.1692 (−2.04)
Car travel time (mins) - class 4	−0.1040 (−3.23)	−0.1383 (−3.93)
Car travel time (mins) - class 5	−0.2688 (−4.74)	−0.0259 (−0.78)
Bus travel time (mins) - class 1	−0.0795 (−8.48)	−0.0626 (−4.81)
Bus travel time (mins) - class 2	−0.0323 (−1.16)	−0.1047 (−7.54)
Bus travel time (mins) - class 3	−0.1113 (−2.94)	−0.1556 (−5.13)
Bus travel time (mins) - class 4	−0.0645 (−4.30)	−0.0527 (−3.87)
Bus travel time (mins) - class 5	−0.0937 (−4.93)	−0.0088 (−2.87)
Rail travel time (mins) - class 1	−0.1017 (−8.28)	−0.0575 (−4.44)
Rail travel time (mins) - class 2	−0.0727 (−2.78)	−0.1583 (−8.65)
Rail travel time (mins) - class 3	−0.0230 (−0.71)	−0.0287 (−1.44)
Rail travel time (mins) - class 4	−0.0380 (−2.20)	−0.0641 (−1.81)
Rail travel time (mins) - class 5	−0.0619 (−2.15)	−0.1505 (−4.52)
Taxi travel time (mins) - class 1	−0.2324 (−5.56)	−0.0748 (−1.55)
Taxi travel time (mins) - class 2	0.0058 (−0.06)	−0.2888 (−6.03)
Taxi travel time (mins) - class 3	−0.3930 (−5.04)	−0.4942 (−5.81)
Taxi travel time (mins) - class 4	−0.1285 (−2.20)	−0.1200 (−3.34)
Taxi travel time (mins) - class 5	−0.1891 (−2.96)	−0.1022 (−1.77)
Cycling travel time (mins) - class 1	−8.8323 (−4.29)	−0.0700 (−2.06)
Cycling travel time (mins) - class 2	−0.3935 (−4.87)	−0.2064 (−2.73)
Cycling travel time (mins) - class 3	−0.1425 (−3.92)	−0.0796 (−2.54)
Cycling travel time (mins) - class 4	−0.0621 (−4.14)	−0.0688 (−5.25)
Cycling travel time (mins) - class 5	−0.1287 (−3.05)	−1.9995 (−6.35)
Walking travel time (mins) - class 1	−0.2024 (−10.31)	−0.2226 (−7.56)
Walking travel time (mins) - class 2	−0.1535 (−5.48)	−0.2126 (−8.78)
Walking travel time (mins) - class 3	−0.1178 (−4.13)	−0.1404 (−5.40)
Walking travel time (mins) - class 4	−0.2245 (−8.75)	−0.1948 (−6.58)
Walking travel time (mins) - class 5	−0.1780 (−5.38)	−0.1394 (−6.82)
Natural logarithm of travel cost (£) - class 1	−0.3611 (−1.75)	−0.9863 (−3.70)
Natural logarithm of travel cost (£) - class 2	−0.5080 (−3.00)	−0.4647 (−2.36)
Natural logarithm of travel cost (£) - class 3	−1.2605 (−3.87)	−1.7901 (−5.83)
Natural logarithm of travel cost (£) - class 4	−1.7146 (−10.47)	−0.7184 (−2.61)
Natural logarithm of travel cost (£) - class 5	−0.2245 (−0.81)	−0.9709 (−4.51)

(*continued on next page*)

Table 2 (*continued*).

Parameter	LCCM	H-LCCM
Class allocation parameters		
Constant - class 1	1.4520 (1.42)	–
Season ticket ownership - class 1	−1.2008 (−2.33)	–
Number of cars in household - class 1	2.2219 (4.19)	–
Age - class 1	−0.0078 (−0.42)	–
Female - class 1	−0.7680 (−1.39)	–
Annual household income (£1,000) - class 1	−0.0068 (−0.69)	–
Constant - class 2	1.3730 (1.28)	–
Season ticket ownership - class 2	0.7042 (1.25)	–
Number of cars in household - class 2	1.8080 (3.15)	–
Age - class 2	−0.0054 (−0.28)	–
Female - class 2	−0.9572 (−1.60)	–
Annual household income (£1,000) - class 2	−0.0287 (−2.44)	–
Constant - class 3	4.2390 (3.68)	–
Season ticket ownership - class 3	−2.2309 (−2.38)	–
Number of cars in household - class 3	1.7204 (2.69)	–
Age - class 3	−0.1023 (−3.40)	–
Female - class 3	−0.6773 (−0.95)	–
Annual household income (£1,000) - class 3	−0.0143 (−1.29)	–
Constant - class 4	3.7335 (3.45)	–
Season ticket ownership - class 4	−1.0282 (−1.62)	–
Number of cars in household - class 4	2.2784 (4.34)	–
Age - class 4	−0.0640 (−3.00)	–
Female - class 4	−1.3283 (−2.15)	–
Annual household income (£1,000) - class 4	−0.0180 (−1.58)	–
Clustering distance parameters		
Distance multiplier γ - cluster 1	–	−0.9110 (3.35)
Distance multiplier γ - cluster 2	–	−0.7202 (3.60)
Distance multiplier γ - cluster 3	–	−0.3428 (1.50)
Distance multiplier γ - cluster 4	–	−0.1812 (1.08)
Distance multiplier γ - cluster 5	–	−1.0880 (5.05)
Class/cluster membership probabilities		
Class/cluster 1	0.42	0.20
Class/cluster 2	0.18	0.21
Class/cluster 3	0.10	0.17
Class/cluster 4	0.21	0.19
Class/cluster 5	0.09	0.23

Table 3
Estimated parameters of clustering covariates for the Yorkshire mode choice model.

Parameters	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Constant	−0.1794 (−3.00)	0.9519 (21.70)	1.8562 (42.92)	0.9197 (19.95)	–
Season ticket ownership	0.1162 (4.03)	1.6027 (57.54)	0.2742 (11.60)	0.3584 (16.20)	–
Number of cars	0.3962 (16.11)	−0.1814 (−8.87)	−0.2455 (−11.88)	−0.0987 (−4.70)	–
Age	−0.0342 (−26.28)	−0.0418 (−37.90)	−0.0490 (−42.28)	−0.0405 (−38.49)	–
Female	0.2148 (6.45)	0.2221 (7.82)	−0.2797 (−10.18)	0.6297 (24.00)	–
Annual household income	0.0172 (24.39)	0.0032 (5.36)	0.0027 (5.49)	0.0034 (7.04)	–

heterogeneity resulted again in model fit improvements compared to *MNL-base*, even with the simpler *C-MNL* specification. The *GBM-LCCM* model achieved a fit of −3585.41, while *LCCM* managed to achieve a better one at −3578.01. Finally, the two remaining models achieved the best results with *H-LCCM* achieving an LL of −3573.32 and *GP-LCCM* a slightly better LL of −3573.03 with the same number of parameters. The above are also evident from the adjusted r^2, AIC and BIC statistics presented in Table 6. Finally, it is worth mentioning that during the H-LCCM cluster calibration, the Silhouette score increased by 57.3% from 0.2432 (for the initial randomly sampled centroids) to 0.3825 (for the final calibrated centroids) suggesting that the quality of the clusters improves during calibration.

The estimated parameters of *LCCM* and *H-LCCM* are presented in Table 7. A more detailed version of that Table is included in the Appendix (Table 15) showing also the estimated parameters of *MNL-base*, *C-MNL*, *GBM-LCCM* and *GP-LCCM*, while the lower triangular matrix of the covariates for the GBM-LCCM is presented in Table 18. Focusing on *LCCM* and *H-LCCM*, the estimates of the two specifications are almost identical with only negligible discrepancies. All of the variables were allowed to vary across classes capturing significant taste differences. The shopping destinations at the centre of Leeds were selected as the base alternatives. The specified ASCs were grouped separately for the remaining destinations of Leeds outside of the city centre and the destinations located in the remaining region of Yorkshire. Additional interactions were also specified for season ticket holders and individuals with no car in their household for the destinations outside the city centre and outside Leeds. The purpose of those interactions was to capture the additional disutility of travelling to those places, which are located further away from the city centre with worse

Table 4
Class-specific and weighted Values of Travel Time estimates (£/hr) for the Yorkshire mode choice dataset.

VTT estimate	MNL-base	C-MNL	LCCM	GBM-LCCM	GP-LCCM	H-LCCM
Car - class 1	–	19.11	39.98	25.59	21.25	7.21
Car - class 2	–	22.00	15.75	9.92	489.68	49.74
Car - class 3	–	14.42	15.30	6.27	3.25	8.31
Car - class 4	–	7.64	4.75	20.21	42.55	16.92
Car - class 5	–	13.19	161.63	8.96	3.30	2.33
Car - weighted	12.01	15.87	34.80	17.33	110.74	17.56
Bus - class 1	–	11.66	18.33	11.79	10.21	5.54
Bus - class 2	–	5.24	5.88	6.33	181.41	12.71
Bus - class 3	–	7.11	9.89	3.63	1.30	7.58
Bus - class 4	–	10.03	3.27	5.70	23.81	19.64
Bus - class 5	–	15.44	53.23	6.11	2.03	0.79
Bus - weighted	6.14	7.52	16.93	7.89	43.69	12.71
Rail - class 1	–	64.26	93.82	50.85	38.80	20.84
Rail - class 2	–	32.37	54.05	21.95	959.1	121.76
Rail - class 3	–	53.13	28.82	21.16	18.53	5.73
Rail - class 4	–	34.36	7.60	35.71	126.17	31.91
Rail - class 5	–	55.46	137.87	1.76	16.91	55.38
Rail - weighted	28.20	35.84	68.74	33.55	227.20	49.45
Taxi - class 1	–	215.4	283.60	186.78	161.09	33.58
Taxi - class 2	–	132.10	18.42	96.03	2412	275.20
Taxi - class 3	–	93.38	126.55	14.56	54.57	122.25
Taxi - class 4	–	79.71	32.04	111.29	204.6	73.98
Taxi - class 5	–	125.13	558.70	134.27	−2.23	46.60
Taxi - weighted	82.11	122.81	198.33	127.62	568.24	118.38

Table 5
Class-specific and weighted mode shares for the Yorkshire mode choice dataset.

VTT estimate	MNL-base	C-MNL	LCCM	GBM-LCCM	GP-LCCM	H-LCCM
Car - class 1	–	0.39	0.52	0.52	0.51	0.49
Car - class 2	–	0.26	0.43	0.47	0.12	0.42
Car - class 3	–	0.76	0.41	0.41	0.45	0.37
Car - class 4	–	0.41	0.47	0.45	0.31	0.46
Car - class 5	–	0.70	0.37	0.37	0.44	0.52
Car - weighted	0.48	0.50	0.47	0.46	0.38	0.46
Bus - class 1	–	0.10	0.09	0.09	0.07	0.09
Bus - class 2	–	0.35	0.30	0.04	0.82	0.29
Bus - class 3	–	0.03	0.12	0.14	0.26	0.12
Bus - class 4	–	0.09	0.05	0.31	0.36	0.03
Bus - class 5	–	0.01	0.24	0.19	0.07	0.21
Bus - weighted	0.15	0.13	0.14	0.14	0.29	0.15
Rail - class 1	–	0.03	0.08	0.06	0.08	0.11
Rail - class 2	–	0.13	0.04	0.12	0.01	0.05
Rail - class 3	–	0.02	0.00	0.06	0.02	0.07
Rail - class 4	–	0.03	0.08	0.01	0.01	0.03
Rail - class 5	–	0.03	0.02	0.03	0.10	0.01
Rail - weighted	0.05	0.05	0.06	0.06	0.05	0.05
Taxi - class 1	–	0.02	0.04	0.07	0.06	0.03
Taxi - class 2	–	0.03	0.01	0.03	0.01	0.05
Taxi - class 3	–	0.01	0.07	0.04	0.01	0.06
Taxi - class 4	–	0.06	0.05	0.01	0.06	0.08
Taxi - class 5	–	0.02	0.05	0.07	0.02	0.01
Taxi - weighted	0.03	0.03	0.04	0.05	0.05	0.04
Cycling - class 1	–	0.04	0.00	0.01	0.01	0.00
Cycling - class 2	–	0.02	0.00	0.05	0.00	0.00
Cycling - class 3	–	0.05	0.02	0.15	0.00	0.02
Cycling - class 4	–	0.01	0.09	0.00	0.04	0.17
Cycling - class 5	–	0.03	0.16	0.00	0.13	0.00
Cycling - weighted	0.03	0.03	0.04	0.03	0.02	0.04
Walking - class 1	–	0.42	0.27	0.26	0.27	0.28
Walking - class 2	–	0.21	0.22	0.29	0.05	0.20
Walking - class 3	–	0.12	0.36	0.22	0.26	0.35
Walking - class 4	–	0.41	0.26	0.21	0.22	0.23
Walking - class 5	–	0.15	0.17	0.35	0.25	0.26
Walking - weighted	0.26	0.26	0.26	0.26	0.21	0.26

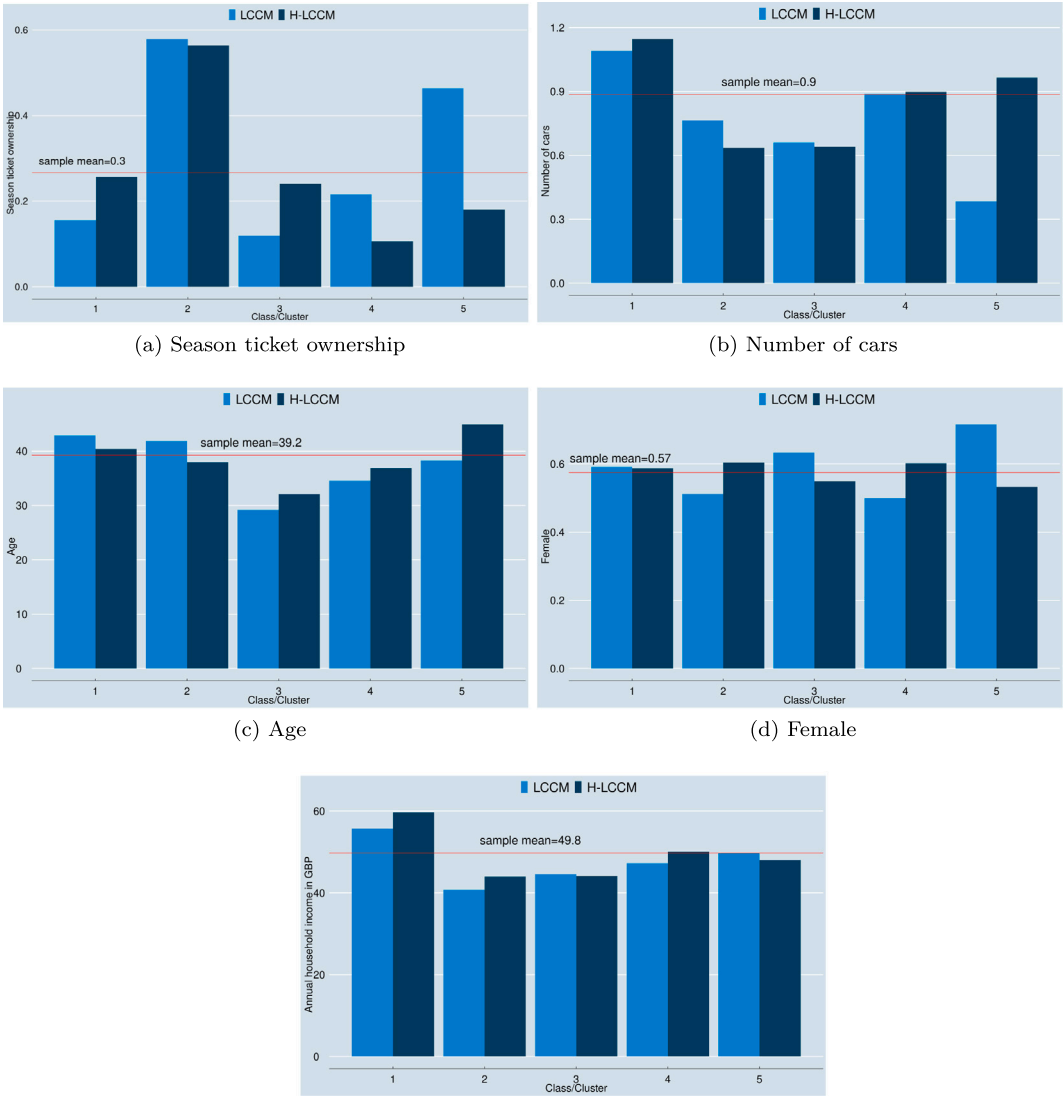


Fig. 4. Class-specific average values of covariates across LCCM and H-LCCM for Yorkshire mode choice model.

provision of PT infrastructure and without the convenience of a private vehicle. Significant non-linearities were captured travel cost and parking areas for car trips using a logarithmic transformation alongside a linear one.

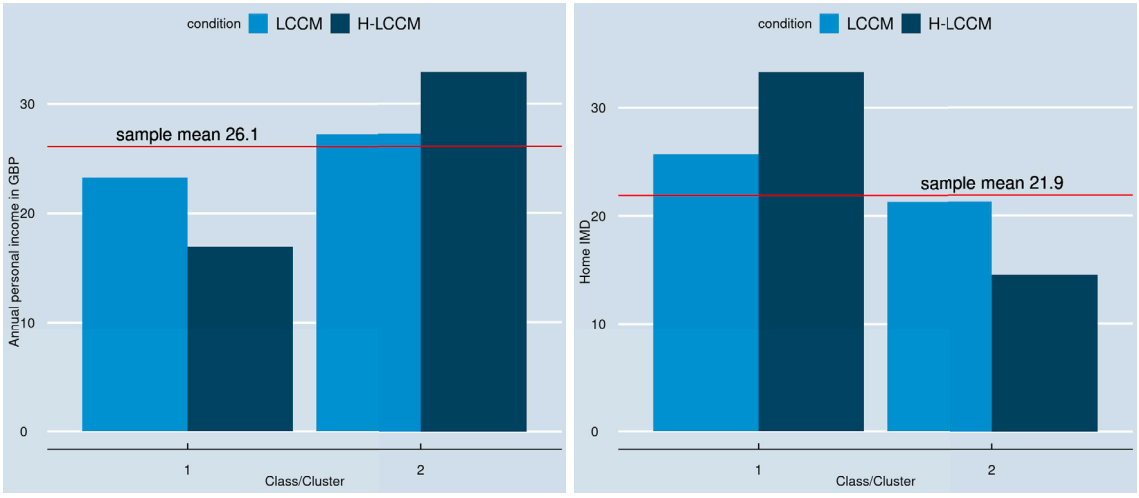
The class allocation of *LCCM* resulted in a sample segmentation with 37% of the sample allocated in class 1 and 63% in class 2. Similarly, the clustering procedure of *H-LCCM* allocated individuals by 42% to class 1 and 58% to class 2. According to the estimated distance multipliers γ of *H-LCCM*, the individuals of cluster 2 are allocated with a higher probability to their class (68.8% to class 2 on average) compared to individuals of class 1 (54.3% to class 1 on average).

The small number of identified classes allows us to perform an easier behavioural profiling compared to the five classes of *Case study 1*. Overall, the insights derived from the covariates of the class allocation, presented in Table 7, and the FMNL model on the cluster-specific probabilities for *H-LCCM*, shown in Table 8, are in agreement with the sensitivities between the two classes/clusters. Individuals allocated into class/cluster 1 have a lower personal income (average personal income=£23,258 from *LCCM* and £16,960 from *H-LCCM*) and are living in more deprived areas (average home IMD=25.7 from *LCCM* and 33.3 from *H-LCCM*), while also showing higher cost sensitivities. Furthermore, they show significantly lower car and public transport time sensitivities for the trip to the shopping location compared to the following, while the opposite is true for the walking sensitivities.

Individuals in class/cluster 2 are more likely to have a higher personal income (average personal income=£27,244 from *LCCM* and £32,900 from *H-LCCM*) and reside in less deprived areas (average home IMD=21.3 from *LCCM* and 14.51 from *H-LCCM*). They are characterised by lower and non significant cost sensitivities and almost equal car time sensitivities for the first and following trips

Table 6
Fit statistics of the Yorkshire shopping destination choice models.

Fit statistics	MNL-base	C-MNL	LCCM	GBM-LCCM	GP-LCCM	H-LCCM
Log-likelihood (0)	−7,961.332					
Log-likelihood (model)	−3,671.03	−3,640.98	−3,578.01	−3,585.41	−3,573.03	−3,573.32
Adjusted ρ^2	0.5369	0.5386	0.5462	0.5442	0.5469	0.5469
AIC	7,374.05	7,345.96	7,226.02	7,256.82	7,214.06	7,214.64
BIC	7,459.49	7,516.85	7,412.93	7,486.45	7,395.63	7,396.21
Number of parameters	16	32	35	43	34	34
Number of individuals	270					
Number of observations	1,541					



(a) Annual personal income (b) Home IMD

Fig. 5. Class-specific average values of covariates across LCCM and H-LCCM for Yorkshire shopping destination choice model.

to a shopping destination. That does not hold, however, for public transport time and walking distance sensitivities with individuals in class 2 having higher sensitivities for the second trip to the shopping destination meaning that they are more likely to choose to shop closer to their following activity location when they choose to travel by public transport or walking. The aforementioned comparison of the behavioural profiling of the estimated classes/clusters is also depicted in Fig. 5.

Regarding the remaining parameters, the presence of parking areas is a significant factor for car trips. That marginal utility is decreasing with the increase of parking spaces for *LCCM*, as captured by the estimated logarithmic parameter, while only linear significant sensitivities were captured with the *H-LCCM*. The directionality of travel is also an important factor with intermediate shopping destinations that require a significant deviation (above 90°) from the straight path between the previous origin and the following destination are less likely to be chosen compared to others. Both models were able to capture significant marginal utilities of the direction of travel for class 2, but not for class 1. Finally, the multiplier of the composite logarithmic term for the size variables is significantly less than 1.0. According to Kristoffersson et al. (2018) that implies the existence of correlation among the elemental alternatives inside the aggregated destination alternatives, thus providing a behavioural meaning behind the alternative aggregation, in that case the implementation of HAC.

4.3. Case study 3: London mode choice

4.3.1. Model specification

For the mode choice model for the London dataset, the *MNL-base* model follows the specification presented in Krueger et al. (2020) and Hancock et al. (2021). In addition to ASCs and interactions with socio-demographic (e.g. gender, number of cars, age etc.) and trip characteristics (e.g. month of the year), the utility function also includes generic in-vehicle travel time parameters for motorised modes (car, transit) and out-of-vehicle time for the access-egress segments of transit, cycling and walking. Moreover, there are parameters capturing the impact of traffic variability for car trips and the number of necessary transfers for transit trips. The specification of *LCCM* was able to identify two latent classes of individuals, while failing to identify a third class. The number of cars owned per household and the age of the individuals were used as covariates in the class allocation model in the absence of any measure of personal or household income. The same socio-demographics were also included in the class-specific mode choice models at the lower level with different parameters specified for each case, similar to the study of Calastri et al. (2018). The estimated parameters of each component will inform the analyst whether a specific socio-demographic attribute might be better at explaining

Table 7
Modelling estimates and t-ratios of LCCM and H-LCCM models for the Yorkshire shopping destination choice context.

Parameter	LCCM	H-LCCM
Alternative-specific constants		
Constant Leeds city centre (base)	–	–
Constant Remaining Leeds - class 1	−0.5628 (−0.55)	−0.8985 (−0.87)
Constant Remaining Leeds - class 2	−0.5195 (−1.34)	−0.5190 (−1.59)
Constant Remaining Leeds shift for season ticket owners/no car ownership - class 1	−1.9416 (−2.74)	−1.6475 (−1.15)
Constant Remaining Leeds shift for season ticket owners/no car ownership - class 2	−0.4993 (−0.61)	−0.3828 (−0.77)
Constant Remaining Yorkshire - class 1	0.3734 (0.31)	−0.1418 (−0.12)
Constant Remaining Yorkshire - class 2	0.1881 (0.52)	0.2503 (0.75)
Constant Remaining Yorkshire shift for season ticket owners/no car ownership - class 1	−0.4916 (−0.31)	−0.8747 (−0.74)
Constant Remaining Yorkshire shift for season ticket owners/no car ownership - class 2	−0.7664 (−0.91)	−0.6232 (−1.23)
LOS parameters		
Travel time car previous trip (mins) - class 1	−0.0523 (−0.76)	−0.0114 (−0.10)
Travel time car previous trip (mins) - class 2	−0.1379 (−3.96)	−0.1319 (−5.48)
Travel time car next trip (mins) - class 1	−0.2350 (−2.29)	−0.1178 (−1.13)
Travel time car next trip (mins) - class 2	−0.1217 (−3.77)	−0.1351 (−3.66)
Travel time PT previous trip (mins) - class 1	−0.0852 (−2.14)	−0.0289 (−0.61)
Travel time PT previous trip (mins) - class 2	−0.0201 (−0.74)	−0.0283 (−0.99)
Travel time PT next trip (mins) - class 1	−0.2748 (−2.15)	−0.1904 (−2.90)
Travel time PT next trip (mins) - class 2	−0.0631 (−1.67)	−0.0708 (−2.69)
Walking distance previous trip (km) - class 1	−2.8828 (−1.17)	−2.4770 (−3.31)
Walking distance previous trip (km) - class 2	−1.1975 (−3.32)	−1.2556 (−4.36)
Walking distance next trip (km) - class 1	−1.8370 (−0.51)	−1.3240 (−1.59)
Walking distance next trip (km) - class 2	−1.8779 (−1.15)	−2.1864 (−3.57)
Travel cost linear (£) - class 1	−2.4312 (−2.56)	−3.7554 (−4.20)
Travel cost linear (£) - class 2	−0.1222 (0.49)	−0.0847 (−1.10)
Travel cost log (£) - class 1	0.2629 (0.52)	0.2961 (0.69)
Travel cost log (£) - class 2	−0.1244 (−1.67)	−0.1078 (−0.76)
Direction of travel		
Presence of angle> 90º between O-S and O-D - class 1	−0.0242 (−0.04)	−0.0624 (−0.26)
Presence of angle> 90º between O-S and O-D - class 2	−0.4832 (−2.98)	−0.4562 (−2.80)
Locational variables		
Parking areas linear (400 m buffer) - class 1	0.0052 (0.58)	0.0087 (2.14)
Parking areas linear (400 m buffer) - class 2	0.0198 (4.93)	0.0195 (8.16)
Parking areas log (400 m buffer) - class 1	0.0663 (2.03)	0.0433 (0.68)
Parking areas log (400 m buffer) - class 2	0.0395 (2.21)	0.0521 (1.17)
Size variables		
Natural logarithm multiplier ϕ - class 1	0.2950 (0.71)	0.3295 (1.21)
Natural logarithm multiplier ϕ - class 2	0.5823 (2.92)	0.5640 (3.62)
Class allocation parameters		
Constant - class 1	−0.6320 (−0.86)	–
Annual personal income (£1,000) - class 1	−0.0150 (−1.18)	–
Home IMD - class 1	0.0211 (0.99)	–
Clustering distance parameters		
Distance multiplier γ - cluster 1	–	−0.1507 (0.27)
Distance multiplier γ - cluster 2	–	−0.7640 (2.06)
Class/cluster membership probabilities		
Class/cluster 1	0.37	0.42
Class/cluster 2	0.63	0.58

Table 8
Estimated parameters of clustering covariates for the Yorkshire shopping destination choice model.

Parameters	Cluster 1	Cluster 2
Constant	−0.3841 (−7.93)	–
Annual personal income (£1,000)	−0.0223 (−12.69)	–
Home IMD	0.0275 (19.76)	–

the allocation of the individuals into the classes/clusters or their observed choices. Increasing the number of clusters in *GBM-LCCM*, *GP-LCCM* and *H-LCCM* resulted in numeric issues in the covariance matrix, hence it was decided to use an equivalent specification for *C-MNL* for evaluation purposes, similar to the previous case studies already discussed.

Table 9
Fit statistics of the London mode choice models.

Fit statistics	MNL-base	C-MNL	LCCM	GBM-LCCM	GP-LCCM	H-LCCM
Log-likelihood (0)	−81,214.67					
Log-likelihood (model)	−44,309.01	−43,854.73	−37,597.26	−37,219.81	−37,527.57	−37,137.11
Adjusted ρ^2	0.4542	0.4596	0.5366	0.5411	0.5375	0.5423
AIC	88,654.03	87,773.46	75,272.52	74,533.62	75,131.14	74,350.23
BIC	88,815.63	88,060.76	75,622.67	74,955.60	75,472.31	74,691.40
Number of parameters	18	32	39	47	38	38
Number of individuals				26,904		
Number of observations				58,584		

4.3.2. Model outputs

For this case study, we have estimated six models in total, namely *MNL-base*, *C-MNL*, *LCCM*, *GBM-LCCM*, *GP-LCCM* and the proposed *H-LCCM*. The *GBM-LCCM* was estimated using a full covariance matrix and a matérn kernel was used for *GP-LCCM*. Due to the large number of observations in the dataset, a much higher estimation time was observed for *GP-LCCM* by a factor of more than 10, compared to the remaining specifications examined. The fit statistics of the six specifications are shown in Table 9. The estimated parameters of *LCCM* and *H-LCCM*, along with their robust t-ratios, are reported in Table 10. A more detailed version of that Table including also the estimates and robust t-ratios of *MNL-base*, *C-MNL*, *GBM-LCCM* and *GP-LCCM* is presented in Table 16 in the Appendix. The lower triangular matrix of the estimated covariances of the continuous covariates of the *GBM-LCCM* are presented in Table 19. Overall, the proposed specification is able to provide significant model fit improvements compared to *LCCM* with 460.15 LL units of improvement, while also having 1 parameter less. The remaining fit statistics of *H-LCCM* are also improved compared to *LCCM*.

Besides the improvements in model fit, however, the advantages of our proposed methodology are more evident in the behavioural interpretation of the estimated classes/clusters (Fig. 6). According to *LCCM*, class 1 is characterised by mostly older individuals (average age=44.6) with a lower than average number of cars in their households (average number of cars=0.89), while the opposite is true for class 2 including younger individuals (average age=39.1) with a higher than average number of cars in their possession (average number of cars=0.98). It is fair to say that the behavioural interpretation of the covariates of *LCCM* is not intuitive enough, since age and the number of cars were used as proxy measures of income. As such, our prior assumption was that older individuals would likely also be in the possession of more cars, relative to younger individuals, acting as a manifestation of an increased income accumulation over their lifetime.

Contrary to this, the clusters of *H-LCCM* represent a more intuitive behavioural profiling. First of all, based on the estimated distance multipliers γ, there is high certainty for allocating individuals into cluster 1 with a probability of 81.5% on average to belong to that cluster, while individuals of cluster 2 have an average probability of 61.4% to be allocated into cluster 2. Secondly, based on the estimated parameters of the FMNL model on the cluster-specific allocation probabilities per individual reported in Table 11, cluster 1 is more likely to include younger individuals (average age=38.9) with a lower than average number of cars (average number of cars=0.88), while older individuals (average age=44.7) with a higher than average number of cars (average number of cars=0.99) are more likely to be allocated into cluster 2. Individuals in cluster 1 are also more cost sensitive and less sensitive for in-vehicle car and transit time compared to individuals in cluster 2. Finally, it should be mentioned that for *H-LCCM* the Silhouette score increased by 37.8% between the first iteration with the randomly selected centroids (Silhouette score = 0.2860) and the final calibrated ones (Silhouette score = 0.3940).

A range of willingness-to-pay (WTP) estimates across models, both class-specific and weighted, are also presented in Table 12, specifically for in-vehicle travel times for car and transit, for out-of-vehicle times for transit, for traffic variability for car and for transit transfers. The values are similar across all models, but an interesting thing to note here is that the ML-assisted LCCMs, namely *GBM-LCCM*, *GP-LCCM* and *H-LCCM*, result in higher WTP for IVTT relative to the rest and specifically compared to *LCCM* by around 5£/hr. Taking into account the VTTs for the Yorkshire mode choice model (see Table 4) and the inherently increased average income of London residents, which is subset by the increased cost of living, the higher VTT for IVTT of *H-LCCM* presented in Table 12 could be considered as a more accurate behavioural representation of the trade-offs that individuals in London are willing to make.

Finally, the class-specific and weighted mode shares are presented in Table 13 with the *LCCM* and *H-LCCM* models achieving weighted shares closer to the ones obtained from the MNL model, which can perfectly match the observed shares in the sample. On the hand, the *GBM-LCCM* and *GP-LCCM* specifications lead to significantly lower car shares and higher walking, cycling and transit shares than the rest.

5. Conclusions

The present paper showcased the integration of a probabilistic clustering algorithm based on data mining techniques into a state-of-the-art econometric framework for the purpose of capturing individual heterogeneity in the sample. The novelty of our approach compared to existing studies in the literature is the transformation of a deterministic clustering algorithm into a probabilistic one in order to effectively take the role of a class allocation model without compromising the computational feasibility. That effectively allows for a simultaneous calibration of the clustering component and the estimation of a choice model at the second stage, which also provides feedback and helps the clustering component to re-adjust the centroids until convergence. It may be noted that there

Table 10
Modelling estimates and t-ratios of LCCM and H-LCCM for the London mode choice context.

Parameter	LCCM	H-LCCM
Alternative-specific constants		
Constant Walking (base)	–	–
Constant Cycling - class 1	−7.6354 (−7.05)	−7.7086 (−9.42)
Shift Cycling for females - class 1	−0.8494 (−0.75)	−7.1144 (−11.08)
Shift Cycling for winter (November–March) - class 1	1.1809 (0.91)	−1.5958 (−1.06)
Shift Cycling for age below 18 or above 64 - class 1	−9.5609 (−12.94)	−0.1846 (−0.14)
Constant Cycling - class 2	−3.2261 (−32.62)	−1.6507 (−11.51)
Shift Cycling for females - class 2	−1.1075 (−11.09)	−1.0203 (−6.50)
Shift Cycling for winter (November–March) - class 2	−0.3090 (−3.47)	−0.3227 (−3.80)
Shift Cycling for age below 18 or above 64 - class 2	−0.6358 (−4.16)	−1.3231 (−5.18)
Constant Transit - class 1	−1.7203 (−13.24)	−2.0758 (−11.47)
Shift Transit for females - class 1	0.3339 (2.50)	0.3186 (4.58)
Shift Transit for age below 18 - class 1	−0.7387 (−0.81)	0.2925 (1.56)
Shift Transit for age above 64 - class 1	0.9919 (5.27)	0.6414 (5.96)
Constant Transit - class 2	−2.2070 (−26.06)	−1.9745 (−11.05)
Shift Transit for females - class 2	0.2583 (3.68)	0.3881 (2.35)
Shift Transit for age below 18 - class 2	0.7975 (3.13)	0.1651 (0.49)
Shift Transit for age above 64 - class 2	0.4306 (4.18)	−0.0379 (−0.15)
Constant Car - class 1	−4.7424 (−18.60)	−6.3640 (−9.91)
Shift Car for females - class 1	0.4735 (3.19)	0.4468 (2.58)
Shift Car for age below 18 - class 1	−1.5875 (−2.57)	−1.7334 (−4.46)
Shift Car for age above 64 - class 1	1.5389 (6.49)	0.6141 (2.29)
Shift Car for car ownership - class 1	4.3621 (29.95)	1.6926 (13.46)
Constant Car - class 2	−5.8963 (−27.31)	−0.6286 (−2.28)
Shift Car for females - class 2	0.3270 (3.42)	0.0701 (0.51)
Shift Car for age below 18 - class 2	−0.7333 (−3.10)	−1.6307 (−6.19)
Shift Car for age above 64 - class 2	−0.1673 (−1.18)	0.1112 (0.58)
Shift Car for car ownership - class 2	2.1917 (23.02)	1.2616 (16.38)
LOS parameters		
Travel cost (£) - class 1	−0.2757 (−13.36)	−0.5484 (−7.07)
Travel cost (£) - class 2	−0.3607 (−11.30)	−0.1891 (−16.67)
Out-of-vehicle travel time for walking, cycling and transit (hrs) - class 1	−13.3327 (−30.10)	−10.1507 (−7.57)
Out-of-vehicle travel time for walking, cycling and transit (hrs) - class 2	−7.1890 (−37.46)	−7.5145 (−5.02)
In-vehicle travel time for transit and car (hrs) - class 1	−6.0404 (−15.92)	−3.6914 (−5.90)
In-vehicle travel time for transit and car (hrs) - class 2	−2.6818 (−13.46)	−6.6554 (−4.90)
Traffic variability for car - class 1	−3.3284 (−13.21)	−7.5255 (−11.31)
Traffic variability for car - class 2	−5.5894 (−20.00)	−2.8760 (−13.69)
Number of transfers for transit - class 1	−0.4640 (−5.79)	−0.3665 (−2.97)
Number of transfers for transit - class 2	−0.0620 (−1.02)	−0.0267 (−0.24)
Class allocation parameters		
Constant - class 1	−0.4306 (−4.53)	–
Number of cars - class 1	−0.1861 (−3.36)	–
Age - class 1	0.0173 (14.71)	–
Clustering distance parameters		
Distance multiplier γ - class 1	–	−1.2584 (21.65)
Distance multiplier γ - class 2	–	−0.4785 (9.80)
Class/cluster membership probabilities		
Class/cluster 1	0.53	0.52
Class/cluster 2	0.47	0.48

Table 11
Estimated parameters of clustering covariates for the London mode choice model.

Parameters	Cluster 1	Cluster 2
Constant	0.8106 (595.0)	–
Number of cars	−0.1945 (−411.3)	–
Age	−0.0133 (−419.1)	–

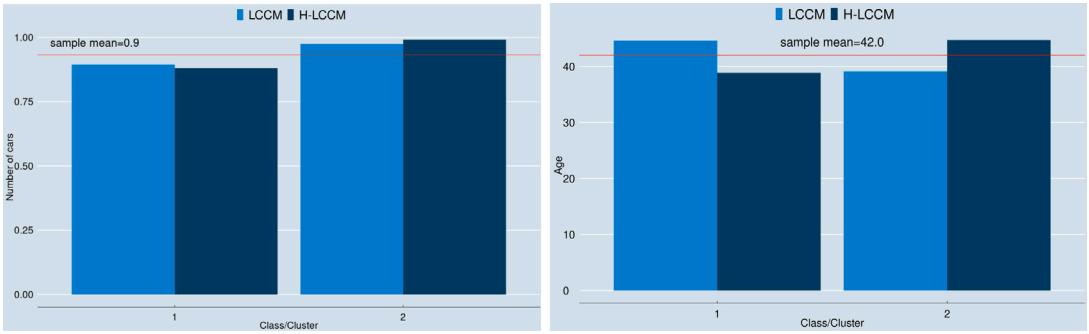
are more advanced clustering algorithms than the K-means inspired approach utilised in the current study. Nonetheless, we focused on probabilistic clustering techniques because we wanted to mimic the traditional econometric LCCM, where the class allocation component, which probabilistically allocates individuals into classes, takes feedback from the choice model and both are jointly estimated. The same methodology developed can also be applied to other similar centroid-based clustering algorithms, such as K-medoids or to more advanced approaches, such as K-harmonic means, and also with different distance measures (e.g. Mahalanobis

Table 12
Class-specific and weighted Willingness-to-pay estimates (£/hr) for the London dataset.

WTP estimate	MNL-base	C-MNL	LCCM	GBM-LCCM	GP-LCCM	H-LCCM
IVTT for Car, Transit - class 1	–	12.94	21.90	6.21	7.52	6.73
IVTT for Car, Transit - class 2	–	21.37	7.44	36.93	31.99	35.19
IVTT for Car, Transit - weighted	17.28	16.93	15.10	20.65	19.02	20.54
OVTT Transit - class 1	–	30.99	48.35	17.81	21.63	18.51
OVTT Transit - class 2	–	38.16	19.93	41.46	37.72	39.74
OVTT Transit - weighted	36.08	34.38	34.98	28.92	29.19	28.80
Car traffic variability - class 1	–	17.42	12.07	13.88	12.44	13.72
Car traffic variability - class 2	–	17.68	15.50	15.18	15.80	15.21
Car traffic variability - weighted	19.00	17.54	13.68	14.49	14.02	14.44
Transit transfers - class 1	–	−0.05	1.68	0.64	1.01	0.67
Transit transfers - class 2	–	0.72	0.17	0.26	0.38	0.14
Transit transfers - weighted	0.28	0.31	0.97	0.46	0.72	0.41

Table 13
Class-specific and weighted mode shares for the London dataset.

Mode	MNL-base	C-MNL	LCCM	GBM-LCCM	GP-LCCM	H-LCCM
Walking - class 1	–	0.20	0.07	0.27	0.26	0.26
Walking - class 2	–	0.13	0.28	0.11	0.12	0.09
Walking - weighted	0.17	0.17	0.17	0.20	0.19	0.18
Cycling - class 1	–	0.04	0.00	0.00	0.00	0.00
Cycling - class 2	–	0.02	0.07	0.12	0.13	0.09
Cycling - weighted	0.03	0.03	0.03	0.06	0.06	0.04
Transit - class 1	–	0.48	0.27	0.66	0.62	0.62
Transit - class 2	–	0.26	0.48	0.17	0.21	0.13
Transit - weighted	0.37	0.38	0.37	0.43	0.43	0.39
Car - class 1	–	0.29	0.66	0.07	0.12	0.12
Car - class 2	–	0.59	0.16	0.60	0.54	0.69
Car - weighted	0.43	0.43	0.43	0.32	0.32	0.39



(a) Number of cars (b) Age

Fig. 6. Class-specific average values of covariates across LCCM and H-LCCM for the London mode choice model.

distance), however requiring significant effort to integrate them into a modelling framework similar to LCCM. The simplicity of the K-Means algorithm made it possible to apply the proposed approach to large sample sizes (case study 3) and to choice problems with large choice sets (case study 2). The same framework and its fundamental principles can also be used to accommodate unconventional data, such as text, in the clustering/class allocation part of the model, as long as those types of data can be decoded in a lower dimensionality representation as numeric vectors, which can be the subject of future research.

Based on all case studies analysed, the proposed methodology is able to achieve at worse comparable results with the traditional econometric specification and with the current state-of-the-art ML-inspired approaches, namely *GBM-LCCM* and *GP-LCCM*, in terms of model fit. Specifically, although the proposed *H-LCCM* shares many similarities with the *GBM-LCCM*, it manages to mimick the properties of *LCCM* while requiring less parameters and it is also able to outperform the *GBM-LCCM* in all of the case studies examined. That in turn leads to less biased estimated parameters, which are more appropriate to be used as guidance for policy making. Among the three case studies, it was evident that more benefits can be achieved with larger samples including more individuals and trips (*Case studies 1, 3*) indicating that a data mining algorithm (soft K-means) might excel at identifying more complex patterns with more available data. Furthermore, it was possible to achieve a more intuitive behavioural interpretation and WTP estimates compared to the traditional econometric model in all of the case studies examined. WTP estimates and especially Values of Travel Time are important from a policy perspective and a more accurate estimation can provide significant benefits for

Table 14
Modelling estimates and t-ratios for MNL-base, C-MNL, LCCM, GBM-LCCM, GP-LCCM and H-LCCM models for the Yorkshire mode choice context..

Parameter	MNL-base	C-MNL	LCCM	GBM-LCCM	GP-LCCM	H-LCCM
Alternative-specific constants						
Constant Car (base)	–	–	–	–	–	–
Constant Bus	−2.2038 (−10.72)	–	–	–	–	–
Constant Bus - class 1	–	−2.9972 (−4.36)	−5.2074 (−11.22)	−5.5633 (−25.31)	−5.2188 (−23.52)	−3.5083 (−4.09)
Constant Bus - class 2	–	−1.5995 (−4.95)	−1.3102 (−3.73)	−3.8408 (−7.66)	−2.6213 (−19.75)	−1.5507 (−3.26)
Constant Bus - class 3	–	−4.0623 (−9.46)	0.3470 (0.33)	−1.3922 (−9.30)	−1.3125 (−12.78)	1.1644 (1.02)
Constant Bus - class 4	–	−1.4891 (−3.82)	−3.7784 (−5.22)	−1.9836 (−25.75)	−1.4529 (−4.28)	−5.0562 (−4.64)
Constant Bus - class 5	–	−2.0604 (−4.73)	−2.4179 (−3.81)	0.4970 (1.67)	−2.9656 (−10.99)	−3.6437 (−5.16)
Constant Rail	−1.6482 (−4.67)	–	–	–	–	–
Constant Rail - class 1	–	−2.5314 (−2.23)	−2.7486 (−2.95)	−4.7580 (−11.25)	−3.6310 (−14.27)	−0.1604 (0.21)
Constant Rail - class 2	–	−1.0855 (−1.97)	−0.4778 (−0.49)	0.4789 (1.25)	−3.8510 (−5.49)	−0.8823 (−1.06)
Constant Rail - class 3	–	−2.3867 (−3.15)	−4.7155 (−3.03)	0.8209 (2.39)	−1.1265 (−2.67)	−3.4326 (−2.25)
Constant Rail - class 4	–	−1.8864 (−1.89)	−1.6476 (−1.35)	−4.2621 (−9.00)	−3.1995 (−1.25)	−3.9105 (−5.95)
Constant Rail - class 5	–	−1.8752 (−2.59)	−8.1766 (−5.02)	−6.8548 (−2.54)	2.1404 (5.70)	−3.7222 (−2.71)
Constant Taxi	−1.9297 (−5.43)	–	–	–	–	–
Constant Taxi - class 1	–	−2.5727 (−2.17)	−4.7217 (−6.57)	−3.8495 (−17.76)	−3.3651 (−17.04)	−3.9144 (−5.59)
Constant Taxi - class 2	–	−2.8653 (−4.53)	−6.5988 (−3.28)	−1.1912 (−1.31)	−4.2626 (−13.92)	−2.1399 (−3.47)
Constant Taxi - class 3	–	−3.1056 (−6.39)	3.4928 (3.06)	−1.9211 (−7.45)	−3.0849 (−3.48)	4.8633 (3.75)
Constant Taxi - class 4	–	0.0466 (0.06)	−0.6965 (−1.18)	−4.2304 (−11.65)	−1.2447 (−4.78)	−2.8189 (−3.58)
Constant Taxi - class 5	–	−2.1538 (−3.44)	−3.9916 (−4.02)	4.2082 (9.32)	−3.5682 (−7.67)	−4.6772 (−6.55)
Constant Cycling	−3.4370 (−8.84)	–	–	–	–	–
Constant Cycling - class 1	–	−2.2870 (−2.63)	39.8735 (3.84)	−9.0325 (−17.96)	−8.3003 (−18.27)	−5.0748 (−5.10)
Constant Cycling - class 2	–	−4.8653 (−8.22)	−1.6347 (−1.85)	−2.0919 (−8.92)	−5.3019 (−6.73)	−4.2522 (−3.45)
Constant Cycling - class 3	–	−3.6812 (−4.19)	−2.0753 (−1.74)	−0.4867 (−2.38)	−16.3865 (−0.01)	−3.6530 (−2.96)
Constant Cycling - class 4	–	−4.6335 (−4.78)	−2.5533 (−4.09)	−5.2292 (−5.88)	−2.9465 (−9.49)	−2.0757 (−2.92)
Constant Cycling - class 5	–	−3.2443 (−2.40)	−1.5036 (−1.04)	1.5004 (−0.01)	−0.5059 (−2.69)	−2.9096 (−1.81)
Constant Walking	1.3025 (5.30)	–	–	–	–	–
Constant Walking - class 1	–	0.3273 (0.45)	0.2920 (0.46)	0.3779 (1.20)	0.6149 (2.22)	2.4618 (2.16)
Constant Walking - class 2	–	1.8671 (4.22)	1.4847 (2.61)	2.2356 (4.36)	−0.4179 (−1.45)	0.8567 (1.30)
Constant Walking - class 3	–	−0.0134 (−0.03)	4.4302 (4.81)	1.5118 (4.62)	2.0694 (7.45)	4.1692 (4.72)
Constant Walking - class 4	–	2.5581 (5.42)	2.1666 (2.80)	0.6317 (3.08)	0.9783 (2.48)	0.8978 (1.02)
Constant Walking - class 5	–	1.1943 (2.09)	−0.4120 (−0.54)	4.4258 (10.39)	2.8548 (5.09)	0.2404 (0.41)
LOS parameters						
Car travel time (mins)	−0.1287 (−9.40)	–	–	–	–	–
Car travel time (mins) - class 1	–	−0.0883 (−1.05)	−0.1712 (−8.78)	−0.1610 (−21.84)	−0.1503 (−30.05)	−0.0809 (−3.12)
Car travel time (mins) - class 2	–	−0.1556 (−7.59)	−0.0794 (−1.19)	−0.1482 (−20.04)	−0.3894 (−44.01)	−0.2629 (−7.09)
Car travel time (mins) - class 3	–	−0.1420 (−4.09)	−0.1312 (−3.08)	−0.0946 (−18.71)	−0.0349 (−5.17)	−0.1692 (−2.04)
Car travel time (mins) - class 4	–	−0.1162 (−4.59)	−0.1040 (−3.23)	−0.0826 (−14.98)	−0.3184 (−46.71)	−0.1383 (−3.93)
Car travel time (mins) - class 5	–	−0.1190 (−4.59)	−0.2688 (−4.74)	−0.1509 (−14.49)	−0.0492 (−7.98)	−0.0259 (−0.78)
Bus travel time (mins)	−0.0663 (−8.12)	–	–	–	–	–
Bus travel time (mins) - class 1	–	−0.0284 (−0.72)	−0.0795 (−8.48)	−0.0748 (−11.08)	−0.0728 (−10.60)	−0.0626 (−4.81)
Bus travel time (mins) - class 2	–	−0.0720 (−9.26)	−0.0323 (−1.16)	−0.0953 (−7.22)	−0.1454 (−38.19)	−0.1047 (−7.54)
Bus travel time (mins) - class 3	–	−0.0535 (−2.37)	−0.1113 (−2.94)	−0.0552 (−17.87)	−0.0141 (−6.90)	−0.1556 (−5.13)
Bus travel time (mins) - class 4	–	−0.0804 (−4.81)	−0.0645 (−4.30)	−0.0235 (−16.77)	−0.1796 (−14.96)	−0.0527 (−3.87)
Bus travel time (mins) - class 5	–	−0.0804 (−6.56)	−0.0937 (−4.93)	−0.1037 (−11.52)	−0.0305 (−7.23)	−0.0088 (−2.87)
Rail travel time (mins)	−0.0743 (−8.56)	–	–	–	–	–
Rail travel time (mins) - class 1	–	−0.0411 (−1.11)	−0.1017 (−8.28)	−0.0787 (−8.41)	−0.0675 (−14.23)	−0.0575 (−4.44)
Rail travel time (mins) - class 2	–	−0.0883 (−7.56)	−0.0727 (−2.78)	−0.0807 (−11.28)	−0.1876 (−12.60)	−0.1583 (−8.65)
Rail travel time (mins) - class 3	–	−0.0828 (−4.06)	−0.0230 (−0.71)	−0.0785 (−13.45)	−0.0489 (−5.22)	−0.0287 (−1.44)
Rail travel time (mins) - class 4	–	−0.0723 (−3.61)	−0.0380 (−2.20)	−0.0359 (−4.34)	−0.2322 (−4.25)	−0.0641 (−1.81)
Rail travel time (mins) - class 5	–	−0.0658 (−3.83)	−0.0619 (−2.15)	−0.0074 (−0.21)	−0.0620 (−10.52)	−0.1505 (−4.52)
Taxi travel time (mins)	−0.1747 (−6.35)	–	–	–	–	–
Taxi travel time (mins) - class 1	–	−0.1973 (−2.43)	−0.2324 (−5.56)	−0.2333 (−18.57)	−0.2262 (−18.46)	−0.0748 (−1.55)
Taxi travel time (mins) - class 2	–	−0.1405 (−3.64)	0.0058 (−0.06)	−0.2849 (−3.82)	−0.3808 (−19.37)	−0.2888 (−6.03)
Taxi travel time (mins) - class 3	–	−0.1648 (−3.35)	−0.3930 (−5.04)	−0.0436 (−4.54)	−0.1162 (−2.16)	−0.4942 (−5.81)
Taxi travel time (mins) - class 4	–	−0.2402 (−4.88)	−0.1285 (−2.20)	−0.0903 (−4.16)	−0.3039 (−26.29)	−0.1200 (−3.34)
Taxi travel time (mins) - class 5	–	−0.2627 (−4.89)	−0.1891 (−2.96)	−0.4490 (−11.09)	0.0066 (0.44)	−0.1022 (−1.77)
Cycling travel time (mins)	−0.0846 (−6.58)	–	–	–	–	–
Cycling travel time (mins) - class 1	–	−0.0676 (−2.21)	−8.8323 (−4.29)	−0.0363 (−2.65)	−0.0477 (−3.73)	−0.0700 (−2.06)
Cycling travel time (mins) - class 2	–	−0.0596 (−4.51)	−0.3935 (−4.87)	−0.1526 (−12.14)	−0.2455 (−5.65)	−0.2064 (−2.73)
Cycling travel time (mins) - class 3	–	−0.0829 (−3.36)	−0.1425 (−3.92)	−0.0695 (−8.16)	−1.0650 (−0.01)	−0.0796 (−2.54)
Cycling travel time (mins) - class 4	–	−0.0782 (−1.78)	−0.0621 (−4.14)	−0.0802 (−1.60)	−0.1416 (−9.27)	−0.0688 (−5.25)
Cycling travel time (mins) - class 5	–	−0.1071 (−2.47)	−0.1287 (−3.05)	−1.5004 (−0.01)	−0.0619 (−7.40)	−1.9995 (−6.35)
Walking travel time (mins)	−0.1519 (−15.05)	–	–	–	–	–
Walking travel time (mins) - class 1	–	−0.0804 (−2.82)	−0.2024 (−10.31)	−0.2380 (−17.37)	−0.2080 (−18.33)	−0.2226 (−7.56)
Walking travel time (mins) - class 2	–	−0.1895 (−11.35)	−0.1535 (−5.48)	−0.1748 (−7.72)	−0.2393 (−13.22)	−0.2126 (−8.78)
Walking travel time (mins) - class 3	–	−0.1375 (−7.57)	−0.1178 (−4.13)	−0.1510 (−10.57)	−0.1350 (−11.36)	−0.1404 (−5.40)
Walking travel time (mins) - class 4	–	−0.1726 (−9.29)	−0.2245 (−8.75)	−0.1437 (−13.98)	−0.1052 (−6.84)	−0.1948 (−6.58)
Walking travel time (mins) - class 5	–	−0.1716 (−8.75)	−0.1780 (−5.38)	−0.1266 (−7.66)	−0.2050 (−8.41)	−0.1394 (−6.82)
Natural logarithm of travel cost (£)	−0.9421 (−11.89)	–	–	–	–	–
Natural logarithm of travel cost (£) class 1	–	−0.3561 (−1.03)	−0.3611 (−1.75)	−0.5531 (−5.56)	−0.6218 (−7.60)	−0.9863 (−3.70)
Natural logarithm of travel cost (£) class 2	–	−0.5825 (−4.43)	−0.5080 (−3.00)	−1.3137 (−11.36)	−0.0699 (−0.76)	−0.4647 (−2.36)
Natural logarithm of travel cost (£)	–	−0.9448 (−4.77)	−1.2605 (−3.87)	−1.3256 (−16.75)	−0.9430 (−12.28)	−1.7901 (−5.83)

(*continued on next page*)

Table 14 (*continued*).

Parameter	MNL-base	C-MNL	LCCM	GBM-LCCM	GP-LCCM	H-LCCM
class 3						
Natural logarithm of travel cost (£)	–	−1.1721 (−6.79)	−1.7146 (−10.47)	−0.3593 (−5.40)	−0.6577 (−5.82)	−0.7184 (−2.61)
class 4						
Natural logarithm of travel cost (£)	–	−0.7953 (−4.44)	−0.2245 (−0.81)	−1.4808 (−10.56)	−1.3103 (−12.27)	−0.9709 (−4.51)
class 5						
Class allocation parameters						
Constant - class 1	–	–	1.4520 (1.42)	0.3839	–	–
Season ticket ownership - class 1	–	–	−1.2008 (−2.33)	0.1284	–	–
Number of cars in household - class 1	–	–	2.2219 (4.19)	0.3489	–	–
Age - class 1	–	–	−0.0078 (−0.42)	0.0978	–	–
Female - class 1	–	–	−0.7680 (−1.39)	0.6667	–	–
Annual household income (£1,000)	–	–	−0.0068 (−0.69)	0.1889	–	–
class 1						
Constant - class 2	–	–	1.3730 (1.28)	0.1883	–	–
Season ticket ownership - class 2	–	–	0.7042 (1.25)	0.2984	–	–
Number of cars in household - class 2	–	–	1.8080 (3.15)	0.0494	–	–
Age - class 2	–	–	−0.0054 (−0.28)	−0.0657	–	–
Female - class 2	–	–	−0.9572 (−1.60)	0.4238	–	–
Annual household income (£1,000)	–	–	−0.0287 (−2.44)	−0.1123	–	–
class 2						
Constant - class 3	–	–	4.2390 (3.68)	0.1142	–	–
Season ticket ownership - class 3	–	–	−2.2309 (−2.38)	0.3460	–	–
Number of cars in household - class 3	–	–	1.7204 (2.69)	−0.1372	–	–
Age - class 3	–	–	−0.1023 (−3.40)	0.0206	–	–
Female - class 3	–	–	−0.6773 (−0.95)	0.5138	–	–
Annual household income (£1,000)	–	–	−0.0143 (−1.29)	0.4473	–	–
class 3						
Constant - class 4	–	–	3.7335 (3.45)	0.2144	–	–
Season ticket ownership - class 4	–	–	−1.0282 (−1.62)	0.5013	–	–
Number of cars in household - class 4	–	–	2.2784 (4.34)	−0.4100	–	–
Age - class 4	–	–	−0.0640 (−3.00)	0.2394	–	–
Female - class 4	–	–	−1.3283 (−2.15)	0.5155	–	–
Annual household income (£1,000)	–	–	−0.0180 (−1.58)	−0.3480	–	–
class 4						
Constant - class 5	–	–	–	0.0992	–	–
Season ticket ownership - class 5	–	–	–	0.1430	–	–
Number of cars in household - class 5	–	–	–	−0.3997	–	–
Age - class 5	–	–	–	−0.7950	–	–
Female - class 5	–	–	–	0.6967	–	–
Annual household income (£1,000)	–	–	–	−0.2806	–	–
class 5						
Clustering distance parameters						
Distance multiplier γ - cluster 1	–	–	–	–	–	−0.9110 (3.35)
Distance multiplier γ - cluster 2	–	–	–	–	–	−0.7202 (3.60)
Distance multiplier γ - cluster 3	–	–	–	–	–	−0.3428 (1.50)
Distance multiplier γ - cluster 4	–	–	–	–	–	−0.1812 (1.08)
Distance multiplier γ - cluster 5	–	–	–	–	–	−1.0880 (5.05)
Class/cluster membership probabilities						
Class/cluster 1	–	0.14	0.42	0.38	0.56	0.20
Class/cluster 2	–	0.24	0.18	0.19	0.13	0.21
Class/cluster 3	–	0.22	0.10	0.11	0.13	0.17
Class/cluster 4	–	0.24	0.21	0.21	0.10	0.19
Class/cluster 5	–	0.16	0.09	0.10	0.08	0.23

policy makers. Finally, the models based on the proposed approach were able to offer a more intuitive behavioural profiling for the estimated clusters (e.g. case study 3), which could further lead to policy measures targeting more accurately the underlying population and their needs and constraints.

The proposed methodology is of course subject to certain limitations, the most important being the centroid initialisation process. The estimation is highly dependent on the initial centroid that is randomly selected, and under the K-means++ initialisation process, that initial centroid forms the basis for the selection of the remaining centroids, as well. Prior assumptions regarding the signs of the scaled clustering variables can help to reach a better final LL, but it is difficult to have any meaningful a priori sign directionality assumptions in the presence of a large number of classes/clusters, such as in *Case study 1*. That limitation was addressed by performing multiple estimation runs from different starting points (initial centroids) in order to add confidence to our results. The models presented in the current study are the ones resulting in the most behaviourally intuitive estimates and not solely on the ones with the better model fit. Furthermore, in order to present a fair comparison among the models, the covariates used in the class allocation components are those that resulted in the best performing traditional LCCM in each case study. Therefore, the full range of benefits to be gained from the proposed approach was not examined, i.e. in terms of including more sociodemographic attributes in the clustering stage (class allocation). Finally, it may be noted that K-means assumes well-defined spherical and distinct clusters while more advanced ML techniques based on DNN architectures (that can also handle probabilistic clustering), such as Variational Auto-encoders and Self-Organising Maps, can provide more flexibility and offer an interesting avenue for future research.

The current study aims to build on the increasing literature focusing on the integration of ML and DCM. As illustrated in the case studies presented, there are additional benefits to be achieved by incorporating an ML algorithm into a DCM framework. That approach is able to take the best of both worlds by using a probabilistic data mining approach for identifying patterns in the data more effectively, while also allowing the choice process to be modelled by a DCM, thus providing valuation measures, which are

Table 15
Modelling estimates and t-ratios for MNL-base, C-MNL, LCCM GBM-LCCM, GP-LCCM and H-LCCM models for the Yorkshire shopping destination choice context..

Parameter	MNL-base	C-MNL	LCCM	GBM-LCCM	GP-LCCM	H-LCCM
Alternative-specific constants						
textitConstant Leeds city centre (base)			–			–
Constant Remaining Leeds	−0.4088 (−2.60)	–	–	–	–	–
Constant Remaining Leeds - class 1	–	−0.4293 (−1.59)	−0.5628 (−0.55)	−1.0208 (−6.07)	−0.0480 (−0.22)	−0.8985 (−0.87)
Constant Remaining Leeds - class 2	–	−0.4141 (−2.13)	−0.5195 (−1.34)	−0.3783 (−3.62)	−0.6447 (−6.63)	−0.5190 (−1.59)
Constant Remaining Leeds shift for season ticket owners/no car ownership	−0.9086 (−4.02)	–	–	–	–	–
Constant Remaining Leeds shift for season ticket owners/no car ownership - class 1	–	–	–	–	–	–
Constant Remaining Leeds shift for season ticket owners/no car ownership - class 2	–	−0.8214 (−2.27)	−1.9416 (−2.74)	−1.3989 (−9.29)	−2.2954 (−13.78)	−1.6475 (−1.15)
Constant Remaining Yorkshire	0.3228 (1.66)	–	–	–	–	–
Constant Remaining Yorkshire - class 1	–	0.2560 (0.72)	0.3734 (0.31)	−0.4146 (−1.38)	0.7643 (2.41)	−0.1418 (−0.12)
Constant Remaining Yorkshire - class 2	–	0.3273 (1.38)	0.1881 (0.52)	0.4258 (3.06)	0.1019 (0.74)	0.2503 (0.75)
Constant Remaining Yorkshire shift for season ticket owners/no car ownership	−0.9322 (−2.42)	–	–	–	–	–
Constant Remaining Yorkshire shift for season ticket owners/no car ownership - class 1	–	−0.4602 (−0.65)	−0.4916 (−0.31)	−2.5158 (−4.19)	−1.3764 (−3.19)	−0.8747 (−0.74)
Constant Remaining Yorkshire shift for season ticket owners/no car ownership - class 2	–	−1.2005 (−2.64)	−0.7664 (−0.91)	−0.4528 (−2.54)	−0.8167 (−4.32)	−0.6232 (−1.23)
LOS parameters						
Travel time car previous trip (mins)	−0.1217 (−9.08)	–	–	–	–	–
Travel time car previous trip (mins) - class 1	–	−0.1330 (−4.58)	−0.0523 (−0.76)	−0.0363 (−2.53)	−0.0975 (−6.91)	−0.0114 (−0.10)
Travel time car previous trip (mins) - class 2	–	−0.1218 (−8.26)	−0.1379 (−3.96)	−0.1382 (−19.93)	−0.1311 (−18.51)	−0.1319 (−5.48)
Travel time car next trip (mins)	−0.1456 (−10.66)	–	–	–	–	–
Travel time car next trip (mins) - class 1	–	−0.1391 (−4.62)	−0.2350 (−2.29)	−0.1031 (−7.50)		−0.1178 (−1.13)
Travel time car next trip (mins) - class 2	–	−0.1526 (−9.75)	−0.1217 (−3.77)	−0.1563 (−22.21)	−0.1241 (−18.23)	−0.1351 (−3.66)
Travel time PT previous trip (mins)	−0.0356 (−1.75)	–	–	–	–	–
Travel time PT previous trip (mins) - class 1	–	−0.0176 (−0.76)	−0.0852 (−2.14)	0.0674 (8.41)		−0.0289 (−0.61)
Travel time PT previous trip (mins) - class 2	–	−0.0495 (−1.46)	−0.0201 (−0.74)	−0.0744 (−11.21)	−0.0205 (−3.66)	−0.0283 (−0.99)
Travel time PT next trip (mins)	−0.0834 (−4.13)	–	–	–	–	–
Travel time PT next trip (mins) - class 1	–	−0.0849 (−3.78)	−0.2748 (−2.15)	−0.1065 (−10.58)		−0.1904 (−2.90)
Travel time PT next trip (mins) - class 2	–	−0.0893 (−2.70)	−0.0631 (−1.67)	−0.0983 (−14.81)	−0.0678 (−11.52)	−0.0708 (−2.69)
Walking distance previous trip (km)	−1.5821 (−12.47)	–	–	–	–	–
Walking distance previous trip (km) - class 1	–	−1.5599 (−8.95)	−2.8828 (−1.17)	−1.9456 (−20.15)		−2.4770 (−3.31)
Walking distance previous trip (km) - class 2	–	−1.6033 (−8.39)	−1.1975 (−3.32)	−1.3952 (−23.52)	−1.1761 (−22.72)	−1.2556 (−4.36)
Walking distance next trip (km)	−1.7995 (−12.12)	–	–	–	–	–
Walking distance next trip (km) - class 1	–	−1.7365 (−8.76)	−1.8370 (−0.51)	−1.3156 (−15.79)		−1.3240 (−1.59)
Walking distance next trip (km) - class 2	–	−1.9083 (−7.90)	−1.8779 (−1.15)	−2.3969 (−21.71)	−1.7775 (−23.03)	−2.1864 (−3.57)
Travel cost linear (£)	−0.1921 (−2.26)	–	–	–	–	–
Travel cost linear (£) - class 1	–	−0.6076 (−2.05)	−2.4312 (−2.56)	−4.2631 (−32.17)		−3.7554 (−4.20)
Travel cost linear (£) - class 2	–	−0.1417 (−1.64)	−0.1222 (0.49)	−0.0469 (−1.17)	−0.1271 (−3.01)	−0.0847 (−1.10)
Travel cost log (£)	−0.2026 (−3.13)	–	–	–	–	–
Travel cost log (£) - class 1	–	−0.1760 (−1.24)	0.2629 (0.52)	1.3494 (8.72)		0.2961 (0.69)
Travel cost log (£) - class 2	–	−0.1412 (−1.91)	−0.1244 (−1.67)	−0.1008 (−1.00)	−0.2420 (−2.48)	−0.1078 (−0.76)
Direction of travel						
Presence of angle> 90^o between O-S and O-D	−0.3823 (−3.39)	–	–	–	–	–
Presence of angle> 90^o between O-S and O-D - class 1	–	−0.3069 (−1.54)	−0.0242 (−0.04)	−0.1161 (−0.68)		−0.0624 (−0.26)
Presence of angle> 90^o between O-S and O-D - class 2	–	−0.4096 (−3.06)	−0.4832 (−2.98)	−0.3845 (−3.18)	−0.4628 (−4.04)	−0.4562 (−2.80)
Locational variables						
Parking areas linear (400 m buffer)	0.0161 (9.59)	–	–	–	–	–
Parking areas linear (400 m buffer) - class 1	–	0.0167 (5.72)	0.0052 (0.58)	0.0080 (2.39)		0.0087 (2.14)
Parking areas linear (400 m buffer) - class 2	–	0.0160 (7.77)	0.0198 (4.93)	0.0200 (10.62)	0.0207 (11.11)	0.0195 (8.16)
Parking areas log (400 m buffer)	0.0462 (3.34)	–	–	–	–	–
Parking areas log (400 m buffer) - class 1	–	0.0120 (0.71)	0.0663 (2.03)	0.0369 (2.78)		0.0433 (0.68)
Parking areas log (400 m buffer) - class 2	–	0.0718 (4.64)	0.0395 (2.21)	0.0561 (4.26)	0.0550 (4.28)	0.0521 (1.17)
Size variables						
Natural logarithm multiplier ϕ	0.4829 (12.43)	–	–	–	–	–
Natural logarithm multiplier ϕ - class 1	–	0.4004 (6.61)	0.2950 (0.71)	0.3890 (11.22)		0.3295 (1.21)
Natural logarithm multiplier ϕ - class 2	–	0.5382 (10.74)	0.5823 (2.92)	0.5353 (19.84)	0.4858 (19.72)	0.5640 (3.62)
Class allocation parameters						
Constant - class 1	–	–	−0.6320 (−0.86)	–	–	–
Annual personal income (£1,000) - class 1	–	–	−0.0150 (−1.18)	–	–	–
Home IMD - class 1	–	–	0.0211 (0.99)	–	–	–
Clustering distance parameters						
Distance multiplier γ - cluster 1	–	–	–	–	–	−0.1507 (0.27)
Distance multiplier γ - cluster 2	–	–	–	–	–	−0.7640 (2.06)
Class/cluster membership probabilities						
Class/cluster 1	–	0.41	0.37	0.39	0.34	0.42
Class/cluster 2	–	0.59	0.63	0.61	0.66	0.58

Table 16
Modelling estimates and t-ratios for MNL-base, C-MNL, LCCM, GBM-LCCM, GP-LCCM and H-LCCM for the London mode choice context.

Parameter	MNL-base	C-MNL	LCCM	GBM-LCCM	GP-LCCM	H-LCCM
Alternative-specific constants						
Constant Walking (base)	–	–	–	–	–	–
Constant Cycling	−3.1143 (−52.04)	–	–	–	–	–
Constant Cycling - class 1	–	−3.0798 (−41.19)	−7.6354 (−7.05)	−7.6834 (−26.43)	−7.3312 (−29.96)	−7.7086 (−9.42)
Constant Cycling - class 2	–	−3.1119 (−31.05)	−3.2261 (−32.62)	−1.6516 (−38.83)	−1.5820 (−36.47)	−1.6507 (−11.51)
Shift Cycling for females	−1.0892 (−13.26)	–	–	–	–	–
Shift Cycling for females - class 1	–	−1.1500 (−11.18)	−0.8494 (−0.75)	−18.2637 (−0.01)	−21.4619 (−0.01)	−7.1144 (−11.08)
Shift Cycling for females - class 2	–	−0.9907 (7.29)	−1.1075 (−11.09)	−1.0266 (−18.24)	−0.9438 (−16.41)	−1.0203 (−6.50)
Shift Cycling for winter (November–March)	−0.2767 (−3.66)	–	–	–	–	–
Shift Cycling for winter (November–March) - class 1	–	−0.2372 (−2.57)	1.1809 (0.91)	−1.5562 (−1.44)	−0.3659 (−1.48)	−1.5958 (−1.06)
Shift Cycling for winter (November–March) - class 2	–	−0.3377 (−2.56)	−0.3090 (−3.47)	−0.3258 (−5.76)	−0.3430 (−5.83)	−0.3227 (−3.80)
Shift Cycling for age below 18 or above 64	−0.8067 (−6.92)	–	–	–	–	–
Shift Cycling for age below 18 or above 64 - class 1	–	−0.8982 (−5.37)	−9.5609 (−12.94)	−1.0760 (−1.05)	2.5345 (9.36)	−0.1846 (−0.14)
Shift Cycling for age below 18 or above 64 - class 2	–	−0.7876 (−4.80)	−0.6358 (−4.16)	−1.2155 (−15.22)	−1.7820 (−16.59)	−1.3231 (−5.18)
Constant Transit	−1.5191 (−37.56)	–	–	–	–	–
Constant Transit - class 1	–	−1.4128 (−26.64)	−1.7203 (−13.24)	−2.0240 (−42.45)	−2.0316 (−47.61)	−2.0758 (−11.47)
Constant Transit - class 2	–	−1.6906 (−26.08)	−2.2070 (−26.06)	−2.0390 (−31.32)	−1.9905 (−30.98)	−1.9745 (−11.05)
Shift Transit for females	0.1898 (4.94)	–	–	–	–	–
Shift Transit for females - class 1	–	0.1378 (2.82)	0.3339 (2.50)	0.3138 (8.00)	0.3027 (8.38)	0.3186 (4.58)
Shift Transit for females - class 2	–	0.2862 (4.64)	0.2583 (3.68)	0.3900 (6.00)	0.2552 (4.16)	0.3881 (2.35)
Shift Transit for age below 18	0.3379 (5.66)	–	–	–	–	–
Shift Transit for age below 18 - class 1	–	0.2267 (3.50)	−0.7387 (−0.81)	0.1767 (2.42)	0.3069 (5.29)	0.2925 (1.56)
Shift Transit for age below 18 - class 2	–	–	0.7975 (3.13)	0.2516 (2.76)	0.8854 (12.52)	0.1651 (0.49)
Shift Transit for age above 64	0.5704 (10.43)	–	–	–	–	–
Shift Transit for age above 64 - class 1	–	–	0.9919 (5.27)	0.6417 (10.14)	0.6653 (11.70)	0.6414 (5.96)
Shift Transit for age above 64 - class 2	–	0.7249 (11.26)	0.4306 (4.18)	0.0688 (0.75)	−0.0985 (−0.87)	−0.0379 (−0.15)
Constant Car	−2.6890 (43.83)	–	–	–	–	–
Constant Car - class 1	–	−2.9646 (−35.11)	−4.7424 (−18.60)	−5.8323 (−53.67)	−6.3941 (−68.94)	−6.3640 (−9.91)
Constant Car - class 2	–	−2.0763 (−22.32)	−5.8963 (−27.31)	−0.5142 (−9.26)	−0.5488 (−9.39)	−0.6286 (−2.28)
Shift Car for females	0.1040 (2.37)	–	–	–	–	–
Shift Car for females - class 1	–	0.1047 (1.72)	0.4735 (3.19)	0.4173 (5.60)	0.4930 (8.64)	0.4468 (2.58)
Shift Car for females - class 2	–	0.1281 (2.04)	0.3270 (3.42)	0.0643 (1.64)	0.0643 (1.61)	0.0701 (0.51)
Shift Car for age below 18	−1.0723 (−14.17)	–	–	–	–	–
Shift Car for age below 18 - class 1	–	−1.1846 (−13.14)	−1.5875 (−2.57)	−1.4609 (−8.68)	2.4311 (38.29)	−1.7334 (−4.46)
Shift Car for age below 18 - class 2	–	–	−0.7333 (−3.10)	−1.4118 (−22.51)	−3.4111 (−51.40)	−1.6307 (−6.19)
Shift Car for age above 64	0.5649 (9.10)	–	–	–	–	–
Shift Car for age above 64 - class 1	–	–	1.5389 (6.49)	0.1827 (1.36)	0.9181 (9.58)	0.6141 (2.29)
Shift Car for age above 64 - class 2	–	0.2998 (4.42)	−0.1673 (−1.18)	0.0229 (0.38)	0.3991 (5.27)	0.1112 (0.58)
Shift Car for car ownership	1.5191 (67.12)	–	–	–	–	–
Shift Car for car ownership - class 1	–	1.8517 (43.34)	4.3621 (29.95)	1.4156 (26.90)	1.5012 (32.67)	1.6926 (13.46)
Shift Car for car ownership - class 2	–	1.0510 (32.64)	2.1917 (23.02)	1.1139 (37.19)	1.2873 (37.60)	1.2616 (16.38)
LOS parameters						
Travel cost (£)	−0.1863 (−22.77)	–	–	–	–	–
Travel cost (£) - class 1	–	−0.2088 (−14.66)	−0.2757 (−13.36)	−0.5532 (−21.63)	−0.4279 (−20.53)	−0.5484 (−7.07)
Travel cost (£) - class 2	–	−0.1864 (−16.72)	−0.3607 (−11.30)	−0.1877 (−25.55)	−0.1968 (−26.57)	−0.1891 (−16.67)
Out-of-vehicle travel time for walking, cycling and transit (hrs)	−6.7220 (−63.76)	–	–	–	–	–
Out-of-vehicle travel time for walking, cycling and transit (hrs) - class 1	–	−6.4701 (−47.55)	−13.3327 (−30.10)	−9.8516 (−231.26)	−9.2565 (−233.46)	−10.1507 (−7.57)
Out-of-vehicle travel time for walking, cycling and transit (hrs) - class 2	–	−7.1115 (−42.15)	−7.1890 (−37.46)	−7.7818 (−135.12)	−7.4236 (−126.48)	−7.5145 (−5.02)
In-vehicle travel time for transit and car (hrs)	−3.2195 (−30.48)	–	–	–	–	–
In-vehicle travel time for transit and car (hrs) - class 1	–	−2.7022 (−19.21)	−6.0404 (−15.92)	−3.4366 (−24.87)	−3.2187 (−28.43)	−3.6914 (−5.90)
In-vehicle travel time for transit and car (hrs) - class 2	–	−3.9825 (−22.98)	−2.6818 (−13.46)	−6.9322 (−74.97)	−6.2960 (−69.19)	−6.6554 (−4.90)
Traffic variability for car	−3.5401 (−37.41)	–	–	–	–	–
Traffic variability for car - class 1	–	−3.6361 (−26.16)	−3.3284 (−13.21)	−7.6756 (−26.24)	−5.3219 (−26.64)	−7.5255 (−11.31)
Traffic variability for car - class 2	–	−3.2943 (−25.09)	−5.5894 (−20.00)	−2.8498 (−23.89)	−3.1088 (−25.40)	−2.8760 (−13.69)
Number of transfers for transit	−0.0523 (−2.05)	–	–	–	–	–
Number of transfers for transit - class 1	–	0.0111 (0.31)	−0.4640 (−5.79)	−0.3524 (−5.79)	−0.4309 (−9.02)	−0.3665 (−2.97)
Number of transfers for transit - class 2	–	−0.1348 (−3.44)	−0.0620 (−1.02)	−0.0481 (−1.00)	0.0747 (1.59)	−0.0267 (−0.24)

(*continued on next page*)

Table 16 (*continued*).

Parameter	MNL-base	C-MNL	LCCM	GBM-LCCM	GP-LCCM	H-LCCM
Class allocation parameters						
Constant - class 1	–	–	−0.4306 (−4.53)	0.5283	–	–
Number of cars - class 1	–	–	−0.1861 (−3.36)	−0.3909	–	–
Age - class 1	–	–	0.0173 (14.71)	−0.1190	–	–
Constant - class 2	–	–	–	0.4717	–	–
Number of cars - class 2	–	–	–	0.4379	–	–
Age - class 2	–	–	–	0.1333	–	–
Clustering distance parameters						
Distance multiplier γ - class 1	–	–	–	–	–	−1.2584 (21.65)
Distance multiplier γ - class 2	–	–	–	–	–	−0.4785 (9.80)
Class/cluster membership probabilities						
Class/cluster 1	–	0.53	0.53	0.53	0.53	0.52
Class/cluster 2	–	0.47	0.47	0.47	0.47	0.48

Table 17
Estimated covariances for the continuous covariates of GBM-LCCM for the Yorkshire mode choice study. .

Covariate	*Number of cars*	*Age*	*Annual household income*
Class 1			
Number of cars	1.1506		
Age	0.4973	1.0154	
Annual household income	0.3557	0.1293	1.2077
Class 2			
Number of cars	0.7448		
Age	0.2560	0.8357	
Annual household income	0.3046	0.1337	0.4924
Class 3			
Number of cars	0.6492		
Age	0.2124	0.9152	
Annual household income	0.3425	−0.1187	2.0533
Class 4			
Number of cars	0.8410		
Age	0.2167	1.1076	
Annual household income	0.1590	−0.1822	0.4044
Class 5			
Number of cars	0.6288		
Age	0.1457	0.3159	
Annual household income	0.2206	0.1457	0.5017

Table 18
Estimated covariances for the continuous covariates of GBM-LCCM for the Yorkshire destination choice study. .

Covariate	*Annual household income*	*Home IMD*
Class 1		
Annual household income	0.5727	
Home IMD	0.0271	1.3313
Class 2		
Annual household income	1.1929	
Home IMD	−0.2477	0.5139

Table 19
Estimated covariances for the continuous covariates of GBM-LCCM for the London mode choice study.

Covariate	*Number of cars*	*Age*
Class 1		
Number of cars	0.9128	
Age	0.0199	1.0107
Class 2		
Number of cars	0.7347	
Age	0.0490	0.9544

Table 20
Train/Test validation for the C-MNL based models and the H-LCCM for the Yorkshire mode choice study. .

Model type	*LL (full sample)*	*LL (80% train set)*	*LL (20% test set)*
C-MNL (K-means Euclidean)	−4,928.32	−4,011.62	−910.50
C-MNL (K-means Mahalanobis)	−4,999.54	−4,119.24	−995.08
C-MNL (K-Harmonic means)	−4,958.71	−4,248.38	−1,137.60
C-MNL (DBSCAN)	−4,895.39	−4,068.07	−847.66
H-LCCM (with Euclidean distance)	−3,940.00	−3,235.24	−742.39

Table 21
Train/Test validation for the C-MNL based models and the H-LCCM for the Yorkshire destination choice study. .

Model type[a]	*LL (full sample)*	*LL (80% train set)*	*LL (20% test set)*
C-MNL (K-means Euclidean)	−3,640.98	−2,922.72	−726,20
C-MNL (K-means Mahalanobis)	−3,640.74	−2,921.59	−726.34
C-MNL (K-Harmonic means)	−3,641.88	−2,922.72	−726.03
H-LCCM (with Euclidean distance)	−3,573.32	−2,880.22	−711.13

[a] DBSCAN was not evaluated in that study because one of the clusters had only 10 trips and it was decided not to estimate a model on such a small sample.

Table 22
Train/Test validation for the C-MNL based models and the H-LCCM for the London mode choice study. .

Model type	*LL (full sample)*	*LL (80% train set)*	*LL (20% test set)*
C-MNL (K-means Euclidean)	−43,854.73	−34,109.46	−8,956.70
C-MNL (K-means Mahalanobis)	−43,053.58	−34,109.45	−8,956.69
C-MNL (K-Harmonic means)	−43,865.30	−33,971.49	−8,945.23
C-MNL (DBSCAN)	−42,948.96	−34,013.65	−8,955.67
H-LCCM (with Euclidean distance)	−37,137.11	−29,614.85	−7,327.07

important for policy making. The proposed approach thus aims to make ML relevant for policy in transport by highlighting the benefits to be gained from its use. More studies are expected to take these approaches even further as the ML-DCM literature keeps developing.

CRediT authorship contribution statement

Panagiotis Tsoleridis: Conceptualization, Formal analysis, Writing – original draft, Validation, Methodology, Data curation, Writing – review & editing, Visualization, Software, Investigation. **Charisma F. Choudhury:** Project administration, Funding acquisition, Writing – review & editing, Conceptualization, Methodology, Supervision. **Stephane Hess:** Writing – review & editing, Project administration, Conceptualization, Supervision, Methodology, Funding acquisition.

Acknowledgements

The current research was funded by the Advanced Quantitative Methods (AQM) scholarship of the Economic and Social Research Council (ESRC). Charisma Choudhury acknowledges the financial support of her UK Research and Innovation (UKRI) Future Leader Fellowship MR/T020423/1-NEXUS. Stephane Hess acknowledges the financial support by the European Research Council through the consolidator Grant 615596-DECISIONS and the advanced grant 101020940-SYNERGY. We want to thank Dr. Georges Sfeir for his help and guidance in using his code for the GBP-LCCM and GP-LCCM models.

Appendix A

See Tables 14–22.

References

Anda, C., Erath, A., Fourie, P.J., 2017. Transport modelling in the age of big data. Int. J. Urban Sci. 21 (sup1), 19–42. http://dx.doi.org/10.1080/12265934.2017.1281150.

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

Arthur, D., Vassilvitskii, S., 2007. K-means++: The advantages of careful seeding. In: Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms. SODA 2007. New Orleans, Louisiana, USA, January 7-9, 2007, http://dx.doi.org/10.1145/1283383.1283494.

Batley, R., Bates, J., Bliemer, M., Börjesson, M., Bourdon, J., Cabral, M.O., Chintakayala, P.K., Choudhury, C.F., Daly, A., Dekker, T., Drivyla, E., Fowkes, T., Hess, S., Heywood, C., Johnson, D., Laird, J., Mackie, P., Parkin, J., Sanders, S., Sheldon, R., Wardman, M., Worsley, T., 2019. New appraisal values of travel time saving and reliability in great britain. Transportation 46, 583–621. http://dx.doi.org/10.1007/s11116-017-9798-7.

Ben-Akiva, M., Lerman, S., 1985. Discrete Choice Analysis: Theory and Application To Travel Demand. MIT Press, Cambridge, Massachusetts.

Bishop, C., 2006. Pattern Recognition and Machine Learning. Springer, Singapore.
Bora, D.J., Gupta, A.N., 2014. Effect of different distance measures on the performance of K-means algorithm: An experimental study in matlab. Int. J. Comput. Sci. Inf. Technol. 5 (2), 2501–2506.
Calastri, C., Crastes dit Sourd, R., 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, 175–201. http://dx.doi.org/10.1007/s11116-018-9858-7.
Calastri, C., Hess, S., Choudhury, C.F., Daly, A., Gabrielli, L., 2018. Mode choice with latent availability and consideration: Theory and a case study. Transp. Res. Part B: Methodol..
Cantarella, G.E., de Luca, S., 2005. Multilayer feedforward networks for transportation mode choice analysis: An analysis and a comparison with random utility models. Transp. Res. Part C: Emerg. Technol. 13, 121–155.
Chakraborty, A., Ghosh, J.K., 2011. AIC, BIC and recent advances in model selection. Philos. Stat. 7, 583–605. http://dx.doi.org/10.1016/B978-0-444-51862-0.50018-6.
Crawford, F., 2017. Methods for Analysing Emerging Data Sources To Understand Variability in Traveller Behaviour on the Road Network (Ph.D. thesis). University of Leeds.
Daly, A., 1982. Estimating choice models containing attraction variables. Transp. Res. Part B: Methodol. 16B (1), 5–15.
Daly, A., 2010. Cost Damping in Travel Demand Models - Report of a Study for the Department for Transport. RAND Europe.
Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B Stat. Methodol. 39 (1), 1–38. http://dx.doi.org/10.1177/019262339101900314.
Ester, M., Kriegel, H.P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. pp. 226–231.
Ghorbani, H., 2019. Mahalanobis distance and its application for detecting multivariate outliers. Facta Univ. Ser. Math. Informatics 34 (3), 583–595. http://dx.doi.org/10.22190/FUMI1903583G.
Hafezi, H.H., Liu, L., Millward, H., 2019. A time-use activity-pattern recognition model for activity-based travel demand modeling. Transportation 46, 1369–1394. http://dx.doi.org/10.1007/s11116-017-9840-9.
Hagenauer, J., Helbich, M., 2017. A comparative study of machine learning classifiers for modelling travel mode choice. Expert Syst. Appl..
Hancock, T.O., Hess, S., Choudhury, C.F., Tsoleridis, P., 2021. Decision field theory: An extension for real-world settings. In: Australasian Transport Research Forum 2021 Proceedings. 8-10 December, Brisbane, Australia.
Hasan, S., Ukkusuri, S.V., 2014. Urban activity pattern classification using topic models from online geo-location data. Transp. Res. Part C: Emerg. Technol. 44, 363–381. http://dx.doi.org/10.1016/j.trc.2014.04.003.
Hasan, S., Ukkusuri, S.V., 2015. Location contexts of user check-ins to model urban geo life-style patterns. PLoS One 10 (5), e0124819.
Hasan, S., Ukkusuri, S.V., 2018. Reconstructing activity location sequences from incomplete check-in data: A semi-Markov continuous-time Bayesian network model. IEEE Trans. Intell. Transp. Syst. 19 (3), 687–698.
Hensher, D.A., Ton, T.T., 2000. A comparison of the predictive potential of artificial neural networks and nested logit models for commuter mode choice. Transp. Res. Part E 36, 155–172. http://dx.doi.org/10.1016/S1366-5545(99)00030-7.
Hess, S., 2014. Latent class structures: Taste heterogeneity and beyond. In: Hess, S., Daly, A. (Eds.), Handbook of Choice Modelling. Edward Elgar Publishing Limited, Cheltenham, Uk, pp. 311–329.
Hillel, T., Elshafie, M.Z.E.B., Ying, J., 2018. Recreating passenger mode choice-sets for transport simulation: A case study of London, UK. Proc. Inst. Civ. Eng. - Smart Infrastruct. Constr. 171 (1), 29–49. http://dx.doi.org/10.1680/jsmic.17.00018.
Joh, C.-H., Arentze, T.A., Timmermans, H.J.P., 2001. A position-sensitive sequence alignment method illustrated for space–time activity-diary data. Environ. Plan. A 33 (2), 313–338.
Kamakura, W.A., Russell, G., 1989. A probabilistic choice model for market segmentation and elasticity structure. J. Mark. Res. 26, 379–390.
Kristoffersson, I., Daly, A., Algers, S., 2018. Modelling the attraction of travel to shopping destinations in large-scale modelling. Transp. Policy 68, 52–62. http://dx.doi.org/10.1016/j.tranpol.2018.04.013.
Krizek, K., Waddell, P., 2003. Analysis of lifestyles choices: Neighborhood type, travel patterns, and activity participation. Transp. Res. Rec. 1807, 119–128.
Krueger, R., Bansal, P., Bierlaire, M., Gasos, T., 2020. Robust discrete choice models with t-distributed kernel errors. arXiv:2009.06383.
Lanzendorf, M., 2002. Mobility styles and travel behavior application of a lifestyle approach to leisure travel. Transp. Res. Rec. 1807, 163–173.
Lloyd, S.P., 1982. Least squares quantization in pcm. IEEE Trans. Inform. Theory 28 (2), 129–136.
MacKay, D.J.C., 1998. Introduction to gaus- sian processes. In: Bishop, C.M. (Ed.), Neural Networks and Machine Learning. Springer, pp. 133—166.
McFadden, D., 1973. Conditional logit analysis of qualitative choice behavior. In: Zarembka, P. (Ed.), Frontiers in Econometrics. Academic Press, New York, pp. 105–142.
McFadden, D., 1978. Modelling the choice of residential location. In: Karlqvist, A., Lundqvist, L., Snickars, F., Weibull, J. (Eds.), Spatial Interaction Theory and Planning Models. North Holland, Amsterdam, pp. 75–96.
McFadden, D., 2000. Disaggregate behavioural travel demand's RUM side: A 30-year retrospective. In: Hensher, D. (Ed.), Travel Behaviour Research: The Leading Edge. Pergamon Press, Oxford, UK, pp. 17–63, 2000.
McFadden, D., Train, K., 2000. Mixed MNL models for discrete response. J. Appl. Econometrics 15, 447–470.
Milne, D., Watling, D., 2019. Big data and understanding change in the context of planning transport systems. J. Transp. Geogr. 76, 235–244. http://dx.doi.org/10.1016/j.jtrangeo.2017.11.004.
Papke, L.E., Wooldridge, J.M., 1996. Econometric methods for fractional response variables with an application to 401(k) plan participation rates. J. Appl. Econometrics 11, 619–632. http://dx.doi.org/10.1002/(SICI)1099-1255(199611)11:6<619::AID-JAE418>3.0.CO;2-1.
Ripley, B.D., 2009. Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge, Massachusetts.
Salomon, I., Ben-Akiva, M., 1983. The use of the life-style concept in travel demand models. Environ. Plan. A 15, 623–638.
Sekhar, C.R., Minai, Madhu, E., 2016. Mode choice analysis using random forrest decision trees. Transp. Res. Procedia 17, 644–652. http://dx.doi.org/10.1016/j.trpro.2016.11.119.
Sfeir, G., Abou-Zeid, M., Rodrigues, F., Pereira, F.C., Kaysi, I., 2021. Latent class choice model with a flexible class membership component: A mixture model approach. J. Choice Model. 41, 100320. http://dx.doi.org/10.1016/j.jocm.2021.100320.
Sfeir, G., Rodrigues, F., Abou-Zeid, M., 2022. Gaussian process latent class choice models. Transp. Res. Part C: Emerg. Technol. 136, 103552. http://dx.doi.org/10.1016/j.trc.2022.103552.
Sifringer, B., Lurkin, V., Alahi, A., 2020. Enhancing discrete choice models with representation learning. Transp. Res. Part B: Methodol. 140, 236–261. http://dx.doi.org/10.1016/j.trb.2020.08.006.
Singh, A., Yadav, A., Rana, A., 2013. K-means with three different distance metrics. Int. J. Comput. Appl. 67 (10), 13–17. http://dx.doi.org/10.5120/11430-6785.
Smith, T., Noble, M., Noble, S., Wright, G., McLennan, D., Plunkett, E., 2015. The English Indices of Deprivation 2015. Technical Report, Department for Communities and Local Government, London.
Song, Y., Ren, S., Wolfson, J., Zhang, Y., Brown, R., Fan, Y., 2021. Visualizing, clustering, and characterizing activity-trip sequences via weighted sequence alignment and functional data analysis. Transp. Res. Part C: Emerg. Technol. 126, 103007.
Train, K., 2009. Discrete Choice Methods with Simulation. Cambridge University Press, Cambridge, Massachusetts.

Tsoleridis, P., Choudhury, C.F., Hess, S., 2021. Utilising activity space concepts to sampling of alternatives for mode and destination choice modelling of discretionary activities. J. Choice Model. 42 (1), 100336. http://dx.doi.org/10.1016/j.jocm.2021.100336.

Tsoleridis, P., Choudhury, C.F., Hess, S., 2022a. Deriving transport appraisal values from emerging revealed preference data. Transp. Res. Part A: Policy Pr. 165, 225–245. http://dx.doi.org/10.1016/j.tra.2022.08.016.

Tsoleridis, P., Hess, S., Choudhury, C.F., 2022b. Accounting for distance-based correlations among alternatives in the context of spatial choice modelling using high resolution mobility data. Transp. A: Transp. Sci. 1–45. http://dx.doi.org/10.1080/23249935.2024.2401425

van Cranenburgh, S., Wang, S., Vij, A., Pereira, F., Walker, J., 2022. Choice modelling in the age of machine learning - Discussion paper. J. Choice Model. 42, 100340. http://dx.doi.org/10.1016/j.jocm.2021.100340.

Vij, A., Carrel, A., Walker, J., 2013. Incorporating the influence of latent modal preferences on travel mode choice behavior. Transp. Res. Part A: Policy Pr. 54, 164–178. http://dx.doi.org/10.1016/j.tra.2013.07.008.

Wang, S., Mo, B., Hess, S., Zhao, Z., 2021a. Comparing hundreds of machine learning classifiers and discrete choice models in predicting travel behavior: An empirical benchmark. arXiv:2102.01130.

Wang, S., Mo, B., Zhao, Z., 2021b. Theory-based residual neural networks: A synergy of discrete choice models and deep neural networks. Transp. Res. Part B: Methodol. 146, 333–358. http://dx.doi.org/10.1016/j.trb.2021.03.002.

Wong, M., Farooq, B., 2021. ResLogit: A residual neural network logit model for data-driven choice modelling. Transp. Res. Part C: Emerg. Technol. 126, 103050. http://dx.doi.org/10.1016/j.trc.2021.103050.

Xie, C., Lu, J., Parkany, E., 2003. Work travel mode choice modeling using data mining: Decision trees and neural networks. Transp. Res. Rec.: J. Transp. Res. Board 1854 (1), 50–61.

Zhang, B., Hsu, M., Dayal, U., 1999. K-Harmonic Means–A Data Clustering Algorithm. Technical Report HPL-1999-124, Hewlett–Packard Laboratories, 1999.

Zhang, Y., Xie, Y., 2008. Travel mode choice modeling with support vector machines. Transp. Res. Rec.: J. Transp. Res. Board 2076, 141–150. http://dx.doi.org/10.3141/2076-16.