Computer vision-enriched discrete choice models to explore the importance of street-level factors and commute travel times in the residential location choice behaviour

Sander van Cranenburgh[1]
Francisco Garrido Valenzuela[1]
[1]CityAI lab, Transport and Logistics Group, Delft University of Technology

Keywords
Machine learning, Computer vision, Discrete choice modelling, Residential location choice, Commute

1. Introduction

Residential location choices are strongly coupled with travel demand. Therefore, travel behaviour researchers have extensively examined and modelled residential location choices, often combined with mobility choices (Pinjari et al. 2009; Lee and Waddell 2010; Pinjari et al. 2011). The residential location choice is typically considered a complex and multi-dimensional trade-off involving at least (1) travel-related factors, such as the commute mode and commuting time; (2) accessibility-related factors, such as amenities like schools, stores, hospitals and playgrounds and (3) street-level factors (a.k.a. local environmental factors), such as the building typology, openness, presence of trees, parking conditions, disorders (e.g. due to litter, graffiti, or weeds). (Giles-Corti et al. 2013). The importance of street-level factors became strikingly clear during the COVID-19 pandemic. Without the need to consider their commute, millions of white-collar workers fled the inner cities, often for suburban areas with better local conditions (Economist 2022; Lee and Huang 2022).

In hitherto studies, street-level factors have mostly been ignored by (transport) researchers. Previous studies into residential location choice behaviour have predominantly used zones as the most disaggregate level of analysis. These studies typically have used data from bureaus of statistics, including variables such as e.g. demographics, land use, population density, ethnic compositions, accessibility measures and transport network-related measures. Yet, these data usually do not contain high-level semantic concepts that describe local street-level conditions. In fact, a concept like "openness of the residential space" is hard to define and capture into a numeric value or ascribe a meaningful semantic category.

Images are the one of the pre-eminent medium to capture high-level semantic concepts that describe local street-level conditions. The fact that nearly all housing platforms and real-estate agencies use street-level images on their websites and apps suggests two things. Firstly, it suggests that street-level factors are indeed crucial for residential location choices, and secondly, images are an effective data format to communicate them. In support of the latter point, an abundance of evidence from cognitive psychology shows images are a much more effective medium to convey information to humans, e.g. about scales, textures, and shapes, than text or numbers (Gregory 1973; Pinker 1990). Fortunately, images containing information on street-level factors are nowadays widely available from map services of tech firms like Google and Apple, and researchers have increasingly started to use them to investigate, e.g. safety perceptions of the urban space and the people's density in urban places (Dubey et al. 2016; Garrido-Valenzuela et al. 2022).

However, images have not been used to study residential location choice behaviour. The primary reason for not having used these images is of methodological nature. Discrete Choice Models (DCMs) are the chief methodology to describe and explain resident location choices. But, current DCMs cannot handle image data. That is, at present, there is no algorithmic way to incorporate the information contained in images associated with an alternative directly into a choice model. This inability of DCMs is remarkable, given that visual imagery is indispensable to many of today's multi-attribute decision situations. Nowadays, it is hard to imagine booking a hotel or searching for a house online without having access to images. Webshops often show multiple images per product.

This research aims to develop computer vision-enriched discrete choice model. Computer vision-enriched DCMs are models capable of handling choice tasks in which each alternative comprises a set of numeric attributes and an image. Computer vision-enriched DCMs aim to maintain the DCMs' strengths while extending the realm of application of DCMs to choice situations consisting of a combination of images and numeric attributes. We demonstrate the workings of the proposed computer vision-enriched discrete choice models by shedding light on the importance of street-level factors to residential location choice behaviour. To this end, we developed and administered a novel stated choice experiment involving a trade-off between commute time, monthly housing cost (numeric attributes) and street-level factors (image). Thereby, this research methodologically contributes to the recent work in the travel behaviour field that aims to bring machine learning and DCMs closer together (e.g. Rossetti et al. 2019; Sifringer et al. 2020; Arkoudi et al. 2021; Ramírez et al. 2021; van Cranenburgh et al. 2021).

2. Methodology

2.1. Computer vision

In the last decade, significant developments have taken place in Computer Vision (CV). State-of-the-art CV models can accurately detect scenes and objects in images (Gu et al. 2018). Nowadays, numerous applications use CV, ranging from detecting faults in production processes to tumours in MRI scans. Moreover, many pre-trained CV models are available nowadays, lowering the computational burden and the need for large amounts of data.

Most CV models conceptually consist of two parts: a future extractor which produces a feature map from the image, and a classifier which is used to make predictions (see Figure 1). A feature map is a lower-dimensional representation of the original image. It contains (most of) the information of the image but is more compact in form. Usually, a feature map is a flat array of floating points. Traditionally, researchers have used Convolution Neural Networks (CNNs) to extract feature maps from images. A CNN typically comprises a series of convolution layers, reducing the dimensionality of the data. More recently, Vision transformers have started to dominate the CV field (Dosovitskiy et al. 2020). Vision transformers are XXXX.

Like CNN, vision transformers also produce feature maps. Finally, the classifier (typically a fully connected multilayer perceptron) maps the image onto a class, taking the extracted feature maps from a CNN or vision transformer as inputs.

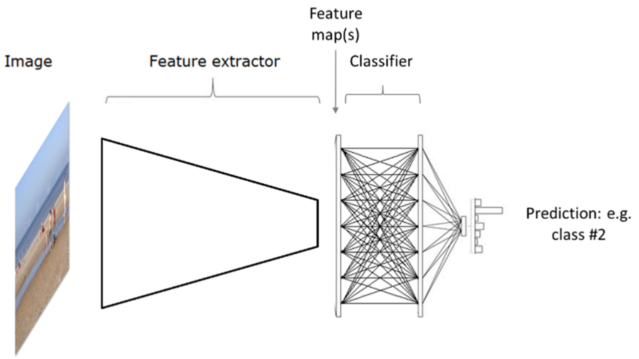


Figure 1: Typical CV model architecture

2.2. Modelling framework

Throughout this paper, a choice task C comprises alternatives $j = \{1 \ldots J\}$. Each alternative j is described by a M continuous numeric attributes $X_j = \{x_{j1}, x_{j2}, \ldots, x_{jM}\}$ and an image: S_j. Images consist of pixels, the basic building blocks of images. Each pixel has a spatial location (h x w) and a colour value. Most colour images nowadays use three colour channels: Red (R), Green (G), Blue (B), and 8 bits per colour channel (implying three 0-225 values). Therefore, mathematically, images are represented as 3D tensors: Height x Width x Colour channel.

In our modelling framework, we assume decision-makers, denoted n, make decisions based on Random Utility Maximising (RUM) principles (McFadden 1974). For each alternative j, a decision-maker experiences utility, denoted U_j, from the numeric attributes X_j and the image S_j, see Equation 1. Furthermore, to account for the fact that the analyst observes not everything that matters to the decision-makers utility, an error term ε_{jn} is included. The error term represents the unobserved part of utility by the decision-maker n on the alternative j ,

$$U_{jn}\left(X_{jn}, S_{jn}\right) = V\left(X_{jn}, S_{jn}\right) + \varepsilon_{jn} \qquad \text{Equation 1}$$

We make three additional assumptions to develop discrete choice models capable of handling choice tasks comprising images. Firstly, we assume that the utility derived from the numeric attributes and the image are independent of the other alternatives and additive in utility space, see Equation 2.

$$U_{jn}\left(X_{jn}, S_{jn}\right) = f\left(X_{jn}\right) + g\left(S_{jn}\right) + \varepsilon_{jn} \qquad \text{Equation 2}$$

The second assumption that we make is that utility is linear with numeric attributes as well as with images' feature maps, which we denote $Z_j = \{z_{j1}, z_{j2}, \ldots, z_{jK}\}$. Hence, the feature map of each image linearly enters the utility function. We let feature maps enter the utility function as opposed to individual pixel values for two reasons. Firstly, the pixel values in and of themselves are meaningless. Rather, the joint distribution of many pixels together produces higher-level meaningful concepts that can be expected to generate utility. Secondly, a medium-sized colour image of 640 x 480 pixels contains almost 1 million values. This high dimensionality renders using each pixel directly in the utility function infeasible. Therefore, we assume that the relevant information in the images can be embedded in their feature maps that linearly map onto utility, see Equation 3. In Equation 3, the function φ maps image tensor S_j onto a lower dimensional feature space Z_j (which has a dimensionality of 1 x K). Furthermore, w_k denotes the weight associated with the k^{th} feature of Z_j; β_m denotes the marginal utility associated with attribute m, and x_{jmn} denotes the attribute level of numeric attribute m of alternative j, as faced by decision-maker n.

$$U_{jn} = \overbrace{\sum_m \beta_m x_{jmn}}^{\substack{\text{Utility derived}\\ \text{from numeric attributes}}} + \overbrace{\sum_k w_k z_{jkn}}^{\substack{\text{Utility derived}\\ \text{from image feature map}}} + \varepsilon_{jn} \qquad \text{where } Z_{jn} = \varphi\left(S_{jn} \mid w_r\right)$$

Equation 3

Finally, in line with common practice in choice modelling, we assume ε_{jn} is independent and identically Extreme Value Type II distributed with a variance of $\pi^2/6$, resulting in the well-known and convenient closed-form logit formula for the choice probabilities, given in Equation 4.

$$P_{in} = \frac{e^{V_{in}}}{\sum_{j \in C_n} e^{V_{jn}}}$$

Equation 4

Figure 2 provides a graphical representation of the structure of the proposed CV-enriched DCM. An essential aspect of the CV-enriched DCM structure is that the upper (associated with the left alternative) and lower (associated with the right alternative) parts of the network are identical in architecture and the networks' weights. Thereby, the left and right alternatives are mathematically treated in the same way, allowing us to interpret the values on the nodes of the last layer as a utility. This network structure (bar the numeric attributes) is called a siamese network and is typically used to determine the similarity between two images. However, even though we can interpret the last layer as utility, it is essential to note we cannot interpret the weights $\boldsymbol{w}$ in the same way as the traditional choice model parameters $\boldsymbol{\beta}$. Like traditional choice model parameters, the weights $\boldsymbol{w}$ can be conceived as marginal utilities –after all, they reflect the marginal effect on utility. But, because the meaning and unit of the features, z_k, is unclear, they do not carry a behavioural meaning.

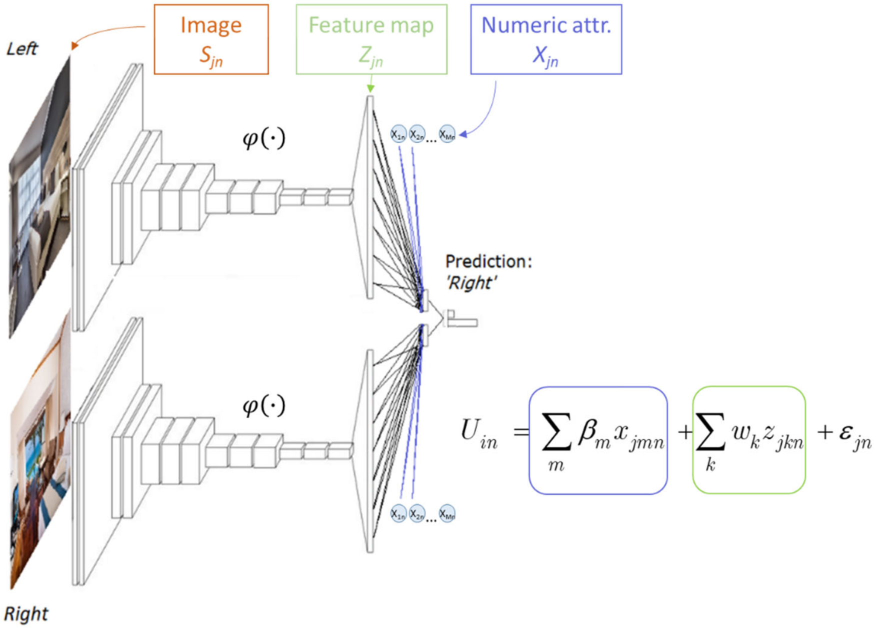


Image
S_{jn}
Feature map
Z_{jn}
Numeric attr.
X_{jn}
Left
Right
$\varphi(\cdot)$
$\varphi(\cdot)$
x_{1n} x_{2n} ... x_{Mn}
x_{1n} x_{2n} ... x_{Mn}
Prediction:
'Right'
$U_{in} = \boxed{\sum_m \beta_m x_{jmn}} + \boxed{\sum_k w_k z_{jkn}} + \varepsilon_{jn}$

Figure 2: CV-enriched DCM framework

2.3. Computer-vision model & training

In this paper, we use transfer learning (Bengio 2012). Training state-of-the-art CVs is computationally intensive and requires expertise and vast amounts of data and time. The idea of transfer learning is to use a pre-trained network as the starting point for developing another network for a closely related task. Thus, rather than retraining the whole network (typically consisting of millions of weights) from scratch, the model is trained from an already good starting point. Thereby, transfer learning lowers the computational burden and the need for large amounts of data.

We use the DeiT base (Touvron et al. 2021) as the CV model in our CV-enriched DCMs (siamese part). DeiT is a data-efficient vision transformer-based model which produces competitive capabilities on benchmark data sets, such as ImageNet, at a much lower computational cost and data requirements than many of its competitors (Touvron et al. 2021). DeiT base model consumes a relatively modest 86 million weights.

3. Data collection

To demonstrate the merits of CV-enriched DCM, we collect residential location choice data through a novel stated choice experiment. In the residential location choice, both numeric attributes and visual information can be expected to be important to the choices. A requirement to effectively train the CV part of the model is that we have sufficient variation in the images. Sufficient variation enables the model to learn the high-level characteristics of images that generate (dis)utility. Therefore, we built our stated choice experiment to involve many images.

3.1. Stated choice experiment

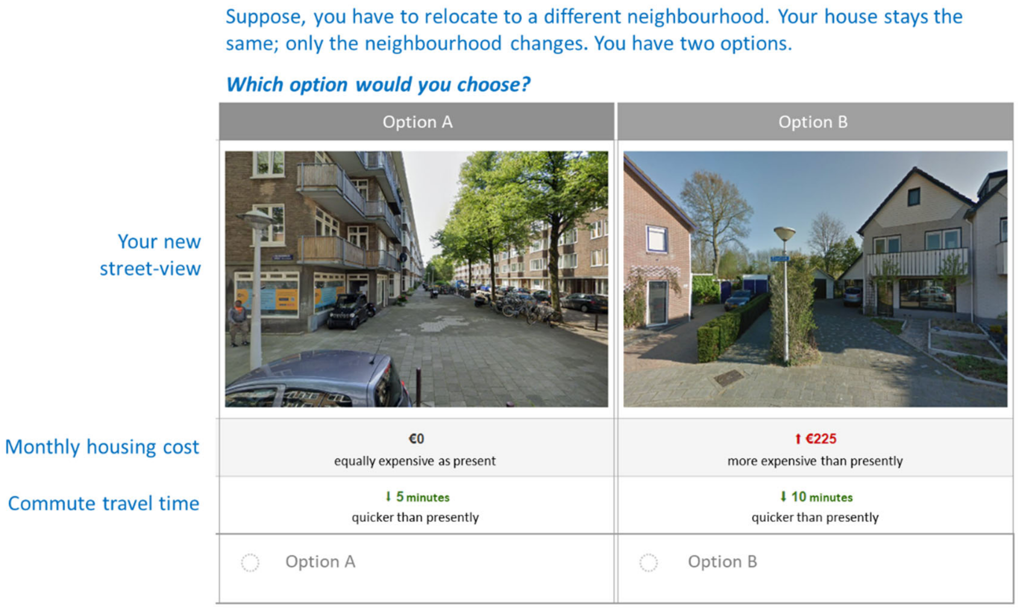
Specifically, we conduct a Stated Choice (SC) experiment in which we ask respondents to envision they were forced to relocate to a new neighbourhood. Furthermore, respondents were told: (1) the house they would move to is identical to their current home in terms of size, type, and built-year, etc. (2) the monthly housing cost (including rent, mortgage, taxes, insurance, etc.) may go up or go down. (3) the new neighbourhood is relatively close to the respondent's current residential location, but the commute time may go up or down. The commute time is shown for their current mode of transport. (4) in all other aspects, the situation stays the same (e.g. in terms of distances to amenities, schools, and the general practitioner). (5) The images show each alternative's window view at ground level on the street side. Figure 3 shows a screenshot of the experiment.

We choose monthly housing costs (*hhc*) and commute travel time (*tti*) as the salient numeric attributes for several reasons. Firstly, they are highly important to the residential location choice. Secondly, they apply to virtually everyones' residential location choice. Thirdly, they help to interpret our empirical results. The combination of cost and time attributes allows us to compute the Value-of-Travel-Time (VTT) and, in turn, shed light on the relative importance of the street-level factors.

As can be seen in Figure 3, we decide for a pivoted experimental design. We used a pivoted design to present respondents with as realistic choice situations as possible. Using absolute levels instead of pivoted levels would presumably render many choice tasks unrealistic because of the considerable variation across respondents' current situations, especially regarding housing costs. For the attribute housing cost, we used seven pivoted levels. For the attribute travel time, the number of levels and ranges presented to the respondent depended on the respondent's current travel time, see Table 1. The ranges of both attributes were determined through a small pilot conducted before the actual survey.

Table 1: Attribute levels Stated Choice experiment

Current commute travel time of the respondent (TT_n)	Attribute levels	
	Housing cost (*hhc*) [€]	Commute travel time (*tti*) [minutes]
TT_n < 10 minutes	N/A	
10 minutes < TT_n < 20 minutes	-225, -150, -75, 0, +75, +150, +225	-5, 0, +5, +10, +15
20 minutes < TT_n < 30 minutes		-10, -5, 0, +5, +10, +15
30 minutes < TT_n		-15, -10, -5, 0, +5, +10, +15



Suppose, you have to relocate to a different neighbourhood. Your house stays the same; only the neighbourhood changes. You have two options.

Which option would you choose?

	Option A	Option B
Your new street-view		
Monthly housing cost	€0 equally expensive as present	↑ €225 more expensive than presently
Commute travel time	↓ 5 minutes quicker than presently	↓ 10 minutes quicker than presently
	○ Option A	○ Option B

Figure 3: Screenshot of the pivoted stated choice experiment (translated to English)

3.1.1. Streetview image collection

Besides monthly housing costs and commute travel time, each alternative has a street-view image. This image is randomly sampled from a database of street-view images we created before conducting the stated choice experiment. A major effort went into the construction of this database with street-view images. Specifically, we took the following steps to construct it. First, we randomly selected 50 municipalities (of about 350) in the Netherlands. We capped the number of municipalities to 50 because using more would lead to collecting many more images than we would need for our experiment. Second, we created a grid of points with 150-metres spacing within areas designated as the residential area within the selected municipalities. Third, we retrieved the nearest street-view image urls for each point on the grid using Google's API. If multiple years were available, we collected images of all available years. Each street-view image id corresponds to a 360-degree panorama view at the street level. Fifth, from each panorama, we generated two image urls with 90-degree angles to the direction of the street. This latter ensures the images are 'window views' (e.g. as opposed to views parallel to the driving direction of the Google car). Sixth, images of poor quality were removed using XXXX [REF]. More specifically, black images, blurred images and images with tilted horizons were removed. All street-view images are stored using png format with 900 x 600 pixels and 8 bits per colour channel (implying

16.7m colour values per pixel). Finally, we excluded all images taken before 2020. The final database of images contains (the url links to) a little over 60k randomly sampled street-view images of residential streets from 50 municipalities in the Netherlands.

Importantly, for each image in our database, we also stored the month of the year in which the image was taken. The Netherlands lies in temperate zones, having four distinct seasons. Even though Google collects images on dry days only, due to the seasonality, street-view images taken in the winter may look different from those taken in summer. These differences could, in turn, impact the utility experienced by the respondent from the depicted local environment (and thus must be accounted for in our models).

3.1.2. Experimental design

We used a random experimental design. Because the images do not possess ordinal or categorical levels, adopting an orthogonal or efficient experimental design strategy was not feasible. Therefore, we took a two-step approach to construct the choice tasks. First, we randomly pulled a pair of images from our image database. The only requirement imposed on the drawing was that the drawn images were not from the municipality where the respondent lives. We determined each respondent's municipality based on the postcode (which we elicited at the start of the survey). We excluded images from the respondent's municipality to avoid unobserved heterogeneity entering our experiment derived from respondents' knowledge of places the images were taken. Unobserved utilities flowing into stated choice experiments could lead to biased modelling outcomes if not econometrically accounted for (see, e.g. Train and Wilson 2008; Van Cranenburgh et al. 2014; Guevara and Hess 2019). While excluding images from respondents' own municipalities does not guarantee that respondents do not recognise the places the street-view image were taken, it lowers the probability.

Second, we added the housing cost (*hhc*) and travel time (*tti*) levels. To do so, we randomly pulled a choice task from one of the three tables we generated before conducting the SC experiment. Each table was created by taking the following steps. First, a full-factorial design was created based on the attribute levels shown in Table 1. Second, we excluded choice tasks that did not involve a trade-off between housing costs and travel time. Removing such (partially) dominating choice tasks is possible because we have strong prior beliefs for the expected sign of the preference parameters for housing cost and travel time. Third, we excluded all choice tasks where one or more attribute levels were equal. As a result of this choice task construction approach, each choice task necessarily consists of a trade-off between housing cost and travel time.

3.2. Data collection and sample description

The survey was implemented in SurveyEngine software and conducted in September 2022. The survey started with a few questions to determine respondents' eligibility for the survey. In particular, we elicited respondents' age, gender, province and current commute travel time. Then came the SC experiment, in which each respondent was presented with 15 choice tasks. The survey ended with a series of questions regarding the respondents' current housing situation (e.g. housing costs, rating of the current visual environment) and commute situation (e.g. mode of transport, number of commute days). Noteworthy, we also asked respondents how important the three attributes (housing cost, travel time and the image) were for their decisions on a scale from 1 to 10. Although it is well-known that direct elicitation of preferences is treacherous (Nisbett and Wilson 1977), it still can provide first (albeit inconclusive) evidence of the importance of the presented street-view images relative to the numeric attributes for the residential location choices.

The target population for the survey was the Dutch population of 18 years and older, with ten or more minutes of commute travel time. The latter requirement was necessary because we chose to use a pivoted experimental design. Because of this latter condition, no official population statistics exist to compare our sample against, but we do not expect this condition to affect the population statistics substantially. Therefore, care was taken that the sample was, by and large, representative of the Dutch 18 years and older population in terms of gender, age and spatial distribution across the Netherlands. Cint[1], a panel data provider, provided the panel of respondents. In total, 800 respondents completed our survey.

Table 2 shows the sample statistics. Overall, the sample is representative of the target population. Also, for the variables that are not explicitly considered during the data collection, such as, e.g. the modal split and household composition, the statistics are close to the population data (e.g. Ton et al. 2019). Further, looking at the reported monthly housing cost, we notice that the largest share of the respondents has a housing cost below €750. This seems reasonable since the average net housing cost of rental houses in the Netherlands is around €700 p/m; homeowners' average net housing cost is slightly above €900 p/m (Stuart-Fox et al. 2022).

Table 2: Sample statistics

Socio-demographic variable	**Category**	**Distribution**
Age	18 - 29 year	21%
	30 - 39 year	19%
	40 - 49 year	20%
	50 - 59 year	22%
	60 - 69 year	17%
	+70 year	1%
Gender	Male	50%
	Female	50%
Province	North (Groningen, Friesland, Drenthe) 1,3,5	12%
	East (Gelderland, Overijssel)	23%
	South (Limburg, Noord-Brabant, Zeeland)	24%
	West (N-Holland, Z-Holland, Utrecht, Flevoland)	41%
Current commute travel time (TT)	10 minutes < TT < 20 minutes	35%
	20 minutes < TT < 30 minutes	31%
	30 minutes < TT < 45 minutes	20%
	45 minutes < TT	14%
Primary mode for commute	Bike, E-bike, Scooter, Moped	30%
	Bus, Metro, Tram	8%
	Train	10%
	Car, Motor bike	52%
Commuting days per week	1 day per week	8%
	2 days per week	15%
	3 days per week	20%
	4 days per week	22%
	5 or more days per week	35%
Household composition	One-person household	26%
	Multiple-person household without children	40%
	Multiple-person household with children	34%

[1] See www.cint.com

House type	Flat, gallery, porch, apartment Terraced house Corner house Semidetached house Detached house	23% 31% 16% 14% 15%
Current monthly housing cost (HC)	HC < 750 p/m 750 p/m < HC < 1,250 p/m 1,250 p/m < HC < 1,750 p/m 1,750 p/m < HC I do not want to report	36% 33% 16% 6% 9%
Rating of own visual environment	1 (worst) 2 3 4 5 (best)	1% 6% 21% 46% 26%

3.3. Descriptive analysis

Figure 4 shows histograms of the self-reported importance levels of the street-view images (left), monthly housing costs (middle) and commute travel times (right). Figure 4 shows that the street-view images and monthly housing costs are, on average, considered equally important to the residential location choice and more important than commute travel times. The variance in the ratings across respondents is higher for the street-view images than for the monthly housing cost – suggesting a considerable amount of preference heterogeneity is present in the importance of street-level factors. However, we observe the most variance for the commute travel time. Noteworthily, the importance rating for the street-view images is weakly negatively correlated with the importance ratings for monthly housing costs (ρ = -0.10) and uncorrelated with the ratings for commute travel time (ρ = 0.02). In contrast, the importance ratings for monthly housing costs and commute travel times are strongly positively correlated (ρ = 0.36). This strong positive correlation reveals that people who find housing costs important usually also find commute travel time important, and vice versa.

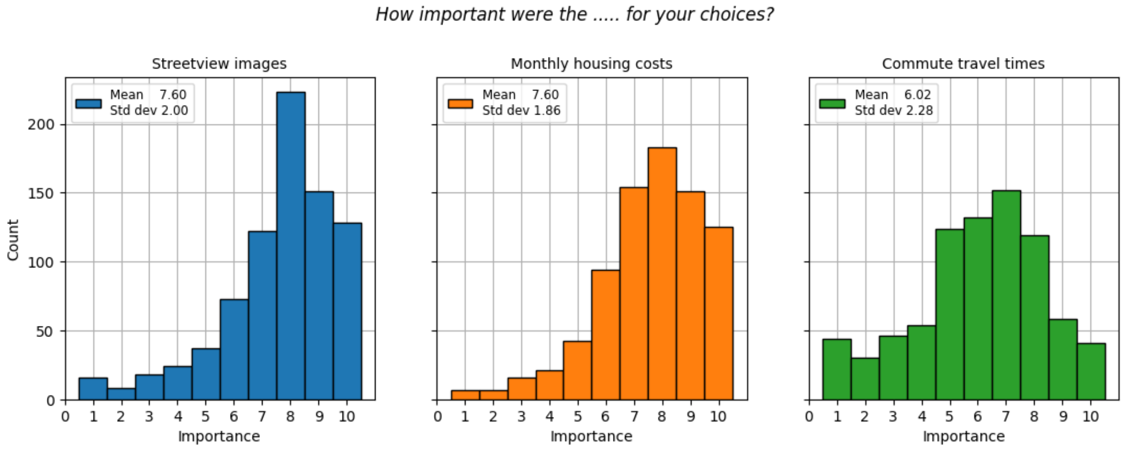


Figure 4: Self-reported importance levels of attributes in the SC experiment

Figure 5 shows the Pearson correlation coefficients between importance ratings and a selection of respondent characteristics. Interestingly, the top row shows that the importance of the street-view images correlates strongest with the self-reported rating of respondents' current visual environment.

This strong positive correlation suggests that people living in visually attractive neighbourhoods consider their visual environment relatively more important than people living in visually less attractive places. Moreover, we see that the importance of the street-view images positively correlates with living in a detached or semi-detached house. A self-selection mechanism could explain this effect: people caring about their visual environment are more likely to choose an attractive residential location. Finally, perhaps somewhat counter to expectations, we see that variables such as gender and monthly housing costs do not strongly correlate with the importance given to the street-view images.

Furthermore, Figure 5 reveals that the importance of the monthly housing cost (middle row) correlates strongest with living in house type 'Flat, gallery, porch, or apartment'. This correlation seems in line with intuition, given that low-income people are more likely to live in this type of housing. Finally, we see that the importance of the commute travel time (bottom row) positively correlates with age class 18-39 years. Since this age class sits in the centre of the working-age population, it makes sense commute travel time is essential to this group. Altogether, the correlations reported in Figure 5 seem plausible.

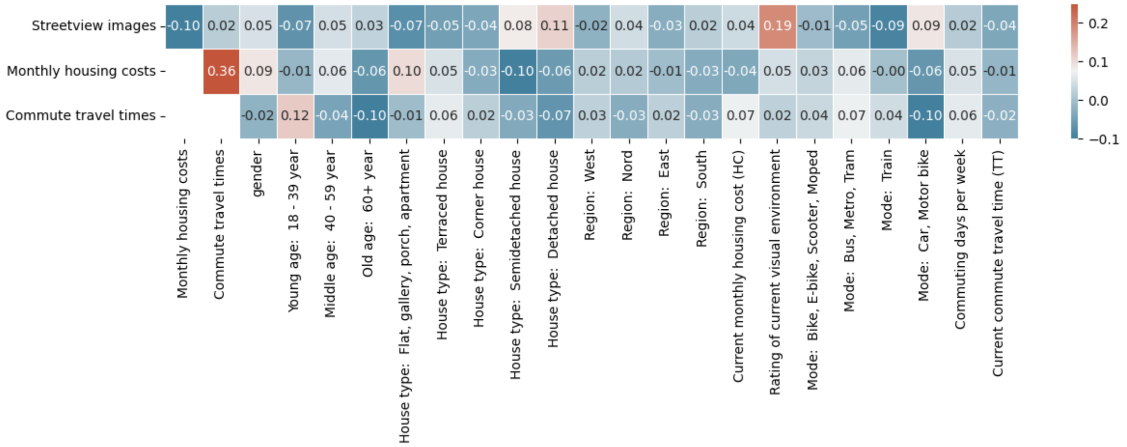


Figure 5: Pearson correlation coefficients between importance ratings and respondent characteristics

Next, we analyse the images that are used in the stated choice experiment. Although our street-view image database comprises 60k images, only slightly over 7.5k unique images are used in the stated choice experiment. Because images are drawn randomly from our image database with replacement, we expect that some images are sampled more than once. Figure 6 depicts the distribution of the number of times images are used in the SC experiment. It shows that most images are used once in line with our experimental design intention. But, counter to our design intention, Figure 6 shows that a small number of images are used 20 times or more. Observing the distribution of Figure 6 is highly improbable if the images are indeed drawn randomly. A possible underlying cause could be the survey software's use of seed numbers. But, regardless of the issue's origin, when we deal with the issue carefully during the training of our models (see section 3.2), it does not have to impact our (substantive) findings.

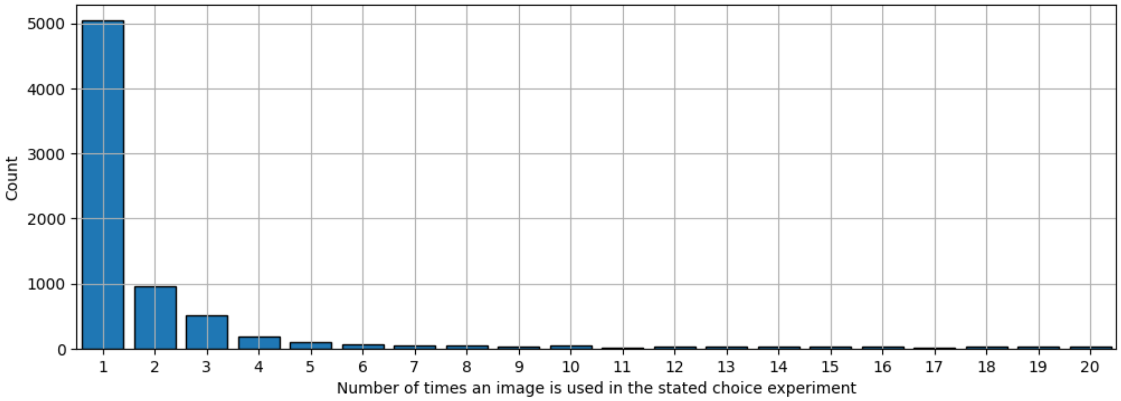


Figure 6: Distribution of the number of times images are used in the stated choice experiment

Finally, Figure 7 shows the distribution of the month of the year of the images used in the survey. In line with expectations, the images are not evenly distributed over the year. We see that most images are taken in spring and summer (March to September). Furthermore, we notice that images are sampled for all 12 months. This implies we can account for the impact of the seasons on the utility derived from the street-view images by estimating constants for all months (except one, which we need to fix to zero for normalisation).

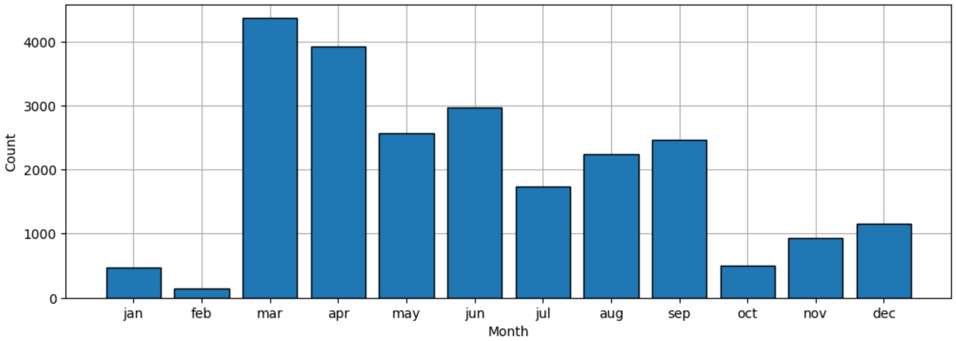


Figure 7: Distribution of images used in the stated choice experiment over the months of the year

4. Results

4.1. Training

Our CV-enriched DCM is implemented and trained in PyTorch. PyTorch is a Python-based machine learning package commonly used for deep learning computer vision research because it supports GPU computing. Training the CV-enriched DCM involves finding the weights of the model (β, w) that minimise the loss function. For the CV-enriched DCM, the loss function equals the cross-entropy loss[2] plus an L2 regularisation term, see Equation 5. The L2 regularisation aims to reduce the chance of model overfitting by penalising the magnitude of the weights in the model. γ governs the strength of the regularisation. Note that we apply regularisation only to w, thus not to our models preference parameters β.

[2] Note that minimising the cross-entropy loss is equivalent to maximising the log-likelihood of the data given the model.

$$w^{*},\beta^{*} = \operatorname{argmin}_{w,\beta}\left[\overbrace{\frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{J} y_{nj}\log\left(P_{nj}\mid x_{nj},\beta,w\right)}^{\text{cross-entropy loss}} + \overbrace{\gamma\sum_{r=1} w_{r}^{2}}^{\text{L2 regularisation}}\right] \qquad \text{Equation 5}$$

We have made the model's source code and data available to support model-building and validation practices.[3] We hope our data can become a benchmark data set for studying choice behaviour in the presence of visual stimuli.

4.1.1. Train-test split

Splitting the data into a train set and a test set is essential for training virtually all machine learning models because their high capacity makes them prone to overfitting. As the name suggests, the train set is used for training the model; the test is unseen by the model during training and used to evaluate (test) the model's generalisation performance after training. If a trained model overfits the data, a gap in the performance between the train and test set will tell.

The most common way to create the train-test split is by randomly allocating observations to the two sets. But, it is important to avoid 'data leakage'. Data leakage happens when the model has access to information during training that it does not have when deployed after training. For this study, we split our data across images to create the train and test sets. By doing so, we aim to avoid potential data leakage from learning the utility levels of specific streetview images, rather than generalisable high-level utility-generating features that the images depict. Making such a split is, however, a nontrivial network problem. Every image is connected to at least one other image (the other street-view image presented in the choice task). But some images are connected to dozens of others images (see XXX). Hence, when we assign one image to the train set, we must place all directly connected images in the train data set too. Some of these directly connected images may have been used in other choice tasks. Hence, they are connected to more images that also must be placed in the train set, and so on.

Given the above network problem, we took the following procedure to create the train and test sets. We randomly picked one choice task, comprising two images, and put this choice task and *all* choice tasks connected to this one in the train set. We repeated the random picking of choice tasks until 80% of the data were used. The remaining data (20%) make the test data set. The train and test data sets comprise *N* = 9,784 and *N* = 1,948 choice observations. Due to our splitting strategy, observations of the same individual may be present in the train and test data set. However, it is unlikely to cause data leakage because no socio-demographic variables (that would be needed to identify observations of the same respondent) are used in the training of the CV-enriched DCM.

4.1.2. Image transformation and feature scaling

In line with common practice in computer vision, we transform and augment images during training the CV-enriched DCM. Specifically, we conduct the two operations. First, we downsize images to 224 x 224 pixels. This downsizing operation ensures that images have the input dimensions expected by the CV model (i.e. DeiT base model). Second, we randomly flip images horizontally. This data augmentation operation reduces the model's ability to remember images, thus lowering the chance of overfitting the training data. Furthermore, we scale the numeric features. Scaling the features helps the optimiser to avoid getting stuck in local minima. The most common type of scaling in machine learning

[3] Github/

involves shifting and scaling the features to a zero mean and a unit variance. We use another commonly used scaling technique to scale the housing cost and travel time features, called min-max scaling. This scaling entails scaling the features to a range of [-1,1]. The advantage of this scaling technique is that it is straightforward and facilitates easy interpretation of the model's parameters. To facilitate interpretation, we have used the same scaling for all data (thus ignoring that the minimum travel time level varied across respondents, see Table 1). Specifically, all housing costs are divided by 225 and travel times are divided by 15.

4.1.3. Hyperparameter tuning

We have determined the hyperparameters of the CV-enriched DCM using a 'heuristic search approach'. We have tried various combinations of optimisation algorithms, learning rates, batch sizes, and regularisation settings to determine the optimal hyperparameters. Ideally, we would have conducted a full-fledged hyperparameter tuning in which optimisation algorithms, learning rates, batch sizes, and regularisation parameters would be tested. But, given the computational cost of training CV-enriched DCM (and CV models more generally), we refrained from conducting such a hyperparameter tuning. We found the following hyperparameters to work best (Table 3).

Table 3: Hyperparmeters CV-enriched DCM

Hyperparameter	**Value**
Optimisation algorithm	Stochastic Gradient Descent
Batch size	20
L2 weight decay (γ)	0.1
Learning rate	1e-6

4.2. Estimation results

Table 4 shows the results of the estimations for the three models we estimated/trained, and Equation 6 shows the associated utility functions. Models 1 and 2 are standard linear-additive RUM-MNL models used as benchmark models to compare the proposed CV-enriched DCM (Model 3). All models assume decision-makers experience (dis)utility from the housing cost and travel time in a linear and additive fashion. Model 1 ignores the images completely, while Model 2 takes into account the month in which the image is taken by estimating constants, denoted β_{mo}, for each month. If where and when images are collected are uncorrelated, we expect that images taken in spring and summer, on average, attain a higher utility than images taken in autumn or winter. Model 3 is the proposed CV-enriched DCM and takes the monthly housing cost (*hhc*), commute travel time (*tti*) and the month of the year as numeric input attributes in the same way as Model 2 does, but also takes the feature maps of the images as inputs. For each model, Table 4 shows the performance on the train and test sets, using three (related) metrics: the log-likelihood, rho-square and cross-entropy. A good performance on the test set implies the model generalises well to unseen data. In case we observe a gap between the performance on the train and test set commonly, overfitting is likely taking place.

$$U_{in} = \beta_{hhc}hhc_{in} + \beta_{tti}tti_{in} + \varepsilon_{in} \qquad \text{Model 1}$$

$$U_{in} = \beta_{hhc}hhc_{in} + \beta_{tti}tti_{in} + \sum_{mo}\beta_{mo}I_{S_j} + \varepsilon_{in} \qquad \text{Model 2}$$

$$U_{in} = \beta_{hhc}hhc_{in} + \beta_{tti}tti_{in} + \sum_{mo}\beta_{mo}I_{S_j} + \sum_{k} w_k z_{ikn} + \varepsilon_{in} \qquad \text{Model 3} \qquad \text{Equation 6}$$

$$\text{where} \quad I_{S_j} = \begin{cases} 1 \text{ if } mo = S_j^{mo} \\ 0 \text{ else} \end{cases}, z_{ikn} = \phi\left(S_{in} \mid w_r\right),$$
$$\varepsilon_{in} \sim \text{i.i.d. Extreme Value Type II}$$

Based on Table 4, we can draw two main conclusions. Firstly, the CV-enriched DCM can extract relevant information from the street-view images to predict the choice behaviour. Looking at the model generalisation performance, we see that CV-enriched DCM outperforms the two benchmark models by a large margin. Specifically, the CV-enriched DCM improves the log-likelihood on the test set by 57 log-likelihood points. The rho-square jumps from 0.116 to 0158. Secondly, the month of the year carries limited information regarding the utility generated by the images, at least when used in isolation from other information from the images, as in Model 2. Comparing Models 1 and 2, we observe that Model 2 outperforms Model 1 by 23 log-likelihood points on the train set but performs equally well on the test set. Hence, the plain incorporation of the month of the year in the utility function does not improve the generalisability of the standard RUM-MNL models. Thirdly, despite having to train 86 million weights, the extra computational time does not render the CV-enriched DCM impractical; 4.5 hours of training time is in the same order of magnitude as the estimation time of conventional mixed logit models.

We can make the following observations by looking at the estimated preference parameters in Table 4. Firstly, housing cost and commute travel time are highly relevant attributes to the residential location choice. In line with expectations, β_{tti} and β_{hhc} are highly significant, and their minus signs align with behavioural intuition. Based on β_{tti} and β_{hhc}, we also compute the Value-of-Travel-Time (VTT).[4] In the context of our SC experiment, the VTT gives the (mean) Willingness-to-Pay (WTP) per month for a one-hour travel time reduction. A VTT between €217 and €228 per hour per month seems reasonable. Furthermore, we observe that the VTT is stable across the three models. We expect stable β_{tti}/β_{hhc} ratios because our experimental design is constructed such that images and numeric attribute levels within choice tasks are entirely uncorrelated. Cramer (2005) shows that ratios of logit model estimates are unaffected by omitted variables if the omitted variables are uncorrelated with other explanatory variables.

Secondly, although Model 2 does not generalise better than Model 1, the signs of the estimates associated with the months of the year are mostly intuitive. These estimates reflect the average utility difference between an image taken in that month with images taken in December (which is fixed to zero). The positive and mostly significant estimates for spring and summer months can be explained by the notion that images taken in these months are more likely to look more attractive to live than images

[4] $VTT = 60\left(\frac{225}{15}\right)\frac{\beta_{tti}}{\beta_{hhc}}$ Note that the factor (225/15) comes from the fact that the attributes are scaled before training.

taken in winter, for instance, because the weather is better. However, the positive and significant estimate for January counters this line of argumentation. Now we turn to the estimates associated with the months of the year of Model 3. Importantly, these estimates do not carry the same interpretation as under Model 2. The utility derived from an image in Model 3 is the sum of the utility from the feature map and the month of the year. As a result, we cannot see the estimates associated with these two utility sources in isolation. One noteworthy observation concerning the month-of-the-year estimates in Model 3 is that fewer estimates are significant in Model 3 than in Model 2. This observation aligns with statistical expectations because images' feature maps already contain information about the month of the year. For instance, trees that have shed their leaves indicate the image is probably taken in winter. Therefore, the month-of-the-year variables contain little additional information for the model to explain the residential location choices.

Table 4: Estimation results

		Model 1			**Model 2**			**Model 3**		
	Model type	lin-add RUM-MNL			lin-add RUM-MNL			CV-enriched DCM		
	Number of parameters	2			13			86m		
	Estimation time	<1 sec[I]			<1 sec[I]			4.5 hr.[II]		
Train set N = 9784	Log-Likelihood	-5,954			-5,931			-5,724		
	ρ^2	0.120			0.130			0.156		
	Cross-entropy	0.609			0.606			0.585		
Test set N = 1948	Log-Likelihood	-1,194			-1,194			-1,137		
	ρ^2	0.116			0.116			0.158		
	Cross-entropy	0.613			0.613			0.585		
		est	*s.e.*	*p-val*	*est*	*s.e.*	*p-val*	*est*	*s.e.*[III]	*p-val*[III]
	β_{hhc}	-0.86	0.025	0.00	-0.87	0.024	0.00	-0.96	0.025	0.00
	β_{tti}	-0.21	0.023	0.00	-0.21	0.025	0.00	-0.24	0.026	0.00
	β_{jan}				0.46	0.129	0.00	0.25	0.136	0.07
	β_{feb}				0.02	0.228	0.91	-0.40	0.240	0.10
	β_{mar}				0.10	0.080	0.23	0.05	0.084	0.58
	β_{apr}				0.25	0.080	0.00	0.36	0.084	0.00
	β_{may}				0.28	0.084	0.00	0.08	0.088	0.39
	β_{jun}				0.17	0.084	0.04	-0.12	0.088	0.16
	β_{jul}				0.21	0.094	0.02	0.31	0.098	0.00
	β_{aug}				0.24	0.087	0.01	0.12	0.092	0.17
	β_{sep}				0.19	0.085	0.03	0.33	0.089	0.00
	β_{oct}				0.46	0.131	0.00	0.40	0.138	0.00
	β_{nov}				-0.11	0.106	0.31	-0.04	0.111	0.74
	β_{dec}				0.00	--fixed		0.00	--fixed	
	Value-of-Travel-Time [€/hr month]	216.7	28.26	0.00	217.2	28.35	0.00	228.5	26.73	0.00

[I] Using 4 CPUs (Xeon @ 3.60 GHz)

[II] Using GPU (GeForce RTX 2080Ti)

[III] Obtained though computing the hessian while keeping the utility derived from the image

Lastly, we analyse the contributions to utility differences between the right and left-hand side alternatives derived from the images' feature maps. Figure 8 shows three kernel density plots. The left-hand side plot shows the total utility difference as predicted by the trained CV-enriched DCM; the middle plot shows the utility difference arising from the numeric attributes; and the right-hand side plot shows the utility difference arising from the images. We make several observations based on Figure 8. Firstly, looking at the x-axes of the middle and right-hand side plots, we see that the utility differences arising from the images and numeric attributes are in the same range. This tells us that the part-worth utilities derived from housing costs and commute travel times are of the same magnitude as those derived from street-level factors embedded in the street-view images (given the ranges of the numeric attributes). This observation adds to the evidence that street-level factors are important to residential location choice behaviour and can effectively be captured and modelled using images and CV-enriched

DCMs. Secondly, we notice that the distributions of utility differences are virtually equal for the test and train sets. This observation indicates the model does not overfit the training data, and the data are adequately split into train and test sets. Therefore, the CV-enriched DCM must have learned to extract high-level generalisable features from the images that generate utility. Thirdly, we see that the distribution of the utility differences stemming from the images is comparatively more bell-shaped than those of the numeric attributes. At first sight, this may look odd, but it can be explained by how the choice task are constructed. Recall that we removed choice tasks without trade-offs between the numeric attributes (see section 3.1.2 for more details). This removal leads to the bi-modal shape of the utility difference.

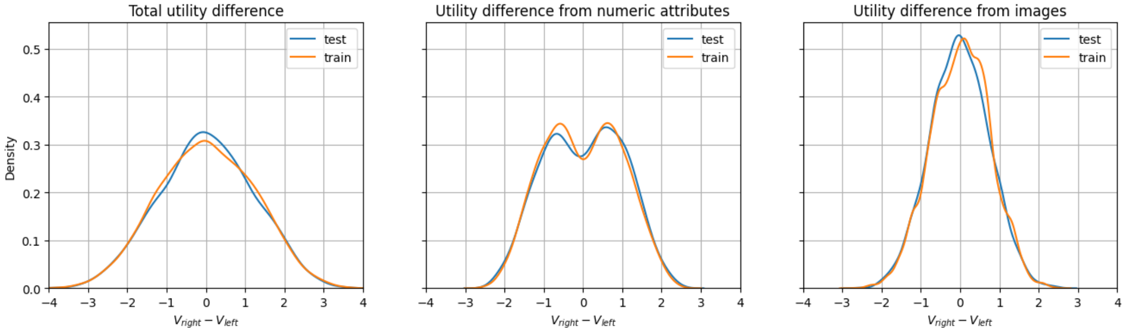


Figure 8: Utility differences

4.3. Face validity: what has the CV-enriched DCM learned about images?

The considerable performance improvement by the CV-enriched DCM compared to the benchmark model (see Table 4) supports the notion that the CV-enriched DCM can extract relevant information from images to predict choice behaviour. But, as the weights of the CV part of the model do not carry behavioural meaning, it does not directly provide insights about what the CV-enriched DCM has learned about what decision-makers find relevant for their residential location choices. To shed light on this, we show two collages of images (taken from the test set), to which the trained CV-enriched DCM assigns the highest (Figure 9) and lowest (Figure 10) utility levels. Note that the utility level is stamped in the top left of each image.[5] These utility levels are 'uncorrected' for the month of the year. Hence, the top left image yields a utility of 1.63 if the image was taken in December, while it produces a utility of 1.63-0.12 = 1.51 if it was taken in August (which it is).

What catches the eye in Figure 9 is that the images all look spacious, green and often water-abundant. We see many trees, grassland and detached houses. In the authors' view, these street-view images indeed are highly attractive residential locations. In sharp contrast, the images in Figure 10 look cramped, greyish, and urbanised and often have hallmarks of transportation, such as overhead wires, bus stops, parked bikes, and cars. In the authors' views, these street-view images are highly unattractive as residential locations.

[5] Note that the mean utility derived from images across all images is 0.02 (thus not exactly zero). Utility has no absolute scale of level (Train 2003).

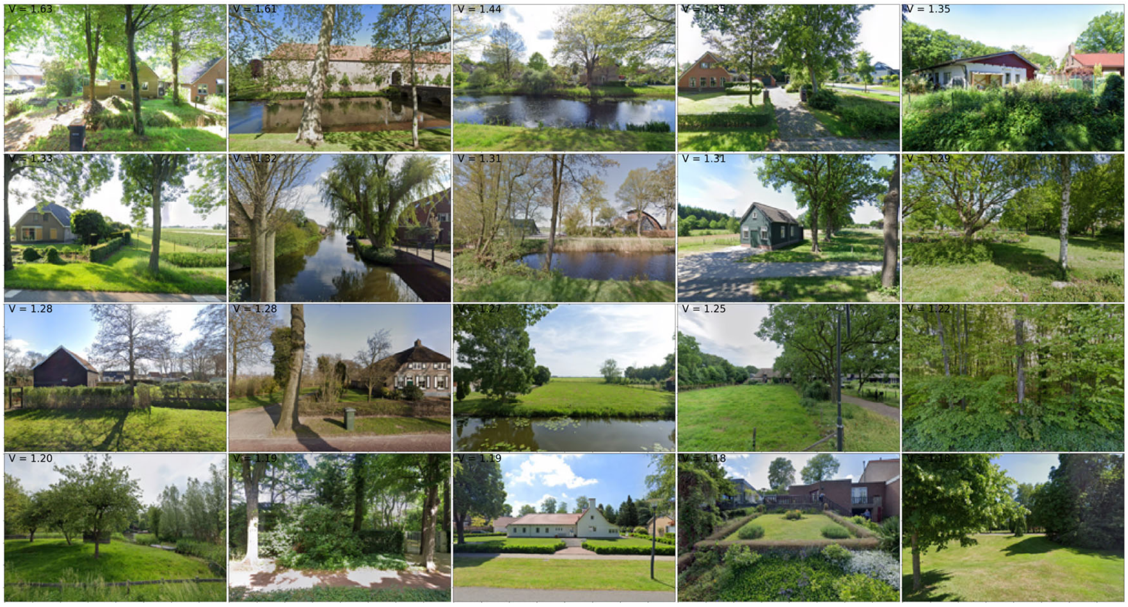


Figure 9: Images from the test data set with the highest predicted utility levels

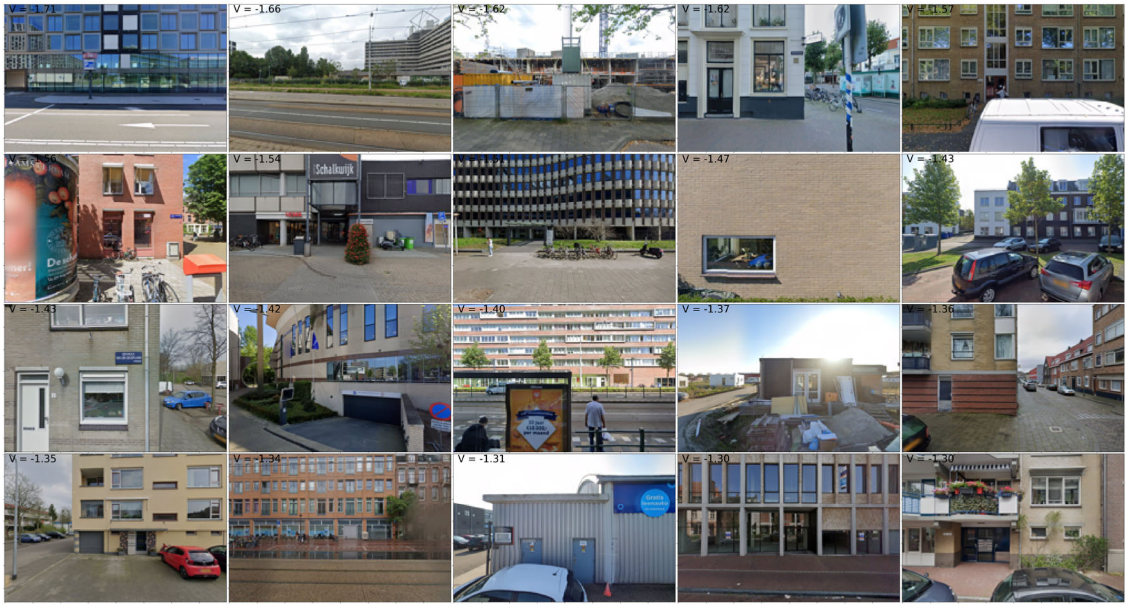


Figure 10: Images from the test data set with the lowest predicted utility levels

4.4. Heterogeneity in residential location choice preferences

To further demonstrate the merits of the CV-enriched DCM for modelling residential location choice behaviour and, more generally, modelling choice behaviour in the context of visual and numeric stimuli, we show how it can be used to investigate preference heterogeneity in residential location choice behaviour. For this purpose, we use the Latent Class (LC) modelling framework (Greene and Hensher 2003; Hess 2014). Because the class membership function in LC models can be parameterised based on the socio-demographic characteristics of decision-makers, they are particularly well-suited to uncover profiles of groups with different preferences. Equation 7 shows the general log-likelihood function of the (panel) LC model, where $P\left(i_{nt} \mid \beta_s\right)$ is the probability predicted by class *s* of the chosen alternative in choice task *t* of decision maker *n*. In LC models, a class membership function, $f\left(\gamma, g_n\right)$, probabilistically allocates respondents to different classes, and each class *s* has its own sets of preference parameters $\beta_s = \left\{\beta_{1s}, \beta_{2s}, \ldots, \beta_{Ms}\right\}$. The probability decision maker *n* belongs to class *s*, π_{ns}, is thus a function of the individual's characteristics, $g_n = \left\{g_{n1}, g_{n2}, \ldots, g_{nQ}\right\}$, and estimable parameters $\gamma = \left\{\gamma_1, \gamma_2, \ldots, \gamma_Q\right\}$.

$$LL\left(\beta, \pi\right) = \sum_{n=1}^{N} \ln\left[\sum_{s=1}^{S} \pi_{ns} \prod_{t=1}^{T_n} P\left(i_{nt} \mid \beta_s\right)\right] \qquad \text{Equation 7}$$

$$\text{where } \pi_{ns} = f\left(\gamma, g_n\right)$$

The estimation of a full-fledged CV-enriched DCM in LC is currently technically beyond reach. The already high number of weights (86m) would further increase, scaling linearly with the number of classes. To avoid the number of weights exceeding technical constraints[6], we freeze the CV part of the model. We fix the weights associated with the CV part, i.e. w_r and w_k, to their trained values. Consequently, the average utility level of image S_{int}, denoted V_{int}^{emb}, can be computed before estimation and enters the utility function as an explanatory attribute, see Equation 8. The associated preference parameter, β_{emb}^{s}, represents the importance given to the images' utility levels.

$$P\left(i_{nt} \mid \beta_s\right) = \frac{e^{V_{int}}}{\sum_j e^{V_{jnt}}}, \quad \text{where } V_{int} = \beta_{hhc}^{s} hhc_{int} + \beta_{tti}^{s} tti_{int} + \beta_{emb}^{s} V_{int}^{emb}, \qquad \text{Equation 8}$$

$$V_{int}^{emb} = \sum_k w_k \phi\left(S_{int} \mid w_r\right)$$

We estimated the CV-enriched LC model of Equation 8, using Apollo (Hess and Palma 2019), for two to five latent classes using the training set. In line with the practices in machine learning, the optimal number of classes is determined based on generalisation performance (as opposed to statistical metrics, such as BIC and AIC) (Parady et al. 2020). To do so, we applied the trained LC models to our test data set and determined the log-likelihood (i.e. LL_{test}). Although the 4-class model marginally outperforms the 3-class LC model, we choose the 3-class model for further development for reasons of interpretability. We parameterised the class membership function using the respondents' characteristics, g_n, reported in Table 2. To determine the final set of respondent characteristics, a series of LC models were estimated with different parametrisations onf the membership model in which we tested whether characteristics significantly contributed to the membership function, at least with a critical level of significance of α = 0.1.

[6] e.g. available RAM on the GPU

Table 5 shows the final LC estimation results, based on which we can make the following inferences. Firstly, accounting for preferences heterogeneity increases the model performance substantially. Specifically, the log-likelihood on the test data sets drops 77 LL points compared to Model 3; the best single-class model (see Table 4). Secondly, the three latent classes represent three highly distinct sets of preferences. Specifically, respondents typical to Class A (the largest class) consider the street-level factors and commute travel time to be the most important attributes and monthly housing costs to be the least important. Respondents typical to Class B consider monthly housing cost to be of greatest importance, attributing comparatively little importance to street-level factors and commute travel times. Respondents typical to Class C seem to be only sensitive to the monthly housing costs. But, the small sizes of the estimates of this class indicate highly stochastic choice behaviour.

Table 5: LC estimation results

		Model 4
Model type		CV-enriched LC model
Number of parameters		35
Estimation time		1.5 min[I]
Train set N = 9784	Log-Likelihood	-5,158
	ρ^2	0.239
	Cross-entropy	0.527
Test set N = 1948	Log-Likelihood	-1,060
	ρ^2	0.215
	Cross-entropy	0.544

	Class A [45%]				**Class B [41%]**				**Class C [14%]**			
Class description	Residents with high WTP for street level factors and high VTT				Residents with low WTP for street level factors and moderate VTT				Residents only caring about monthly cost			
Choice model estimates	*est*	*s.e.*	*t-val*	*p-val*	*est*	*s.e.*	*t-val*	*p-val*	*est*	*s.e.*	*t-val*	*p-val*
β_{hhc}	-0.51	0.056	-9.07	0.00	-2.59	0.138	-18.84	0.00	-0.20	0.091	-2.19	0.03
β_{tti}	-0.24	0.050	-4.75	0.00	-0.54	0.068	-7.86	0.00	-0.14	0.094	-1.47	0.14
β_{emb}	1.58	0.094	16.79	0.00	0.86	0.084	10.28	0.00	0.08	0.142	0.59	0.56
VTT[II]	€422	101.9		0.00	€186	31.4		0.00	€628	637.1		0.32
$WTP_{street\text{-}level\ factors}$[II]	€701	114.2		0.00	€75	10.8		0.00	€95	187.8		0.61
Class membership model estimates												
δ	0.00	--fixed			0.36	0.557	0.64	0.52	0.57	0.980	0.58	0.56
γ_female	0.00	--fixed			0.34	0.220	1.53	0.13	-0.77	0.414	-1.86	0.06
γ_age_young (age≤39yr.)	0.00	--fixed			0.00	--fixed			0.00	--fixed		
γ_age_middle (40yr.≤age≤59yr.)	0.00	--fixed			-0.37	0.245	-1.51	0.13	-1.73	0.511	-3.39	0.00
γ_age_old (60yr.≤age)	0.00	--fixed			-0.28	0.313	-0.90	0.37	-2.37	0.799	-2.97	0.00
γ_housetype_flat_porch_apartment	0.00	--fixed			0.00	--fixed			0.00	--fixed		
γ_housetype_terraced_house	0.00	--fixed			-1.03	0.339	-3.03	0.00	-1.18	0.575	-2.05	0.04
γ_housetype_corner_house	0.00	--fixed			-1.07	0.385	-2.78	0.01	-0.40	0.603	-0.67	0.51
γ_housetype_semidetached_house	0.00	--fixed			-1.75	0.407	-4.30	0.00	-1.58	0.766	-2.07	0.04
γ_housetype_detached_house	0.00	--fixed			-2.03	0.412	-4.93	0.00	-2.52	0.790	-3.19	0.00
γ_region_West	0.00	--fixed			0.00	--fixed			0.00	--fixed		
γ_region_Nord	0.00	--fixed			0.10	0.387	0.27	0.79	0.30	0.580	0.51	0.61
γ_region_East	0.00	--fixed			0.13	0.272	0.50	0.62	-1.87	0.745	-2.51	0.01
γ_region_South	0.00	--fixed			-0.06	0.279	-0.22	0.83	-0.64	0.555	-1.15	0.25
γ_commuting_days_per_week	0.00	--fixed			0.15	0.081	1.83	0.07	0.20	0.158	1.25	0.21

[I] Using 4 CPUs (Xeon @ 3.60 GHz)

[II] $VTT = 60\left(\frac{225}{15}\right)\frac{\beta_{tti}}{\beta_{hhc}}, WTP = -225\frac{\beta_{emb}}{\beta_{hhc}}$

Now we turn to the class membership model. The first thing that catches the eye is the characteristics shown in Figure 5 that are not reported in Table 5. Based on the bi-variate Pearson correlations shown in Figure 5, we would, for instance, expect to find the rating of the current visual environment and modes of transport to co-explain the class membership of decision makers. But, these variables do not significantly affect the membership to the classes (and thus are dropped from the final model, reported in Table 5). Instead, Table 5 shows that age and house type are strongly associated with class memberships.

To further facilitate the interpretation of the membership model, Figure 11 visualises the across-class (left) and within-class profiles (right). As the name suggests, the across-class profile shows the distribution for each characteristic level across the classes. In the across-class profile, the probabilities in each row sum up to one. The across-class profile, for instance, tells us how males are distributed across the three classes. The within-class profile shows the distribution across respondent characteristics (e.g. gender) within a class. It sheds light on the share of males and females across the respondents within a given class. Hence, in the within-class profile, the probabilities column-wise sum up to one (for each characteristic).

Based on Figure 11, we make several observations. Firstly, the across-class profile plot shows that most males are in Class A, while most females are in Class B. But, the within-class profile also reveals that males are particularly overrepresented in class C. Secondly, most middle and older respondents are in Class A, while most young respondents are in class B. But, looking at the within-class profiles, we see that young respondents are particularly overrepresented in Class C. Thirdly, most respondents living in a flat, gallery, porch or apartment are in Class B, while respondents living in other house types are mostly in Class A. Within the classes we, however, see that respondents living in a flat, gallery, porch or apartment are particularly overrepresented in Class C. Fourthly, and somewhat surprisingly, we see a noteworthy difference between the four regions. Specifically, unlike respondents living elsewhere, most respondents living in the West of the Netherlands are in class B. Moreover, respondents living in the West of the Netherlands are strongly overrepresented in class C. Finally, most respondents commuting three days or less are in Class A, while most respondents commuting more than three days are in class B. Looking at the within-class distribution, we observe that respondents commuting five days are particularly overrepresented in class C.

Altogether, a coherent picture emerges: respondents in Class A seem to be older, living in more expensive house types and commuting less frequently (possibly because of working-from-home practices). They have a comparatively high willingness-to-pay for street-level factors and a high Value-of-Travel-Time. Respondents in class B mostly live in flats, galleries, and porches and commute somewhat more often than respondents in class A. They have a comparatively lower willingness-to-pay for street-level factors and travel time savings than respondents in class A. Finally, respondents in class C are primarily young, commute four or five days a week, and live in the West of the Netherlands. They have little interest in street-level factors.

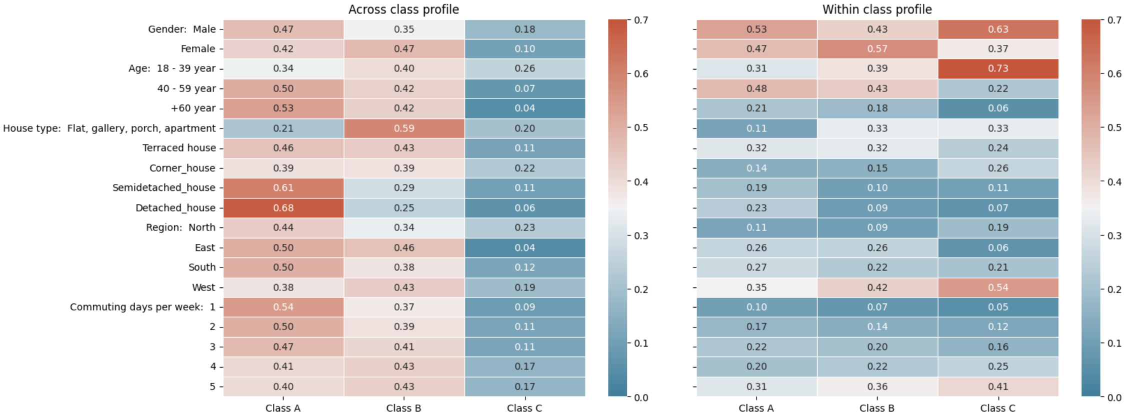


Figure 11: Across and within class profiles from latent class model

5. Conclusion and discussion

In conclusion, this research proposed in new computer vision-based choice model for modelling choice behaviour in the presence of visual and numeric stimuli. Thereby, this paper methodologically expands the realm choice models. It contributes to the recent methodological progress made in the fields of transportation and choice modelling that aims to bring machine learning and DCMs closer together (e.g. Rossetti et al. 2019; Sifringer et al. 2020; Arkoudi et al. 2021; Ramírez et al. 2021; van Cranenburgh et al. 2021). In this paper, we have demonstrated its merits by applying it to residential location behaviour, which is strongly coupled with travel demand. To do so, we have administered a novel SC experiment in which we presented to respondents trade-offs between street-level actors embedded in images and commute travel time and monthly housing costs. We have shown that using the model we can acquire novel insights into preferences over the visual environment using our proposed model. For instance, we have uncovered which residential places people find most and least attractive and what people are willing to pay to improve street-level factors in their residential area.

There are numerous avenues for further research.

Future research avenues

- Apply Explainable AI to better grasp what concepts in the street-level images are particularly generating positive of negative utilities
- Using street-level utilities in hedonic price analysis
- Clustering of street-view images in terms of the utility they generate
- Using the training CV-enriched DCM to monitor changes in the attractiveness of residential areas. For instance, to assess whether a policy aimed to improve a neighbourhood

To capitalise on the complementary information embedded in images and effectively pursue the above avenues for future research, the modelling toolbox of transport modellers needs a significant push. Our field's estimation software, survey platforms, computational resources and data handling practices are not geared towards working with large numbers of images. Developing is required not only on the hardware and software sides but also on the side of the researcher. Working with a large number of images requires a higher technical level of programming and data handling skills. These hurdles are surmountable, but in the short term, they may slow scientific progress. A lot of the required skills and good practices can readily be learned from the computer vision field. By making our code and data available, we hope to encourage research the cross-fertilisation between the transport, behavioural modelling and computer vision fields.

Acknowledgement

This work is supported by the TU Delft AI Labs programme.

References

Arkoudi, I., Azevedo, C. L. & Pereira, F. C. (2021). Combining Discrete Choice Models and Neural Networks through Embeddings: Formulation, Interpretability and Performance. *arXiv preprint arXiv:2109.12042*.

Cramer, J. S. (2005) Omitted variables and misspecified disturbances in the logit model, *Tinbergen Institute Discussion Paper*,Manuscript

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G. & Gelly, S. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dubey, A., Naik, N., Parikh, D., Raskar, R. & Hidalgo, C. A. (2016). Deep Learning the City: Quantifying Urban Perception at a Global Scale, Cham, Springer International Publishing.

Economist, T. (2022). How the pandemic has changed American homebuyers' preferences. The Economist. UK.

Garrido-Valenzuela, F., van Cranenburgh, S. & Cats, O. (2022). Enriching geospatial data with computer vision to identify urban environment determinants of social interactions. *AGILE: GIScience Series*, *3*, 72.

Giles-Corti, B., Bull, F., Knuiman, M., McCormack, G., Van Niel, K., Timperio, A., Christian, H., Foster, S., Divitini, M. & Middleton, N. (2013). The influence of urban design on neighbourhood walking following residential relocation: longitudinal results from the RESIDE study. *Social science & medicine*, *77*, 20-30.

Greene, W. H. & Hensher, D. A. (2003). A latent class model for discrete choice analysis: contrasts with mixed logit. *Transportation Research Part B: Methodological*, *37(8)*, 681-698.

Gregory, R. L. (1973). *Eye and brain: The psychology of seeing*: McGraw-Hill).

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J. & Chen, T. (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, *77*, 354-377.

Guevara, C. A. & Hess, S. (2019). A control-function approach to correct for endogeneity in discrete choice models estimated on SP-off-RP data and contrasts with an earlier FIML approach by Train & Wilson. *Transportation Research Part B: Methodological*, *123*, 224-239.

Hess, S. (2014). Latent class structures: taste heterogeneity and beyond. In (Eds.), *Handbook of choice modelling* (pp. 311-330). Edward Elgar Publishing.

Hess, S. & Palma, D. (2019). Apollo: A flexible, powerful and customisable freeware package for choice model estimation and application. *Journal of choice modelling*, *32*, 100170.

Lee, B. H. Y. & Waddell, P. (2010). Residential mobility and location choice: a nested logit model with sampling of alternatives. *Transportation*, *37(4)*, 587-601.

Lee, J. & Huang, Y. (2022). Covid-19 impact on US housing markets: evidence from spatial regression models. *Spatial Economic Analysis*, *17(3)*, 395-415.

McFadden, D. (1974). *The measurement of urban travel demand*. (Berkeley: Institute of Urban & Regional Development, University of California).

Nisbett, R. E. & Wilson, T. D. (1977). Telling More Than We Can Know - Verbal Reports on Mental Processes. *Psychological Review*, *84(3)*, 231-259.

Parady, G., Ory, D. & Walker, J. (2020). The overreliance on statistical goodness-of-fit and under-reliance on model validation in discrete choice models: A review of validation practices in the transportation academic literature. *Journal of Choice Modelling*, 100257.

Pinjari, A., Pendyala, R., Bhat, C. & Waddell, P. (2011). Modeling the choice continuum: an integrated model of residential location, auto ownership, bicycle ownership, and commute tour mode choice decisions. *Transportation*, *38(6)*, 933-958.

Pinjari, A. R., Bhat, C. R. & Hensher, D. A. (2009). Residential self-selection effects in an activity time-use behavior model. *Transportation Research Part B: Methodological*, *43(7)*, 729-748.

Pinker, S. (1990). A theory of graph comprehension. *Artificial intelligence and the future of testing*, *73*, 126.

Ramírez, T., Hurtubia, R., Lobel, H. & Rossetti, T. (2021). Measuring heterogeneous perception of urban space with massive data and machine learning: An application to safety. *Landscape and Urban Planning*, *208*, 104002.

Rossetti, T., Lobel, H., Rocco, V. & Hurtubia, R. (2019). Explaining subjective perceptions of public spaces as a function of the built environment: A massive data approach. *Landscape and Urban Planning*, *181*, 169-178.

Sifringer, B., Lurkin, V. & Alahi, A. (2020). Enhancing discrete choice models with representation learning. *Transportation Research Part B: Methodological*, *140*, 236-261.

Stuart-Fox, M., Kleinepier, T., Ligthart, D. & Blijie, B. (2022) *Wonen langs de meetlat*. Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, DEN HAAG.

Ton, D., Duives, D. C., Cats, O., Hoogendoorn-Lanser, S. & Hoogendoorn, S. P. (2019). Cycling or walking? Determinants of mode choice in the Netherlands. *Transportation research part A: policy and practice*, *123*, 7-23.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A. & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. International Conference on Machine Learning, PMLR.

Train, K. & Wilson, W. W. (2008). Estimation on stated-preference experiments constructed from revealed-preference choices. *Transportation Research Part B: Methodological*, *42(3)*, 191-203.

Train, K. E. (2003). *Discrete choice methods with simulation*. (New York: Cambridge University Press).

Van Cranenburgh, S., Chorus, C. G. & Van Wee, B. (2014). Vacation behaviour under high travel cost conditions – A stated preference of revealed preference approach. *Tourism Management*, *43(0)*, 105-118.

van Cranenburgh, S., Wang, S., Vij, A., Pereira, F. & Walker, J. (2021). Choice modelling in the age of machine learning-discussion paper. *Journal of Choice Modelling*, 100340.