Why did you predict that ?
Towards explainable artificial neural networks for travel demand analysis

Ahmad Alwosheel[a1], Sander van Cranenburgh[a], Caspar G. Chorus[a]

[a]*Transport and Logistics Group, Department of Engineering Systems and Services, Delft University of Technology*

Abstract:

Artificial Neural Networks (ANNs) are rapidly gaining popularity in transportation research in general and travel demand analysis in particular. While ANNs typically outperform conventional methods in terms of predictive performance, they suffer from limited explainability. That is, it is very difficult to assess whether or not particular predictions made by an ANN are based on intuitively reasonable relationships. As a result, it is difficult for the analyst to gain trust in ANNs. We show that often-used approaches using perturbation (sensitivity analysis) are ill-suited for understanding the inner workings of ANNs. Subsequently, we introduce to the domain of transportation an alternative method, inspired by recent progress in the field of computer vision. This method is based on a re-conceptualisation of the idea of heat maps to explain predictions of a trained ANN. To create a heat map a prediction of an ANN is propagated backward in the ANN towards the input variables, using a technique called Layer-wise Relevance Propagation (LRP). The resulting heat map shows the contribution of each input value; for example the travel time of a certain mode, for a given travel mode choice prediction. By doing this, the heat map reveals the rationale behind the prediction in a way that is understandable to humans. If the rationale makes sense to the analyst, she or he will gain trust in the prediction, and, by extension, in the trained ANN as a whole. If not, the analyst may choose to adapt or re-train the ANN or decide not to use it. We show that by reconceptualising the LRP methodology towards the travel demand analysis context, it can be put to effective use in application domains well beyond computer vision.

1. Introduction

Artificial Neural Networks (ANNs) are emerging as an indispensable tool for many applications in the field of transportation. Recent examples include modelling lane-changing behaviour of drivers (Xie, Fang, Jia, & He, 2019), predicting mode choice behaviour (Sun et al., 2018), predicting traffic flow (Polson & Sokolov, 2017), and investigating travellers' decision rules (Alwosheel, van Cranenburgh, & Chorus, 2017; van Cranenburgh & Alwosheel, 2019). This increase in ANNs' popularity in transportation research is mainly driven by the abundance of data from a variety of emerging sources (Chen, Ma, Susilo, Liu, & Wang, 2016), and the ANN's often impressive predictive performance (Goodfellow, Bengio, & Courville, 2016; Karlaftis & Vlahogianni, 2011; Mckinney et al., 2020).

Although ANNs often obtain superior prediction performance compared to their conventional, more theory-driven counterparts (e.g. discrete choice models in a travel demand analysis context) their opaque nature makes explaining individual predictions which are made by an ANN very difficult. Without sufficient understanding of how and why a model makes a particular prediction, the use of ANNs will mainly be confined to niche settings where prediction performance is highly valued (e.g., short term travel demand predictions) and model transparency is not of great importance. For

[1] a.s.alwosheel@tudelft.nl | +31152783420

justifiable reasons, governments and transport planning agencies put a higher premium on model transparency (which is considered a prerequisite for good governance), than on superior empirical prediction performance.

Recently, the development of techniques for opening up and explaining the ANN's black-box has been the subject of many research efforts in a variety of research fields (Lipton, 2016). Notably, in the computer vision field much progress has been made to shed light on the inner workings of trained ANNs (Montavon, Samek, & Müller, 2018; Samek, Wiegand, & Müller, 2017; Simonyan, Vedaldi, & Zisserman, 2013). The Layer-wise Relevance Propagation (LRP) method has emerged as one of the most popular approaches to inspect the rationale behind ANNs' predictions (Adebayo et al., 2018). The LRP method generates a so called heat map. For example, the heat map of an ANN trained to discriminate between dogs and cats based on pictures, highlights which parts of an image (e.g., pixels representing cat whiskers) were most relevant for the produced prediction (in casu: cat). The generated heat map reveals the rationale of a trained ANN and as such allows for intuitive investigation of what made the model produce a prediction; in case the rationale aligns with the mental map of the analyst, this helps to build trust in that prediction. In case the exhibited rationale does not align, this of course still offers valuable information to the analyst. In principle, LRP (and related techniques) could also be of use to analysts working in other contexts than computer vision, including transportation. But, to the best of the authors' knowledge, no studies have yet investigated the use of heat maps for transportation research in general and travel demand analysis in specific; this is possibly due to the fact that the analogy between picture classification (the original domain of LRP) and travel demand predictions is not directly obvious.

This paper re-conceptualises the use of LRP-based heat map generation and pioneers its use in a transportation (travel demand analysis) context. In particular, we show that by properly reconceptualising the notion of heat maps, they can provide meaningful explanations for predictions made by ANNs which were trained for predicting travel mode choices. As such, our paper presents a method to help analysts gain trust in ANNs' predictions in transportation contexts. Furthermore, we show that by carefully selecting predictions to analyse, the process of heat map generation can be used to build trust in the trained ANN as a whole. For the empirical part of our study, we use a recently collected Revealed Preference (RP) mode choice data dataset (Hillel, Elshafie, & Ying, 2018).

The remainder of this paper is organised as follows: Section 2 introduces the used methodology and establishes the analogy between travel mode choice modelling and image classification. Section 3 presents the dataset used for our analysis and discusses the ANN training procedure. Section 4 presents the results. It shows the heat maps created using LRP. Section 5 draws conclusions and shows directions for future research.

2. Methodology

Before delving into the LRP methodology details, it is useful to present notations and establish the analogy between image classification and discrete (travel mode) choice modelling. Numerous concepts in discrete choice modelling have a counterpart, under a different name, in machine learning. For the reader's convenience Table 1 provides a brief 'translation' table.

In discrete choice analysis the choice data consists of a set of observations $S = ((\mathbf{x}_1,\boldsymbol{y}_1),(\mathbf{x}_2,\boldsymbol{y}_2),...,(\mathbf{x}_n, \boldsymbol{y}_n),...,(\mathbf{x}_N,\boldsymbol{y}_N))$. Each n^{th} observation s_n contains a vector of independent variables $\mathbf{x}_n$ that represent the attributes and a K-dimensional vector of dependent variables $\boldsymbol{y}_n$ that represents the observed choice (i.e., zeros for the non-chosen alternatives, and one for the chosen alternative); K being the size of the choice set. Each vector $\mathbf{x}_n$ consists of I independent variables (annotated as x_i). Since choices are mutually exclusive (i.e., only one alternative can be chosen from the choice set), from a machine learning perspective this is considered a classification problem.

In image classification problems each observation contains an array of pixels of the image and a Q-dimensional vector that represents the image label; Q being the size of the fixed set of categories. For simplicity we consider the case of a greyscale image where each pixel takes a single value that represents intensity within some range (e.g., from 0 (black) to 255 (white)).[2] The task of image classification is to assign an input data to one label from the fixed set of categories. In this setting, an analogy between image classification and discrete choice modelling can be drawn, where pixels are equivalent to attributes (e.g. travel time), and intensity corresponds to the attribute value (e.g. 25 minutes). Further, similar to the observed choice set, the image label set of size Q is finite, collectively exhaustive and mutually exclusive.

Table 1: Basic image processing terminology and discrete choice modelling equivalent

Image processing	**Discrete choice modelling**
Pixel	Attribute
Intensity	Value
Label/Class	Alternative
Label set	Choice set

When using ANNs for classification, the so-called softmax function is used at the output layer to convert values (processed and forwarded by hidden layers) into probabilities. The softmax is essentially a logit function, see Appendix 1 for a brief description of the ANN methodology. Similar to discrete choice models, ANNs make predictions up to a probability. Since this study is primarily concerned with explaining predictions of ANNs by uncovering the relevance of each independent variable to a particular prediction, the values processed and forwarded to the output layer (i.e., softmax function inputs) are annotated $f(x)$ and are henceforth referred to as the *relevance*. Note that the notion of relevance in this context can loosely be conceived as utilities in a discrete choice context.

2.1. Model explainability and trust

Opening the black-box of ANNs has received much attention in a variety of fields (Hall & Gill, 2018). In the literature, several meanings have been attached to the effort of opening an ANN's black-box such as enhancing interpretability, explainability and understandability (Doshi-Velez & Kim, 2017; Lipton, 2016; Rosenfeld & Richardson, 2019). In this study, we focus on explainability, which is defined as the ability of the analyst to inspect the contribution of each input (e.g., attributes or image pixels) for a particular example to produce a prediction (Montavon et al., 2017). By explaining a model prediction, we mean presenting a numerical or visual artefact that provides a qualitative understanding of the relationship between independent variables (e.g., attributes) and the model's prediction (Ribeiro, Singh, & Guestrin, 2016). We consider the ability to explain predictions to be critical to build trust between the analyst and the trained ANN model (see further below).

For an analyst to trust a model prediction and take some actions based on it, it is essential to: 1) understand *why* the model has made this prediction (i.e., prediction explainability, henceforth called the Why part); 2) ensure that is based on 'correct', i.e. intuitive and expected relations (this is called the Domain Knowledge part). Obviously, the latter is domain dependent, and the analyst has the "final say" in this regard. For example, consider a black-box model trained to detect tumours from x-ray

[2] In other image types (e.g., RGB type), each pixels consist of several channels (e.g., red, green and blue channels).

images. For a doctor to trust a model's prediction, (s)he needs to understand on what basis or factors (e.g., which part of the x-ray) the model made that prediction (the Why part), and whether these are correct, intuitive and expected (based on the doctor's Domain Knowledge). In the remaining part of Section 2, we focus on the Why part (i.e., we present an approach that enables an analyst to answer the Why question). The Domain Knowledge part will be elaborated as part of our discussion of the results of our empirical analysis in Section 4.

To address the Why part, so-called saliency methods have emerged as a popular tool to highlight which independent variables deemed relevant or important for an ANN prediction (Adebayo et al., 2018; Kittley-Davies et al., 2019; Simonyan et al., 2013). These methods can be broadly classified into two categories: perturbation- and backpropagation-based methods (Shrikumar, Greenside, & Kundaje, 2017).

Perturbation-based methods aim to measure the effect of applying small changes in each input (or removing it) on the predictions (or probabilities) produced by the trained ANN (Zeiler & Fergus, 2013; Zintgraf, Cohen, Adel, & Welling, 2017). The underlying principle of perturbation-based methods is that the input whose change or removal affects the ANN output most is the one that has the most relative importance (Ancona, Ceolini, Öztireli, & Gross, 2017). In applications of ANNs for travel choice behaviour modelling, most efforts to answer the Why part have indeed been devoted to perturbation-based approaches. For example, several studies conducted (or suggested using) perturbation-based methods – mostly called sensitivity analysis by transportation researchers – to measure the importance of independent variables for different types of trained ANNs (Chiang, Zhang, & Zhou, 2006; Golshani, Shabanpour, Mahmoudifard, Derrible, & Mohammadian, 2018; Hagenauer & Helbich, 2017; Hensher & Ton, 2000; Lee, Derrible, & Pereira, 2018).

While the perturbation-based methods are widely used to answer the Why part, several studies have highlighted their drawbacks and explained that they are fundamentally inappropriate for this aim. The first, more practical, drawback is that these methods can be computationally inefficient as each change requires a separate forward propagation for the ANN (Shrikumar et al., 2017). This aspect of computational (in-) efficiency becomes more important as the complexity and number of parameters of ANNs grow (e.g., an early version of convolution neural network consists of over 60 million parameters (Krizhevsky, Sutskever, & Hinton, 2012)). The second, and more fundamental, drawback of perturbation-based methods is that upon close inspection, they do not actually provide an answer to the Why-question that analysts are looking for. Instead, because the process is based on alternation of independent variables' values, perturbation based methods answer a different question, being *which* independent variable needs to be altered to make the example belong more/less to the predicted alternative. In other words, perturbation-based methods measure the susceptibility of the output to changes in the input which might not necessarily coincide with those inputs on which the network based its prediction (Böhle, Eitel, Weygandt, & Ritter, 2019; Montavon et al., 2018; Shrikumar et al., 2017). This is indeed a fundamental limitation when answering the Why question. A visual illustration of this fundamental point is presented by Samek et al. (2017) where an image of rooster is correctly predicted by the model (see Fig. 1). Changing the pixels' values of the yellow flowers (that block part of the rooster) in a specific way would reconstruct the covered part of the rooster, which may result in an increase in the probability of predicting a rooster. As such, the result of this perturbation process may lead the analyst to believe that pixels that constructed the yellow flowers were important to the prediction of rooster (which is certainly not correct).

In contrast to perturbation methods, backpropagation-based methods operate by propagating the relevance (i.e., softmax function input $f(x)$) backwards from the output neuron backward through the hidden layers towards the input layer (see Appendix. 1 for an overview of ANN structure)) (Springenberg, Dosovitskiy, Brox, & Riedmiller, 2014). One of the most popular of this type of methods in the computer vision field is LRP (Bach et al., 2015). The LRP method leverages the structure of ANNs and the model parameters (i.e., weights) to determine the negative/positive contribution of each independent variable to a particular prediction. It basically asks, for each node, which of the nodes in the preceding layer contributed to what extend to the value in that node. As such, after being applied to the full network, it identifies the independent variables that were pivotal for the ANN's prediction. Thereby, LRP allows the analyst to understand *why* the model has made a particular prediction, given a set of independent variables (Samek, Montavon, Vedaldi, Hansen, & Müller, 2019). Furthermore, as these methods require a single pass to propagate the relevance from the output to the input layer, they are computationally highly efficient (Böhle et al., 2019). Colloquially put, in contrast to perturbation methods which in essence inspect choice probabilities for other than a particular observation (by changing the input variables and looking at changes in choice probabilities), the LRP method only focuses on the particular observation to be explained, studying which input values were particularly crucial for the ANN to arrive at a prediction in the context of the observation.

Before we delve into the technical details, we would like to make a clear distinction between two types of trust: 1) trusting a particular prediction made by an ANN; and 2) trusting the ANN model as a whole. In its core, the LRP method is developed for the former type, but it is worth noting that the method can be also used for the latter type of trust by applying the method to many carefully selected observations (Ribeiro et al., 2016). In this study, we show how to use the method to gain trust regarding multiple ANN predictions (to build trust in each of those predictions). Then, we show a case of how trusting multiple systematically selected ANN predictions can lead to increased levels of trust in the model as a whole (see Section 4).

Fig. 1: Rooster image (Samek et al., 2017). (For interpretation of the references to colour in this figure legend, the reader is referred to the online version of this article.)

2.2. Layer-wise Relevance Propagation method

LRP operates by propagating the activation strength of the node of interest backward, through hidden layers, to the input layer. In this study, we limit our focus on understanding the ANN prediction; hence, we are mainly concerned with back propagating the activation at the *output* nodes backwards through the hidden layers, using local propagation rules, until it allocates a relevance score R_i to each *input* variable x_i (Samek et al., 2017). Each R_i can be interpreted as the contribution an input x_i has made to a prediction (see Fig. 2). Crucially, each output node can have its own LRP-process; for example, in a travel mode choice context, the ANN assigns a probability to each mode, representing the probability, for a particular case, that the traveller chooses, for instance, the bus, train, or car. LRP can then be used for each of these probabilities, what factors were relevant for that prediction. In other words, LRP can be used to explain the choice probability, predicted by the ANN, for the bus mode, and likewise, for the train and car mode. However, in most cases the LRP method is applied to explain the highest choice probability assigned by the ANN; that is, the method explains why the ANN predicts that a particular mode has a higher probability than the others, of being chosen. In our paper, we use LRP in both ways, and we will clearly indicate when the method is used in which way.

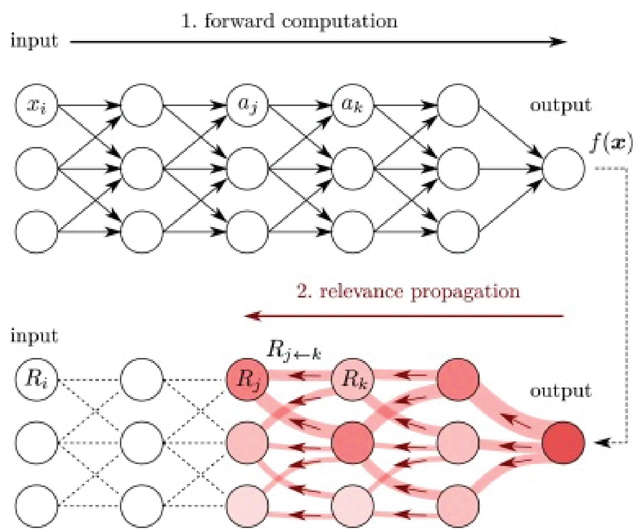


Fig. 2 : Diagram of the LRP procedure (Montavon et al., 2018). Red arrows indicate the relevance propagation flow. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)[3]

The key property of the relevance redistribution process used in LRP is that the total relevance at every layer of the ANN (from the output layer to the input) needs to be maintained; this property is known as relevance conservation and can be described as follows:

$$\sum_i R_i = \dots = \sum_j R_j = \sum_k R_k = \dots = f(\mathbf{x}) \qquad (1)$$

[3] http://www.heatmapping.org/.

where i, j and k are the indices for nodes on the layers, and R_k is the relevance of node k for the relevance $f(\mathbf{x})$. This equation highlights that the method computes the decomposition of $f(\mathbf{x})$ (most right) in terms of the input variables (most left). To ensure Equation (1) holds, two rules need to be imposed:

$$\sum_j R_{j \leftarrow k} = R_k \tag{2}$$

$$R_j = \sum_k R_{j \leftarrow k} \tag{3}$$

where $R_{j \leftarrow k}$ is defined as the share of R_k that is redistributed to node j in the lower layer (see Fig. 2). The redistribution of the relevance resembles the process of forward propagation (used to produce predictions). In forward propagation, the activation function $z(.)$ of the node k generates one output a_k that is fanned out to other neurons and can be described as follows (see Appendix. 1 for comprehensive description of ANN structure):

$$a_k = z\left(\sum_j w_{jk} a_j + w_k\right) \tag{4}$$

Where w_{jk}, w_k are the weight and bias parameter of the neuron. The main principle used by LRP to back propagate the relevance is that what has been received by a node should be redistributed to the nodes at the lower layer proportionally. In the literature, different ways in which relevance is back propagated have been proposed. Empirical studies have shown that some of these rules yield better relevance redistribution depending on many factors such as the used activation function and position of the hidden layer (i.e., the layer deepness). In this study, we use the ϵ-rule (as described in (Samek et al., 2019)), which back propagates the relevance to each neuron as follows:

$$R_j = \sum_k \left(\frac{w_{jk} * a_j}{\sum_j (w_{jk} * a_j) + \epsilon}\right) R_k \tag{5}$$

where ϵ is a fixed constant of small value ($\epsilon = 10^{-7}$) which is added to the denominator to prevent division by zero (not to be confused with the error in discrete choice models). Doing so avoids the relevance values to become too large. This equation shows that the relevance is propagated proportionally depending on: 1) the neuron activation a_j (i.e., more activated neurons receive larger share of relevance), and 2) the strength of the connection w_{jk} (more relevance flows through more strong connection). In this study, we focus only on rule shown in Equation (5), and for more detailed description of LRP and comprehensive discussion on alternative relevance redistribution rules, interested readers are referred to Samek et al. (2019) and Lapuschkin, Binder, Montavon, Müller, and Samek (2016).

2.3. Explaining a prediction using heat map - a computer vision illustration

To further clarify the method, in this subsection we provide a brief illustration of how the LRP method is commonly used in the computer vision field. This particular example is taken from Lapuschkin, Binder, Montavon, Muller, and Samek (2016) whose aim is to explain the predictions of two different machine learning models (these models themselves are not of interest to us in this paper and are not discussed in any detail here). Each of these models is trained using large number of images to discriminate between several output classes, including a horse class. A horse image is presented to the two models, see the left-hand side plot in Fig. 3. Both models produced the correct prediction with high confidence. Then, the prediction is propagated backward using the above-described explanation method (i.e., LRP) to provide an answer to the Why-question (why did the ANN believe that this is a picture of a horse). The analyst can then use the outcome of the LRP process to verify whether the model predictions are based on intuitive and expected rationales (the Domain Knowledge question). To facilitate inspection, the relevance is usually presented as a heat map, where pixels with high positive relevance are shown in red (see colour map on the right side of Fig. 3).

The middle and right-hand side plots in Fig. 3 show the heat maps generated using the LRP method, given the input: the horse image on the left-hand side. Although the predictions produced by both models are correct, the heat maps reveal that the models have a different rationale. For a horse image, we expect (as human analysts with some domain knowledge) a well-trained model to base its prediction on relevant features and distinguishable characteristics of horses, such as e.g. the horse tail. Fig. 3 shows that Model A indeed assigns a high relevance to such horse pixels, while Model B assigns a high relevance to the lower left-hand side corner of the image, where the copyright tag is located. Hence, the heat map reveals that the prediction of model B is largely based on the existence and nature of the copyright tag, rather than the part of the image where distinguishable characteristics of horse are shown. The source of this outcome is that in the training data many horse images were present with the same copyright tag. As a result, Model B has learned that the copyright tag is a good explanatory ‘variable’. This is a clear example of the fact that machine learning methods excel in detecting patterns, regardless of whether these patterns are meaningful, or not (Abu-Mostafa, Magdon-Ismail, & Lin, 2012). Most importantly for the purpose of this paper, this example illustrates that LRP can be used to inspect the model rationale and examine its trustworthiness, using human domain knowledge. In the following, we illustrate how the LRP can be recast and implemented for non-visual contexts, specifically discrete choice analysis (see the analogy between image classification and discrete choice modelling established earlier in this section).

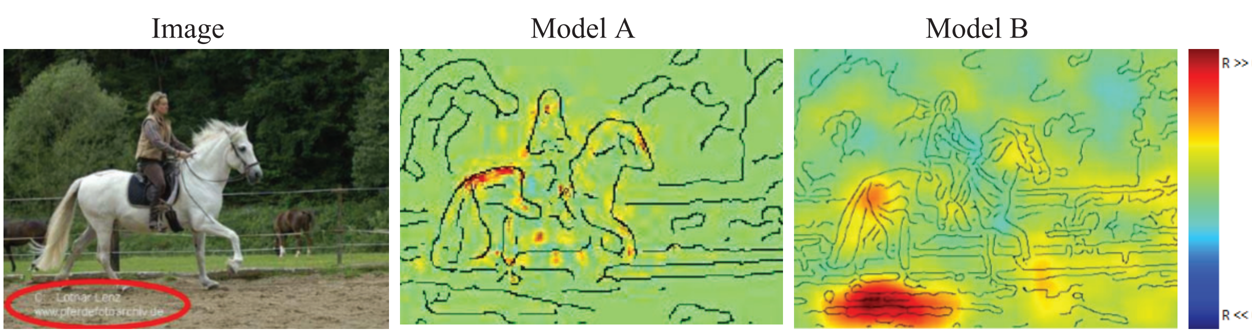


Fig. 3: Left: Image of the horse class, presented to two different models. Middle and right: relevance of each pixel is drawn as heat map. (For interpretation of the references to colour in this figure legend, the reader is referred to the online version of this article.)

2.4. Explaining a prediction in travellers' discrete choice context – A re-conceptualisation using Monte Carlo experiments

This subsection conducts a series of Monte Carlo experiments to get a feeling for how heat maps can be re-conceptualised and used in the context of discrete choice data. Table 2 shows the parametrisations of the three synthetic data sets that we generated. Each data set consists of three alternatives and two generic attributes: X_1 and X_2. Parameters have different values across data sets (we use negative, positive and neutral parameter values). Each data set consists of 10,000 hypothetical respondents, each making a single choice. Attribute levels are generated using a random number generator between zero and one. To create the synthetic choices, the total utility of each alternative is computed and the highest utility alternative is assumed to be chosen, following a Logit (RUM-MNL) model where the random part of utility is distributed Extreme Value type I with variance $\pi^2/6$.

Table 2: Synthetic data specification and parametrisation

Dataset no.	Model specification	Parametrisation	Cross-entropy (RUM-MNL)	ρ^2 (RUM-MNL)	Cross-entropy (ANN)	ρ^2 (ANN)
A1	$V_{un} = \sum_m \beta_m x_{umn}$	$\beta_1 =- 6$ $\beta_2 =- 4$	-0.53	0.51	-0.54	0.50
A2		$\beta_1 =+ 6$ $\beta_2 =+ 4$	-0.53	0.51	-0.54	0.50
A3		$\beta_1 =- 6$ $\beta_2 = 0$	-0.59	0.45	-0.61	0.44

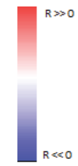
For each data set, a three-layers ANN with 4 hidden nodes on the hidden layers is trained. As has also been found in previous studies (e.g., Alwosheel, van Cranenburgh, and Chorus (2018)), the ANNs are able to learn the a RUM-MNL data generating process with high accuracy, in the sense that the prediction performance of the ANN almost matches that of the true underlying data generating process encoded in a corresponding discrete choice model, see Table 2.

For the first data set (A1), the negative sign of the parameters imposes a dislike for higher attribute values (i.e., the lower the attribute values, the more attractive the alternative becomes). Hence, the attribute values of the *chosen* alternative are expected to contribute negatively to the choice probability prediction for that alternative (as reducing the attribute values would increase the attractiveness of the chosen alternative). In contrast, we expect that high attribute values of the *non-chosen* alternatives contribute positively to the prediction, implying that the attractiveness of these non-chosen alternatives increases as these attribute values increase. These expectations are confirmed in Table 3, where we see the relevance of the attribute values that are computed using the LRP method[4], alongside the choice probabilities predicted by the ANN, for three randomly selected observations from the synthetic data. In this Table, we apply the LRP method to explain the choice probability assigned to the chosen alternative – that is, we do not explain choice probabilities assigned to non-chosen alternatives. In the heat map positive relevance values are depicted red; negative relevance values are depicted blue, and neutral relevance values are depicted white. The colour intensity for each observation is normalised to the maximum absolute value.

Table 3: Results of observations randomly selected from A1 data set

[4] To generate heat maps, the LRP method is used and implemented in the Python environment using the open source library iNNvestigate (Alber et al., 2019).

	Attribute Values						Relevance						True chosen alternative			ANN prob		
	X_1			X_2			X_1			X_2								
	Alt 1	Alt2	Alt3	Alt 1	Alt2	Alt3	Alt 1	Alt2	Alt3	Alt 1	Alt2	Alt3	Alt 1	Alt2	Alt3	Alt 1	Alt2	Alt3
Obs. 1	0.127	0.8887	0.916	0.038	0.871	0.742							1	0	0	0.99	0	0.01
Obs. 2	0.95	0.004	0.97	0.725	0.133	0.75							0	1	0	0.01	0.99	0
Obs. 3	0.936	0.957	0.045	0.882	0.866	0.157							0	0	1	0.01	0	0.99

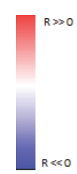


Consider the three observations shown in Table 3, where alternative 1 is chosen in the first observation, alternative 2 is chosen in the second observation, and alternative 3 is chosen in the third observation. Note that the ANN predictions are correct with very high confidence as shown by predicted choice probabilities of 0.99 for the chosen alternative in each of the three observations. The blue diagonal values show that the attribute values of the chosen alterative have contributed negatively toward the predicted probability of the alternative being chosen. In contrast, the off-diagonal cells, which here are associated with the non-chosen alternatives, are coloured red. This means that the attribute values of these unattractive alternatives, which are comparatively high, positively contribute to the prediction that alternative 1 is chosen in observation 1, alternative 2 in observation 2, etc. Hence, increasing the attribute values of the non-chosen alternatives would in this situation further increase the probabilities of the chosen alternatives, which is exactly as expected.

Compared to the first data set, in the second data set (A2) the parameters have flipped signs. Hence, lower attribute values are more attractive than higher ones. Table 4 shows the results for three randomly selected observations (again, from the subset of observations that are correctly predicted by the ANN). We use same colour map and intensity as in Table 3. As can be seen, Table 4 reveals the same patterns as shown in Table 3, but colours are flipped, i.e., cells on the diagonal are red, and cells off the diagonal are blue. This is fully in line with expectations, as here an increase (decrease) in the attribute levels of the chosen (non-chosen) alternative positively contributes to the choice probability that is predicted for the chosen alternative.

Table 4: Results of observations randomly selected from A2 data set

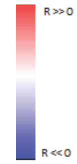
	Attribute Values						Relevance						True chosen alternative			ANN prob		
	X_1			X_2			X_1			X_2								
	Alt 1	Alt2	Alt3	Alt 1	Alt2	Alt3	Alt 1	Alt2	Alt3	Alt 1	Alt2	Alt3	Alt 1	Alt2	Alt3	Alt 1	Alt2	Alt3
Obs. 1	0.97	0.19	0.06	0.95	0.07	0.107							1	0	0	0.99	0.01	0
Obs. 2	0.108	0.98	0.0594	0.063	0.865	0.35							0	1	0	0.01	0.99	0
Obs. 3	0.025	0.05	0.83	0.32	0.305	0.99							0	0	1	0.01	0	0.99



Lastly, Table 5 presents the results for data set A3. Again, three randomly selected observations from the subset of observations that are correctly assigned by the ANN are shown. In this data set, β_2 is zero. This means that the attribute X_2 does not impact the decision makers' choices. As such, we expect the relevancies for these attribute values to have values that are close to zero. In line with expectation, Table 5 shows that all cells for X_2 are (almost) white – meaning that the values of this attribute do neither positively or negatively contribute to the predicted choice probabilities.

Table 5: Results of observations randomly selected from A3 data set

	Attribute Values						Relevance						True chosen alternative			ANN prob		
	X_1			X_2			X1			X2								
	Alt 1	Alt2	Alt3	Alt 1	Alt2	Alt3	Alt 1	Alt2	Alt3	Alt 1	Alt2	Alt3	Alt 1	Alt2	Alt3	Alt 1	Alt2	Alt3
Obs. 1	0.05	0.665	0.979	0.99	0.001	0.757							1	0	0	0.98	0.1	0.1
Obs. 2	0.97	0.05	0.99	0.323	0.385	0.329							0	1	0	0.01	0.99	0
Obs. 3	0.98	0.9	0.00038	0.017	0.154	0.94							0	0	1	0	0.01	0.99



In sum, this application on synthetic data provides a first idea of how the LRP method can be used to inspect the rationale based on which an ANN makes it predictions in a travel mode choice context, and it provides a first sign of face validity of the method. The next sections present an application of the method on a real empirical data set.

3. Empirical data and ANN training

3.1. Data preparation

For this study, we use revealed preference (RP) data from a study conducted for travel mode choice analysis in London city (Hillel, Elshafie, & Jin, 2018).[5] This dataset contains of four alternatives, and a total of 27 features (i.e., attributes of alternatives and characteristics of decision makers). Three processing steps have been executed to prepare the data for this study: First, features that were considered redundant are removed, or merged with others. For instance, rather than using three features to represent car cost (fuel, congestion, and total cost), we merged them into a single one representing the total car cost. Table 6 shows statistics on the attribute levels in the dataset used for this analysis. Second, we noticed that the dataset is highly imbalanced in terms of the chosen mode: walking (17.6%), cycling (3.0%), public transport (35.3%) and driving (44.2%). Such imbalances could affect the reliability of the trained ANNs (Haykin, 2009). As this paper is concerned with explaining ANN predictions (i.e., we do not aim to find the best ANN to predict the mode choices), we considered dealing with this sort of data imbalances out of scope for this paper. Therefore, the data imbalance is 'repaired' by eliminating the cycling alternative from the dataset. Third, we excluded very short trips (i.e., less than two minutes), as these were deemed not to contain a mode trade-off. The resulting dataset that is used for this study consists of 77,638 mode choice observations.

Table 6: Data statistics

No.	Attribute	Description	Type	Range [min, max]	Mean and standard deviation
1	TC_{Drive}	Estimated cost of driving route, including fuel cost and congestion charge	Float (£)	[0.05, 17.16]	(1.91, 3.48)
2	TC_{PubTr}	Estimated cost of public transport route, accounting for rail and bus fare types	Float (£)	[0, 13.49]	(1.56, 1.55)
3	TT_{Drive}	Predicted duration of driving route	Float (minutes)	[0.02, 142]	(17, 15)
4	TT_{PubTr}	Predicted duration of public transportation	Float (minutes)	[0.02, 141]	(28, 19)

[5] The dataset and its description are available online, and can be downloaded from the first author profile at researchgate.net

5	TT_{Walk}	Predicted total duration of walking times for interchanges on public transport route	Float (minutes)	[0.02, 550]	(69, 68)
6	DIS	Straight line distance between origin and destination	Integer (meters)	[96, 40,941]	(4,690, 4,827)
7	$TRAF$	Predicted traffic variability on driving route	Float	[0, 1]	(0.34, 0.20)
8	$INTER$	Number of interchanges on public transport route from directions API	Integer	[0, 4]	(0.38, 0.62)
9	DL	Boolean identifier of a person making trip: 1 if person has driving license, 0 otherwise	Bool	[0, 1]	(0.62, 0.49)
10	CO	Car ownership of household person belongs to: no cars in household (0), less than one car per adult (1), one or more cars per adult (2)	Integer	[0, 2]	(0.99, 0.75)
11	FEM	Boolean identifier of a person making trip: 1 if female, 0 otherwise	Bool	[0, 1]	(0.53, 0.49)
12	AG	Age of person making trip	Integer (years)	[5, 99]	(39.5, 19.3)

3.2. ANN development and training

The ANN is implemented in a Python environment, using the open source deep learning library Keras (Chollet, 2015). To train the ANN, the built-in training algorithm (which is used to update weights' values $\mathbf{w}$) known as Adam is used (Kingma & Ba, 2014). Prior to training the ANN, the data are normalised to reduce training time and minimise the likelihood of ending-up with suboptimal local solutions.[6] A conventional three layers (input, output and one hidden layer consist of ten nodes, see Appendix 1 for a similar ANN layout) fully connected ANN structure is used. Unlike the traditional three-layers ANN, we have removed the bias nodes in the hidden and output layer to avoid losing fraction of the relevance values.[7] Note that the bias nodes removal has not impacted the prediction performance of the trained ANN. To train the ANN and test its performance to predict the travel mode choice, we conducted a so-called *k*-fold cross-validation method, with $k = 5$. The data set is randomly partitioned into five equally sized folds of (roughly) 15,528 observations. Then, a single fold is used for testing, while the remaining four folds are used for training. This process is repeated 5 times, where each of the five folds is used only once for testing.

Table 7 shows several performance metrics for the trained ANN. The reported performance metrics are averaged across the five hold-out folds. It shows that ANN achieves a satisfactory prediction performance. For comparison, we also report the performance of a standard linear-additive RUM-MNL model (see

[6] Data normalisation is a common practice for ANN training. In this study, the minimum and maximum values of data are normalised to 0 to +1.

[7] ANN complexity is adjusted using a cross validation approach (see (Alwosheel et al., 2018) for more details). To avoid overfitting, a commonly used rule-of-thumb in ANNs is that the sample size needs to be (at least) 10 times larger than the number of adjustable parameters in the network (Haykin, 2009). A recent study specifically dealing with sample size requirements for using ANNs in the context of choice models is more conservative, and recommends to use a sample size of (at least) 50 times the number of estimable weights (Alwosheel et al., 2018). Our sample size satisfies this condition and, therefore, we safely avoid overfitting issues.

Appendix 2 for the model specifications).[8] As expected based on previous literature, the trained ANN outperforms the discrete choice model by a large margin. Table 8 shows the *k*-folds confusion matrix for the trained ANN. To construct the confusion matrix each observation is assigned to an alternative based on the highest probability as predicted by the ANN. Then, each prediction is compared to the true chosen alternative. The cells on the diagonal show the mean percentage of the observations that are correctly assigned, across the 5 folds. Additionally, the values between parentheses show the average ANN probabilities of the observations that are correctly classified. The off-diagonal cells show the mean percentage of observations that are misclassified, across the 5 folds. Values between parentheses show the average ANN probabilities of these observations.

Table 7: Performance of the trained ANN

Performance metric	Function	Null model	ANN	Linear-additive RUM
Final Log-likelihood	$\sum_{n=1}^{N}\sum_{k=1}^{K} y_{nk}\ln(P_{nk})$	-86,625	-43,477	-50,704
Cross-entropy	$\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} y_{nk}\ln(P_{nk})$	-1.09	-0.56	-0.65
ρ^2	$\rho = 1 - \frac{LL(\hat{\beta})}{LL(0)}$	0	0.50	0.42

Table 8: Confusion matrix

		ANN Classification			
		Driving	Public Transport	Walking	$\sum$
True chosen alternative	Driving	82.55 (0.69)	11.23 (0.19)	6.22 (0.10)	100% (1)
	Public Transport	21.22 (0.25)	72.66 (0.66)	6.12 (0.09)	100% (1)

[8] Note that the *k*-fold method is not used for the RUM model. Rather, the RUM model is estimated one time using the whole dataset.

	Walking	23.72 (0.25)	11.56 (0.17)	64.72 (0.57)	100% (1)

4. Applying the LRP method

4.1. ANN prediction explanation of randomly selected observations

In this subsection, we use the LRP-generated heat maps for multiple observations with the aim to understand *why* the ANN produces certain mode choice predictions. Tables 9 to 11 show the back-propagated relevance extracted for three observations randomly selected from the subset of observations that are correctly assigned by the ANN. It can be seen, that predictions are made with different levels of confidence. In the context of our analyses, a high confidence level means the network assigns a choice probability of more than 0.80 to one the modes, and a low confidence level means that the highest (across travel mode alternatives in the context of a particular observation) predicted choice probability is still below 0.40. As such, for diversification purposes and to build trust in the model as a whole, the three observations are randomly selected as follows: two predictions with high confidence levels and one prediction with low confidence level. Tables 9 to 11 show the ANN probabilities, the attributes' values, and relevancies obtained using LRP for the selected observations. A heat map (using same colour map as in section 2.4 i.e., positive, negative and neutral values are depicted in red, blue and white, respectively) is employed to visualise the relevancies. As in the Monte Carlo analysis, we apply the LRP to the output node with the highest (choice) probability.

Table 9: Results of observation 1 (index: 47,489 and true chosen alternative is Drive)

	Alternatives' Characteristics						Other Characteristics		
	Alt.1: Drive	Alt.2: PubTr	Alt.3: Walk	Alt.1: Drive	Alt.2: PubTr	Alt.3: Walk			
	Attribute value			Relevance			Attribute	Value	Relevance
TT (min)	23.48	137	160				AG	47	
TC (£)	1.58	7.50					FEM	0	
TRAF	0.03						DL	1	
INTER		4					CO	2	
ANN probs	0.99	0	0.01				DIS	9,556	

Table 10: Results of observation 2 (index: 24,618 and true chosen alternative is Walk).

	Alternatives' Characteristics						Other Characteristics		
	Alt.1: Drive	Alt.2: PubTr	Alt.3: Walk	Alt.1: Drive	Alt.2: PubTr	Alt.3: Walk			
	Attribute value			Relevance			Attribute	Value	Relevance
TT (min)	7.38	4	6				AG	26	
TC (£)	10.70	2.40					FEM	1	
TRAF	0.31						DL	1	
INTER		0					CO	0	
ANN probs	0	0.01	0.99				DIS	368	

Table 9 shows an observation of a middle-aged female, who holds a driver license and owns two cars, who chose the driving alternative, which indeed seems to be the most attractive travel alternative in this case as is the fasted and cheapest alternative. In line with intuition, the ANN predicts a choice for the driving alternative with a very high level of confidence (assigning a 0.99 choice probability to that alternative). The relevance values show that car travel time receives a strong negative relevance, as expected (given that lower travel times are preferred). The relatively long travel times offered by the non-chosen alternatives receive a strong positive relevance, as expected (given that the high travel times of these alternatives makes driving alternative more appealing). Furthermore, the number of owned cars (two) and the driving license availability have a positive relevance. Together, these analyses reveal on what basis the ANN model has predicted that this traveller would choose the driving alternative. From a travel behaviour perspective, all these points are in line with expectations. As such, the analyst equipped with the proper domain knowledge can safely trust this prediction.

Moving forward to Table 10, for this observation, a young female traveller chose the walking alternative. As before, the travel time and cost of the non-chosen alternatives and the relatively high traffic on the driving route have high positive relevance for the predicted choice probability for walking. Furthermore, we see that the travel time of the chosen alternative, and the travelled distance have negative relevance values. All these relations are expected from a travel behaviour perspective; hence, this prediction too can be safely trusted. Lastly, in the Table 11 the alternative with highest predicted probability is walking; however, this mode receives a predicted probability which is only one percentage point higher than that of the other mode, implying that the ANN has low confidence in this prediction. The relevance values show that attribute values with negative relevance for the predicted choice probability for the walking alternative, are the relatively long distance and walking travel time, suggesting that shorter distance and less walking time would have made the walking alternative more attractive. This is expected from a travel behaviour perspective. Further, it can be noticed that the red and blue colours associated in this heat map are less bright, meaning that the ANN is less outspoken about what determined its prediction; this too is to be expected, given that the ANN assigns almost equal choice probabilities to each of the three alternatives.

Table 11: Results of observation 3 (index: 1,621 and true chosen alternative is Walk).

	Alternatives' Characteristics						Other Characteristics		
	Alt.1: Drive	Alt.2: PubTr	Alt.3: Walk	Alt.1: Drive	Alt.2: PubTr	Alt.3: Walk			
	Attribute value			Relevance			Attribute	Value	Relevance
TT (min)	7.23	22	24				AG	15	
TC (£)	0.30	0.00					FEM	1	
TRAF	0.44						DL	0	
INTER		0					CO	1	
ANN probs	0.33	0.33	0.34				DIS	1,271	

4.2. Using RUM-MNL and ANN models to guide observation selection

In this subsection, we use RUM-MNL (a highly robust and well understood model for mode choice analysis) and ANN predictions jointly to guide the process of observations selection, instead of relying on ANN predictions only, as in the previous subsection. This allows us to examine the overall workings and rationale of the trained ANN and decide whether we can trust a trained ANN as a whole. Furthermore, we in this section analyse correct as well as incorrect predictions. Note that while explaining correct predictions made by the black-box is important (as shown in previous subsection), it could even be more important to inspect why an ANN makes wrong predictions. This helps to obtain a higher level understanding on the model. Specifically, we are interested to learn whether, or not, these wrong predictions are based on meaningful intuition, or on counter intuitive or flawed relations learned by the ANN. It goes without saying that a mis-prediction by an trained ANN does not necessarily mean it has learned counterintuitive or flawed relations. But, if the ANN has learned such relations they are particularly to show up in mis-predictions.

To select a diverse set of observations (including observations where the trained ANN makes a wrong prediction), we distinguish between three cases, see Table 12. For each case, we randomly sample one or a few observations to analyse using LRP-generated heat maps.

- Case I: The ANN model predicts the true chosen alternative, while the RUM-MNL model makes the wrong prediction. For this case, we randomly select two observations: one for ANN prediction with high probability, and the other for ANN prediction with low probability (see Tables 13 &14).
- Case II: The RUM-MNL model predicts the true chosen alternative, while the ANN misses it. For this case, we randomly select one observation (see Table 15). As explained in 2.2, relevance back-propagation using LRP method can be implemented for any node at the network. In addition to having the relevance for the predicted alternative, we for this case also compute the relevance for the true chosen alternative (that the ANN misses), which allows for additional examination of the black-box rationale.
- Case III: Both the ANN and the RUM-MNL model mispredicts the correct alternative. Under this category, there are two subcases: ANN and RUM model agree (e.g., an observation where both models predict Driving alternative, and the true chosen alternative is Walking), or ANN and RUM disagree (e.g., an observation where the true chosen alternative is Walking, ANN prediction is Public Transport, RUM prediction is Driving alternative). One observation is selected for each subcase. See Tables 16 &17 for the selected observations details.

Table 12: Cases developed using ANN and RUM-MNL prediction. Values between parenthesis are the total number of observations for each case.

		RUM-MNL prediction	
		Correct	Incorrect
ANN prediction	Correct		Case I (6,679)
	Incorrect	Case II (3,588)	Case III (14,792)

Table 13: Results of first observation in Case I (index: 7,011 and true chosen alternative is Walking).

	Alternatives' Characteristics						Other Characteristics		
	Alt.1: Drive	Alt.2: PubTr	Alt.3: Walk	Alt.1: Drive	Alt.2: PubTr	Alt.3: Walk			
	Attribute value			Relevance			Attribute	Value	Relevance
TT (min)	5.32	5	7				AG	18	
TC (£)	0.22	1.50					FEM	1	
TRAF	0.02						DL	0	
INTER		0					CO	2	
ANN probs	0.06	0.01	0.93				DIS	403	
RUM probs	0.47	0.14	0.39						

R>>0 … R<<0

Table 14: Results of second observation in Case I (index: 38,475 and true chosen alternative is Driving).

	Alternatives' Characteristics						Other Characteristics		
	Alt.1: Drive	Alt.2: PubTr	Alt.3: Walk	Alt.1: Drive	Alt.2: PubTr	Alt.3: Walk			
	Attribute value			Relevance			Attribute	Value	Relevance
TT (min)	3.92	9	22				AG	18	
TC (£)	0.21	0.00					FEM	0	
TRAF	0.14						DL	0	
INTER		0					CO	1	
ANN probs	0.35	0.33	0.32				DIS	1,158	
RUM probs	0.3	0.43	0.29						

R>>0 … R<<0

Tables 13 &14 show (for Case I observations) the attribute values, the relevancies produced using LRP method, ANN and RUM-MNL probabilities. Table 13 shows a young female, who owns two cars and chose the walking alternative. For the chosen alternative, the attribute values with negative relevance are the distance and walking travel time. Further, the travel time values of the non-chosen alternatives have resulted in positive relevance values. Both of these points are in line with expectations and have been also noted from the observation shown in Table 11 (where walking alternative was also predicted). For the observation shown in Table 14, the relatively long distance, and longer travel time offered by public transport and walking alternatives have a positive relevance for the prediction of driving, as we expect. Further, also in line with expectations, travel time and cost of the chosen alternative have a negative relevance to the prediction.

It is possible to end up with relevance values that are unexpected or hard to rationalise. For instance, we expect owning a car to have a positive relevance for the choice probability predicted for the driving alternative, but that is not always the case, as the observation shown in Table 14 reveals. Also, the unavailability of driving license only has a negligible contribution to the driving prediction. It should be kept in mind here that, since the ANN itself is a probabilistic technique that is not expected to fit complex data perfectly (Abu-Mostafa et al., 2012), we should not expect the relevancies values that are produced by LRP to always provide a fully accurate description of the contribution of all independent variables on every observation. As such, we advise the analyst to tolerate having some unexpected relevancies' values, but of course (s)he has the 'final say' on deciding to what extent these unexpected relevancies are tolerable or not – leading to a rejection of the trained network in the latter case. For this particular prediction (shown in Table 14), we believe having two unexplainable values (out of 11) is acceptable, given that all other values are in line with expectations based on domain knowledge.[9]

Table 15: Results of Case II observation (index: 32,923 and true chosen alternative is Public Transport).

	Alternatives' Characteristics									Other Characteristics			
	Alt.1: Drive	Alt.2: PubTr	Alt.3: Walk	Alt.1: Drive	Alt.2: PubTr	Alt.3: Walk	Alt.1: Drive	Alt.2: PubTr	Alt.3: Walk				
	Attribute value			Relevance			Relevance from true alt			Attribute	Value	Relevance	Relevance from true alt
TT (min)	58.98	61	206							AG	5		
TC (£)	13.21	0.00								FEM	1		
TRAF	0.71									DL	0		
INTER		2								CO	2		
ANN probs	0.59	0.4	0.01							DIS	14,764		
RUM probs	0.01	0.98	0.01										

Two back-propagated relevancies are extracted for Case II observation, where RUM-MNL model predicts the true chosen alternative (public transport), and the ANN incorrectly predicts that the driving alternative has the highest choice probability. For a deeper examination of the trained ANN rationale, Table 15 shows the relevance values back-propagated from two output nodes: one from the public transport alternative (the true chosen alternative), and another from the driving alternative (which was predicted by ANN to have the highest choice probability). By doing so, we can obtain a higher level of trust in the model, as we are now able to inspect the model reasoning in the case of "mis-prediction" (i.e., we may come to understand what has led the model to mis-predict, and whether this is still based on an intuitive and defensible rationale). For instance, the relevancies of driving alternative shows that relatively high travel times of the other two alternatives have led to the driving prediction, which is to be expected. Further, the high travel cost and time of the predicted alternative have negative contributions, which is also in line with domain knowledge. Inspecting the relevance extracted from the true chosen alternative (public transport), we observe that the long travel time and the high number of owned cars have contributed negatively to the choice probability assigned to the public transport alternative, highlighting that the probability of choosing public transport would have been higher if these values are lowered, which is as expected.

Table 16: Results of Case III first observation (index: 48,999 and true chosen alternative is Walking).

	Alternatives' Characteristics									Other Characteristics			
	Alt.1: Drive	Alt.2: PubTr	Alt.3: Walk	Alt.1: Drive	Alt.2: PubTr	Alt.3: Walk	Alt.1: Drive	Alt.2: PubTr	Alt.3: Walk				
	Attribute value			Relevance			Relevance from true alt			Attribute	Value	Relevance	Relevance from true alt
TT (min)	11.60	38	59							AG	72		
TC (£)	0.61	3.00								FEM	0		
TRAF	0.10									DL	1		
INTER		1								CO	2		
ANN probs	0.99	0	0.01							DIS	2,861		
RUM probs	0.98	0.02	0										

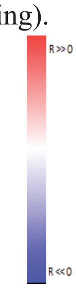


Table 17: Results of Case III second observation (index: 46,900 and true chosen alternative is Walking).

[9] A similar pattern can be also noticed in the computer vision example (presented in subsection 2.3). For instance, we can identify small red spots on non-horse pixels (i.e., pixels showing the background, see Fig. 3).

	Alternatives' Characteristics									Other Characteristics			
	Alt.1: Drive	Alt.2: PubTr	Alt.3: Walk	Alt.1: Drive	Alt.2: PubTr	Alt.3: Walk	Alt.1: Drive	Alt.2: PubTr	Alt.3: Walk				
	Attribute value			Relevance			Relevance from true alt			Attribute	Value	Relevance	Relevance from true alt
TT (min)	14.38	9	38							AG	34		
TC (£)	0.56	2.40								FEM	0		
TRAF	0.57									DL	0		
INTER		0								CO	0		
ANN probs	0.49	0.42	0.09							DIS	3,272		
RUM probs	0.24	0.61	0.15										

Finally, two back-propagated relevancies are presented for the two scenarios of Case III (Table 16 shows when ANN and RUM models agree, and Table 17 when the two models disagree). In Table 16 (ANN and RUM have predicted driving alternative with high probability), the chosen alternative is walking, despite that the travelled distance is relatively long (around 3km). This indicates that in this case, the actual choice for the walking alternative deviates from expectations regarding the length of the average walking trip. The produced relevancies for the ANN are actually as expected. For example, the relevance computed for from the actually chosen alternative (walking) shows that walking travel time and distance have the highest negative relevance, indicating that the walking probability would have been higher if lower walking travel time and distance were lower – this makes sense. A similar point can be also noted for relevancies shown in Table 17 (when RUM and ANN disagree). The true chosen alternative is walking, despite the relatively long travel time and distance. The back-propagated relevancies of these two attributes from the walking node are negative, explaining the reasons for this mis-prediction which turn out to be in line with common sense and domain knowledge.

5. Conclusions and recommendations

This study re-conceptualises the use of heat maps, generated using a Layer-wise Relevance Propagation method, to explain predictions of Artificial Neural Networks in the context of travel demand analysis. We show how heat maps can be applied to provide explanation for the predictions of a trained ANN, thereby helping an analyst to build trust in predictions made by ANNs. Furthermore, we show that by carefully selecting a set of observations, this method can ultimately help building trust in a trained ANN as a whole (or not, in which the ANN can be retrained or adapted).

We would like to point out several limitations to this study, providing potential directions for future research. Firstly, to generate heat maps, this study implemented the widely used ϵ-LRP rule. Several alternative variations to this rule have been proposed in the literature, and some are found to provide better outcomes in specific domains (e.g., natural language processing). Investigating the performance of these alternative rules in the context of transport applications is a fruitful direction for further research. Possibly, this could lead to the discovery of new rules that particularly work well for transportations contexts. Secondly, additional to explaining ANN predictions, it is possible to use the LRP technique in hidden nodes to reveal what concepts and features have been learned by the trained ANN (several researches (e.g., (Olah et al., 2018)) have investigated this in computer vision context). We believe doing so in travel choice behaviour context may provide a deeper understanding of the workings of ANNs and perhaps of the decision-making processes of travellers. Thirdly, the empirical analyses provided in this paper are based on a single empirical data set and a Monte Carlo analysis on synthetic data. It is advisable to repeat these analyses on more data sets with different characteristics (e.g., more and different attributes and alternatives). This will provide a richer view on the generalisability of the proposed method to explain ANNs' predictions. Lastly, to build trust in the ANN model as a whole, predictions of RUM-MNL and ANN models are used to guide the observation selection process. Although we believe this process is very helpful to select diverse observations, it might be rewarding to develop alternative selection strategies that may result in a better representation of the data set.

To conclude, while our analysis suggests that the proposed LRP-based heat map methodology provides a valuable tool to understand the rationale behind ANN predictions in the context of travel choice behaviour, it is important to acknowledge that the proposed method does not completely solve the ANNs' black-box puzzle as it will never completely explain the inner workings of the network. As such, in our view and despite ongoing advances in explainable ANNs, their most natural domain of use in transportation still remains forecasting applications, where complete model transparency is not a prerequisite.

Acknowledgment

The first author would like to thank King Abdulaziz City for Science and Technology (KACST) for supporting this work. The third author would like to acknowledge funding from ERC, consolidator grant BEHAVE – 724431.

Appedix.1. Artificial Neural Networks – An overview

ANNs consist of highly interconnected processing elements, called neurons, which communicate together to perform a learning task, such as classification, based on a set of observations. Fig. A1 shows the layout of the neuron structure.

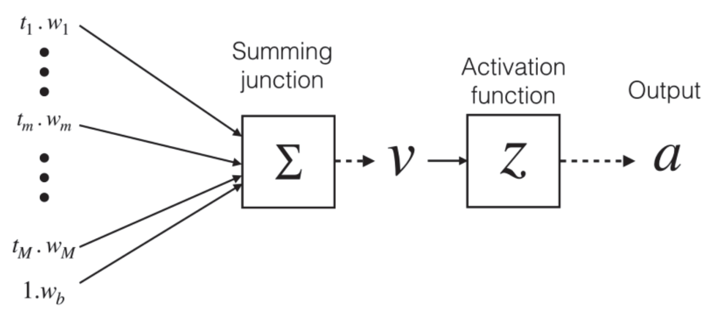


Fig. A1 : A neuron layout

Each neuron in the network receives inputs (t_m) multiplied by estimable parameters known as weights (w_m). The weighted inputs are accumulated and added to a constant (called bias, denoted b) to form a single input v for a pre-defined processing function known as activation function $z(.)$. The bias has the effect of increasing or decreasing the net input of the activation function by a constant value, which increases the ANNs flexibility (Haykin, 2009). The activation function $z(.)$ generates one output a that is fanned out to other neurons. The output a can be described as follows:

$a = z(v) = z(\sum_{m=1}^{M} w_m * t_m + w_b)$, where w_b is the weight associated with the bias.

The neurons are connected together to form a network (Bishop, 2006; LeCun, Bengio, & Hinton, 2015). A widely used ANN structure consists of layers of neurons connected successively, known as multi-layer perceptron (MLP) structure. Typically, the first (input) layer and the output layer depend on the problem at hand. More specifically, input layer neurons represent the independent variables. In the context of choice modelling, these are the alternatives' attributes, characteristics of decision-makers, and contextual factors. The output layer, in a discrete choice context, consists of neurons that provide choice probabilities P for each alternative. Layers in-between are called hidden layers because their inputs and outputs are connected to other neurons and are therefore 'invisible' to the analyst. For illustrative purposes, consider the following hypothetical situation: a person can travel using one of three modes: bus, train, or car; two attributes (travel cost "TC" and travel time "TT") are associated with each alternative. Fig. A2 shows this typical choice situation in a three-layer MLP network with four hidden neurons.

Neurons at the hidden and output layers are represented by circles in Fig. A2, while input and bias neurons are represented by squares. This is to emphasise that the neurons at the hidden and output layers are processing units, meaning that they receive inputs t and return outputs a according to predefined activation function $z(.)$, as illustrated in Fig. A1. Input neurons pass the input signals to the next layer. In Fig. A2, the ANN has a total of 7 processing units.

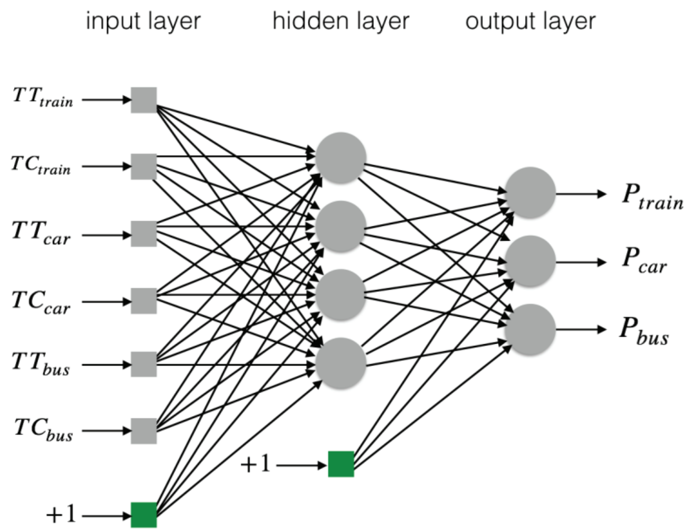


Fig. A2 : Three-layers Artificial Neural Network

For a complete MLP structure, three elements need to be defined:

1) Number of hidden layers: a commonly used structure is three-layers MLP: input, output and one hidden layer. A key property of this structure lies in the ability to approximate, with arbitrary level of precision, any measurable function given that a sufficient number of processing neurons are available at the hidden layer; this property is known as the Universal Approximation Theorem (UAT) (Cybenko, 1989; Hornik, Stinchcombe, & White, 1989).

2) Number of neurons for the hidden layer(s): the UAT holds true only if a sufficient number of hidden neurons are available. Intuitively, ANNs with more hidden neurons have more free parameters (w) and are therefore capable of learning more complex functions.

3) Activation function $z(.)$: As mentioned before, each neuron processes its input via a pre-defined activation function. Neurons at the same layer usually employ identical functions. In the analyses presented in the remainder of this paper, a tangent sigmoidal function has been employed at the hidden layer neurons, as it has been shown to lead to fast training times (LeCun, Bottou, Orr, & Müller, 2012). For the output layer, a so-called softmax function is used (which is essentially a logit) to ensure that the sum of the choice probabilities equals one.

Appedix.2. Specifications of the linear-additive random utility maximisation model

Table A1 shows the estimation results of the linear-additive random utility maximisation (RUM) model used in this study (see Table 1 for attributes' name, notation, and description). The model is estimated in Multinomial Logit (MNL) form.. As can be seen, and as is expected, all parameters have the intuitively correct sign and are highly significantly different from zero. Table A2 shows the observed utility function for RUM model.

Table A1: Estimation results for RUM-MNL model

No. of observations	77,638	
Final LL	-50,716.29	
ρ^2	0.41	
Attribute	*Est.*	*Rob. t-values*
ASC_Drive	0	fixed
ASC_PubTr	1.81	59.60
ASC_Walk	2.64	74.47
TT	-6.10	-95.31
TC	-0.135	-8.85
DL	1.02	46.26
CO	1.38	89.96
INTER	0.765	37.99
TRAF	-2.70	-43.44
AG_TC	-0.128	-4.04
DIS_TC	0.00937	11.59
FEM_TC	-0.0377	-4.31

Table A2: Utility function specifications

$V_{Drive} = ASC_{Drive} + \beta_{TT}TT_{Drive} + \beta_{TC}TC_{Drive} + \beta_{AG\_TC}(AG * TC_{Drive}) + \beta_{DIS\_TC}(DIS * TC_{Drive}) + \beta_{FEM\_TC}(FEM * TC_{Drive}) + \beta_{DL}DL + \beta_{CO}CO + \beta_{TRAF}TRAF$
$V_{PubTr} = ASC_{PublTr} + \beta_{TT}TT_{PubTr} + \beta_{TC}TC_{PubTr} + \beta_{AG\_TC}(AG * TC_{PubTr}) + \beta_{DIS_{TC}}(DIS * TC_{PubTr}) + \beta_{FEM\_TC}(FEM * TC_{PubTr}) + \beta_{INTER}INTER$
$V_{Walk} = ASC_{Walk} + \beta_{TT}TT_{Walk} + \beta_{TC}TC_{Walk} + \beta_{AG\_TC}(AG * TC_{Walk}) + \beta_{DIS\_TC}(DIS * TC_{Walk}) + \beta_{FEM\_TC}(FEM * TC_{Walk})$

Notations

V_i Observed part of utility of alternative *i*

ASC_i Specific constant of alternative *i*

β_{TT}	Taste parameter associated with travel time attribute
β_{TC}	Taste parameter associated with travel cost attribute
β_{AG_TC}	Taste parameter associated with interaction between age and travel cost attribute
β_{DIS_TC}	Taste parameter associated with interaction between travel distance and travel cost attribute
β_{FEM_TC}	Taste parameter associated with interaction between gender and travel cost attribute
β_{DL}	Taste parameter associated with driving license attribute
β_{CO}	Taste parameter associated with number of owned car attribute
β_{TRAF}	Taste parameter associated with traffic variability attribute
β_{INTER}	Taste parameter associated with number of interchanges attribute

References

Abu-Mostafa, Y. S., Magdon-Ismail, M., & Lin, H.-T. (2012). *Learning from data* (Vol. 4): AMLBook New York, NY, USA:.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). *Sanity checks for saliency maps.* Paper presented at the Advances in Neural Information Processing Systems.

Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K. T., Montavon, G., . . . Kindermans, P.-J. (2019). iNNvestigate neural networks! *Journal of Machine Learning Research, 20*(93), 1-8.

Alwosheel, A., van Cranenburgh, S., & Chorus, C. (2017). *Artificial neural networks as a means to accommodate decision rules in choice models.* Paper presented at the International Choice Modelling Conference 2017.

Alwosheel, A., van Cranenburgh, S., & Chorus, C. G. (2018). Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. *Journal of Choice Modelling, 28*, 167-182. doi:https://doi.org/10.1016/j.jocm.2018.07.002

Ancona, M., Ceolini, E., Öztireli, C., & Gross, M. (2017). *A unified view of gradient-based attribution methods for deep neural networks.* Paper presented at the NIPS 2017-Workshop on Interpreting, Explaining and Visualizing Deep Learning.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one, 10*(7), e0130140.

Bishop, C. M. (2006). *Pattern recognition and machine learning*: springer.

Böhle, M., Eitel, F., Weygandt, M., & Ritter, K. (2019). Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. *Frontiers in aging neuroscience, 11*, 194.

Chen, C., Ma, J., Susilo, Y., Liu, Y., & Wang, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies, 68*, 285-299.

Chiang, W.-y. K., Zhang, D., & Zhou, L. (2006). Predicting and explaining patronage behavior toward web and traditional stores using neural networks: a comparative analysis with logistic regression. *Decision Support Systems, 41*(2), 514-531.

Chollet, F. (2015). Keras.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems (MCSS), 2*(4), 303-314.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Golshani, N., Shabanpour, R., Mahmoudifard, S. M., Derrible, S., & Mohammadian, A. (2018). Modeling travel mode and timing decisions: Comparison of artificial neural networks and copula-based joint model. *Travel Behaviour and Society, 10*, 21-32.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* (Vol. 1): MIT press Cambridge.

Hagenauer, J., & Helbich, M. (2017). A comparative study of machine learning classifiers for modeling travel mode choice. *Expert Systems with Applications, 78*, 273-282.

Hall, P., & Gill, N. (2018). *An Introduction to Machine Learning Interpretability-Dataiku Version*: O'Reilly Media, Incorporated.

Haykin, S. S. (2009). *Neural networks and learning machines* (Vol. 3): Pearson Upper Saddle River.

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. *Transportation Research Part E: Logistics and Transportation Review, 36*(3), 155-172.

Hillel, T., Elshafie, M., & Ying, J. (2018). Recreating passenger mode choice-sets for transport simulation: A case study of London, UK. *Proceedings of the Institution of Civil Engineers - Smart Infrastructure and Construction, 171*(1), 29-42. doi:10.1680/jsmic.17.00018

Hillel, T., Elshafie, M. Z., & Jin, Y. (2018). Recreating Passenger Mode Choice-Sets for Transport Simulation. *Proceedings of the Institution of Civil Engineers-Smart Infrastructure and Construction*, 1-49.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks, 2*(5), 359-366.

Karlaftis, M. G., & Vlahogianni, E. I. (2011). Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. *Transportation Research Part C: Emerging Technologies, 19*(3), 387-399.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kittley-Davies, J., Alqaraawi, A., Yang, R., Costanza, E., Rogers, A., & Stein, S. (2019). *Evaluating the effect of feedback from different computer vision processing stages: a comparative lab study.* Paper presented at the Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *Imagenet classification with deep convolutional neural networks.* Paper presented at the Advances in neural information processing systems.

Lapuschkin, S., Binder, A., Montavon, G., Muller, K.-R., & Samek, W. (2016). *Analyzing classifiers: Fisher vectors and deep neural networks.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Lapuschkin, S., Binder, A., Montavon, G., Müller, K.-R., & Samek, W. (2016). The LRP toolbox for artificial neural networks. *The Journal of Machine Learning Research, 17*(1), 3938-3942.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436.

LeCun, Y. A., Bottou, L., Orr, G. B., & Müller, K.-R. (2012). Efficient backprop *Neural networks: Tricks of the trade* (pp. 9-48): Springer.

Lee, D., Derrible, S., & Pereira, F. C. (2018). Comparison of four types of artificial neural network and a multinomial logit model for travel mode choice modeling. *Transportation Research Record, 2672*(49), 101-112.

Lipton, Z. C. (2016). The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*.

Mckinney, S. M., Sieniek, M., Gilbert, F., Godbole, V., Godwin, J., Antropova, N., . . . Corrado, G. C. (2020). International evaluation of an AI system for breast cancer screening. *Nature, 577*, 89-94. doi:10.1038/s41586-019-1799-6

Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing, 73*, 1-15.

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). The building blocks of interpretability. *Distill, 3*(3), e10.

Polson, N. G., & Sokolov, V. O. (2017). Deep learning for short-term traffic flow prediction. *Transportation Research Part C: Emerging Technologies, 79*, 1-17.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *Why should i trust you?: Explaining the predictions of any classifier.* Paper presented at the Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.

Rosenfeld, A., & Richardson, A. (2019). Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems*, 1-33.

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (2019). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*: Springer Nature.

Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). *Learning important features through propagating activation differences.* Paper presented at the Proceedings of the 34th International Conference on Machine Learning-Volume 70.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.

Sun, Y., Jiang, Z., Gu, J., Zhou, M., Li, Y., & Zhang, L. (2018). Analyzing high speed rail passengers' train choices based on new online booking data in China. *Transportation Research Part C: Emerging Technologies, 97*, 96-113.

van Cranenburgh, S., & Alwosheel, A. (2019). An artificial neural network based approach to investigate travellers' decision rules. *Transportation Research Part C: Emerging Technologies, 98*, 152-166.

Xie, D.-F., Fang, Z.-Z., Jia, B., & He, Z. (2019). A data-driven lane-changing model based on deep learning. *Transportation Research Part C: Emerging Technologies, 106*, 41-60.

Zeiler, M. D., & Fergus, R. (2013). Stochastic pooling for regularization of deep convolutional neural networks. *arXiv preprint arXiv:1301.3557*.

Zintgraf, L. M., Cohen, T. S., Adel, T., & Welling, M. (2017). Visualizing deep neural network decisions: Prediction difference analysis. *arXiv preprint arXiv:1702.04595*.