



Delft University of Technology

An artificial neural network based method to uncover the value-of-travel-time distribution

van Cranenburgh, Sander; Kouwenhoven, Marco

DOI

[10.1007/s11116-020-10139-3](https://doi.org/10.1007/s11116-020-10139-3)

Publication date

2020

Document Version

Final published version

Published in

Transportation

Citation (APA)

van Cranenburgh, S., & Kouwenhoven, M. (2020). An artificial neural network based method to uncover the value-of-travel-time distribution. *Transportation*, 48(5), 2545-2583. <https://doi.org/10.1007/s11116-020-10139-3>

Important note

To cite this publication, please use the final published version (if applicable). Please check the document version above.

Copyright

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights. We will remove access to the work immediately and investigate your claim.



An artificial neural network based method to uncover the value-of-travel-time distribution

Sander van Cranenburgh¹ · Marco Kouwenhoven¹

© The Author(s) 2020

Abstract

This study proposes a novel Artificial Neural Network (ANN) based method to derive the Value-of-Travel-Time (VTT) distribution. The strength of this method is that it is possible to uncover the VTT distribution (and its moments) without making assumptions about the shape of the distribution or the error terms, while being able to incorporate covariates and taking the panel nature of stated choice data into account. To assess how well the proposed ANN-based method works in terms of being able to recover the VTT distribution, we first conduct a series of Monte Carlo experiments. After having demonstrated that the method works on Monte Carlo data, we apply the method to data from the 2009 Norwegian VTT study. Finally, we extensively cross-validate our method by comparing it with a series of state-of-the-art discrete choice models and nonparametric methods. Based on the promising results we have obtained, we believe that there is a place for ANN-based methods in future VTT studies.

Keywords Artificial neural network · Value of travel time · Random valuation · Nonparametric methods · Discrete choice modelling

Introduction

The Value-of-Travel Time (VTT) plays a decisive role in the Cost–Benefit Analyses (CBAs) of transport policies and infrastructure projects as well as in travel demand modelling. The VTT expresses travel time changes in monetary values (Small 2012). Due to its importance for transport policies and appraisal, the VTT is one of the most researched notions in transport economics (Abrantes and Wardman 2011). Most Western societies conduct studies to determine the VTT on a regular basis. The focus of such VTT studies is typically not to obtain a single (mean) VTT for all trips, but rather to obtain tables of VTTs which show how the VTT depends on trip characteristics, such as travel purpose and mode.

Despite decades of experience with data collection and VTT inference, the best way to obtain the VTT is still under debate. Early studies predominantly used Revealed Preference (RP) data in combination with Multinomial Logit (MNL) models (Wardman et al.

✉ Sander van Cranenburgh
S.vanCranenburgh@tudelft.nl

¹ Transport and Logistics Group, Delft University of Technology, Delft, The Netherlands

2016). However, despite the well-known advantages of RP data over Stated Choice (SC) data (Train 2003), nowadays RP data are seldom used in VTT studies. The main reason is that while the travellers' choices are observable, their actual trade-offs across alternatives are not—which hampers estimation of the VTT using RP data. More recent VTT studies therefore favour using SC data in combination with discrete choice models that account for (some of the) potential artefacts of the SC experiment (notably size and sign effects) (Fosgerau et al. 2007; Ramjerdi et al. 2010; Börjesson and Eliasson 2014; Kouwenhoven et al. 2014; Hess et al. 2017). In one of the most common kind of VTT SC experiment travellers are presented a binary within-mode choice task where the alternatives are made up out of travel cost and travel time only. Recent VTT studies report particularly good results with so-called Random Valuation models when analysing data from this type of SC experiments (Fosgerau and Bierlaire 2009; Börjesson and Eliasson 2014; Ojeda-Cabral et al. 2018).

Besides discrete choice models, nowadays nonparametric methods are increasingly pioneered in VTT studies (Fosgerau 2006, 2007). These methods are methodologically appealing as they do not make assumptions regarding the shape of the VTT distribution and the structure of the error terms. However, despite their methodological elegance they are typically not used to derive VTTs for appraisal. Rather, they are used as a first, complementary, step to learn about the shape of the distribution of the VTT, after which parametric discrete choice models are estimated to derive VTTs for appraisal. Börjesson and Eliasson (2014) argue that nonparametric methods are not suitable to compute VTTs for appraisal for three reasons. First, they (often) cannot incorporate covariates. Second, they (often) cannot account for panel effects. Third, they (often) do not recover the VTT distribution over its entire domain. That is, the distribution right of the highest VTT bid is not recovered, which hinders computation of the mean VTT.

Very recently, Artificial Neural Networks (ANNs) are gaining ground in the travel behaviour research arena (e.g. Mohammadian and Miller 2002; Cantarella and de Luca 2005; Karlaftis and Vlahogianni 2011; Omrani et al. 2013; Pereira et al. 2015; Alwosheel et al. 2018; Wong et al. 2017; Golshani et al. 2018; Lee et al. 2018; Sifringer et al. 2020; Van Cranenburgh and Alwosheel 2019; Wang et al. 2020). ANNs are mathematical models which are loosely inspired by the structure and functional aspects of biological neural systems. A fundamental difference between discrete choice models and ANNs is the modelling paradigm to which they belong. Discrete choice models are theory-driven, while ANNs are data-driven. Theory-driven models work from the principle that the true Data Generating Process (DGP) is a (stochastic) function, which can be uncovered (Erdem et al. 2005). To do so, the analyst imposes structure on the model. In the context of discrete choice models this is done by prescribing the utility function, the decision rule, the error term structure, etc. Then, the analyst estimates the model's parameters, usually compares competing models, and interprets the results in light of the theory. An advantage of such a theory-driven method is that its theoretical framework provides solid ground to interpret the model and its outcomes. A drawback is however that it heavily relies on potentially erroneous assumptions regarding choice behaviour, i.e. the assumptions may not accurately describe the true underlying DGP—leading to erroneous inferences. Data-driven methods work from the principle that the true underlying process is complex and inherently unknown. In a data-driven modelling paradigm the aim is not to uncover the DGP, but rather to learn a function that accurately approximates the underlying DGP. The typical outcome in a data-driven modelling paradigm is a network which has good prediction performance (Karlaftis and Vlahogianni 2011). A drawback of many data-driven methods is that—without further intervention—they provide limited (behavioural) insights on the underlying DGP, such as the relative importance of attributes, Willingness-to-Pay, or VTT.

This holds especially true for ANNs, which are widely considered black boxes (Castelvecchi 2016), amongst other things, because it is impossible to interpret or diagnose ANNs by looking at the weights of the network. In fact, the weights will tell the analyst nothing about the importance of attributes, or whether the ANN has learned intuitively correct relationships, as opposed to spurious ones (see Alwosheel et al. 2019 for recent work in choice modelling to open-up ANN's black boxes). Yet, these behavioural insights are typically most valuable to travel behaviour researchers and for transport policy-making.

In the field of travel behaviour research there is a general sense that ANNs (and other data-driven models), could complement existing (predominantly) theory-driven research efforts (Chen et al. 2016). In light of that spirit, this paper¹ develops an ANN-based method to investigate the VTT distribution. This method *blends* theory-driven and data-driven modelling paradigms. Specifically, we develop a novel pattern recognition ANN which is able to estimate travellers' individual underlying VTTs based on the behavioural notion of indifference. Our method capitalises on the strong prediction performance of ANNs (see Paliwal and Kumar 2009 for a comprehensive review of articles that involve a comparative study of ANNs and statistical techniques). The strength of this method is that it is possible to uncover the VTT distribution (and its moments) without making assumptions on the underlying behaviour. For instance, it does not prescribe the utility function, the shape of the VTT distribution, or the structure of the error terms. Moreover, the method can incorporate covariates, account for panel effects and does yield a distribution right of the maximum VTT bid. Thereby, it overcomes important limitations associated with some other nonparametric methods. Finally, the method does not require extensive software coding on the side of the analyst as the method is built on a standard MultiLayer Perceptron (MLP) architecture. Hence, the method can be applied using off-the-shelf (open-source) software.

The remainder of this paper is organised as follows. Section "[Methodology](#)" develops the ANN-based method for uncovering the VTT distribution. Section "[Results](#)" conducts a series of Monte Carlo analyses to assess how well the method works. Section "[Application to real VTT data](#)" applies the method to an empirical VTT data set from a recent VTT study. Section "[Cross-validation](#)" cross-validates the method by comparing its results with those obtained using a series of state-of-the-art discrete choice models and nonparametric methods. Finally, Section "[Conclusions and discussion](#)" draws conclusions and provides a discussion.

Methodology

Preliminary

Data format

Throughout this paper we suppose that we deal with data from a classic binary SC experiment, consisting of $T + 1$ choice observations per individual, in which within-mode trade-offs between travel cost (TC) and travel time (TT) are embedded. In this format there is

¹ An earlier version of this paper has been published in the conference proceedings of the 2019 International Work-Conference on Artificial Neural Networks (IWANN) (Van Cranenburgh and Kouwenhoven 2019b). This paper extends this work by presenting (1) a behaviour framework, (2) Monte Carlo analyses, (3) extended empirical results, and (4) a more extensive cross-validation.

always a *fast and expensive* alternative and a *slow and cheap* alternative.² This data format is in line with standard VTT practice in many Western European countries, including the UK (Batley et al. 2017), The Netherlands (HCG 1998; Kouwenhoven et al. 2014), Denmark (Fosgerau et al. 2007), Norway (Ramjerdi et al. 2010) and Sweden (Börjesson and Eliasson 2014). The straightforward format of this type of data makes it fit the proposed methodology well.

In this format, each choice task has what is commonly referred to as a Boundary VTT (BVTT). The BVTT is the implicit price of time. The BVTT is given by Eq. 1 where alternative 1 denotes the slow and cheap alternative and alternative 2 denotes the fast and expensive alternative. The BVTT can be perceived as a valuation threshold as a respondent choosing the fast and expensive alternative reveals a VTT which is (most likely) above the BVTT, while a respondent choosing the slow and cheap alternative reveals a VTT which is (most likely) below the BVTT.

$$BVTT = -\frac{\Delta TC}{\Delta TT} = \frac{-(TC_2 - TC_1)}{(TT_2 - TT_1)} \quad (1)$$

Covariates in VTT studies

It is important to incorporate covariates in models that aim to infer the VTT. Börjesson and Eliasson (2014) provide four reasons for this. Firstly, accounting for covariates in VTT models allows better extrapolating the VTT to new situations. Secondly, accounting for covariates in VTT models allows better understanding what trip characteristics influence the VTT. Thirdly, accounting for covariates in VTT models allows the analyst to remove the influence of undesirable factors, such as income or urbanisation level from the VTT used for appraisal. Fourthly, accounting for covariates in VTT models allows capturing so-called size and sign effect stemming from the experimental design (De Borger and Fosgerau 2008). Size effects are due to the behavioural notion that the VTT is dependent on the size of the difference in travel time and travel cost across alternatives in the choice task (Daly et al. 2014). Sign effects are due to the behavioural notion that losses (e.g. higher travel cost and longer travel time) loom larger than equivalently sized gains (e.g. lower travel cost and shorter travel time) (Ramjerdi and Lindqvist Dillén 2007; De Borger and Fosgerau 2008).

Trade-offs between travel cost and travel time can be classified into four different types,³ which can be visualised into four different quadrants, see Fig. 1 where the origin corresponds with the reference situation. The top-left quadrant consists of ‘Willingness to Pay’ (WTP) type choices. Here, the reference trip is compared to a faster but more expensive alternative. The top-right quadrant consists of ‘Equivalent Loss’ (EL) type choices. Here, a more expensive alternative is compared to a slower alternative (relative to the reference). In the lower-left quadrant are ‘Equivalent Gain’ (EG) type choices. Here, a cheaper alternative is compared to a faster alternative (relative to the reference). Finally, the lower-right

² Some surveys include a choice with a dominant alternative, i.e. a choice between a fast and cheap alternative and a slow and expensive alternative. These observed choices are used to identify respondents which need to be removed from further analysis.

³ Note that trade-offs between travel cost and travel time can also be posed such that they are in more than one quadrant.

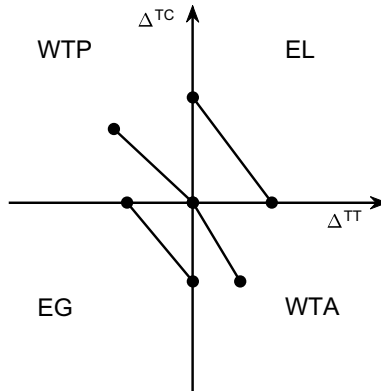


Fig. 1 Sign effects

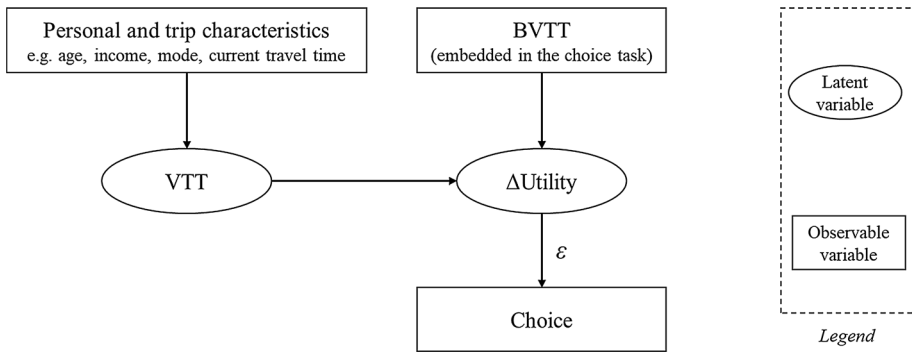


Fig. 2 Conceptual framework

quadrant consists of ‘Willingness to Accept’ (WTA) type choices. Here, the reference alternative is compared to a cheaper but slower alternative.

In our method we make a distinction between generic covariates and what we call experimental covariates. Generic covariates are not dependent on the composition of a choice task in the SC experiment, and typically relate to characteristics of the decision maker or his/her current travel behaviour, such as income, gender, age, and current travel time. Experimental covariates are dependent on the composition of the choice task in the SC experiment. Typical examples are size and sign effects.

Conceptual framework

Figure 2 shows the conceptual framework of this paper. We assume that each decision-maker, when confronted with a binary choice task which embeds a BVTT, comes to his or her choice by comparing the BVTT with his or her own VTT. Personal characteristics as well as current trip characteristics are explanatory variables for the individuals’ VTTs. The comparison of the BVTT and the individual’s VTT results in a utility difference, based on which the choice is made. We assume choice behaviour is stochastic. But, we are agnostic about the source of the stochasticity. That is, the stochasticity can be on the side of the decision maker due to e.g. a trembling hand, boredom, loss of

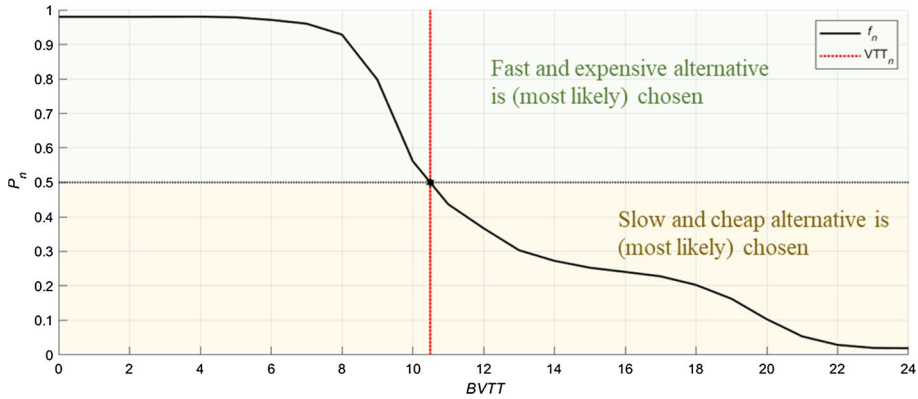


Fig. 3 Choice probability generating function f_n

attention, or other physiological processes—congruent to models in the mathematical psychology (see Luce 2014); or, it can be due to unobserved factors on the side of the analyst—congruent with economist’s view on decision making (McFadden 1974). In our interpretations in the remaining part of this paper we however go with the former.

Thus, in our framework for each individual there exists a function that relates the BVTT to the probability P that he or she chooses the fast and expensive alternative. Henceforth, we call this function the choice probability generating function, denoted f_n . Figure 3 conceptually illustrates this function. It shows that in case the BVTT is smaller than the VTT for decision maker n , the fast and expensive alternative is (most likely) chosen, while in case the BVTT is larger than the decision maker’s VTT, the slow and cheap alternative is (most likely) chosen. In the special case in which the BVTT equals the decision maker’s VTT, the utility difference between the two alternatives equals zero and P_n equals 0.5. We call the BVTT which makes the decision maker indifferent between the two alternative the ‘indifference point’.

Uncovering individual VTTs using an ANN

The ANN-based method is based on three observations. The first observation is that ANNs are universal approximators. That is, ANNs can, under mild assumptions, approximate a wide variety of continuous functions on compact subsets (Cybenko 1989). This ability stems from the versatile structure of ANNs, which allow them to capture non-linearity and interactions between variables (without explicitly being programmed where to find these). The second observation is that we can use an ANN to learn to approximate each decision maker’s choice probability generating function f_n . To do this, we can use all available information we have on each decision maker, including the responses to probed BVTTs and covariates, such as income and age. The third observation is that given f_n we can determine the VTT for each individual by making use of the notion of indifference (see Sect. 2.2). Specifically, under the assumption that the ANN is capable to approximate f_n (for each decision maker), the indifference points reveal the decision makers’ true underlying VTTs.

To do so, we take the following 6 steps:

1. Training and data preparation

The aim of this step is to train an ANN to (probabilistically) predict, for decision maker n the choice in a hold-out choice task $T+1$, based on the tuple consisting of the BVTTs ($BVTT^n$) and responses (Y^n) in choice tasks 1 to T , the probed BVTT in choice task $T+1$ ($bvtt_{T+1}^n$), the experimental covariates in choice task $T+1$ (s_{T+1}^n), and the sets of generic and experimental covariates, denoted D^n and S^n , respectively. That is, we train the ANN to learn the relationships g , see Eq. 2, where P_{T+1}^n denotes the probability of observing a choice for the fast and expensive alternative in choice task $T+1$ for decision maker n .

$$P_{T+1}^n = g(BVTT^n, Y^n, bvtt_{T+1}^n, s_{T+1}^n, D^n, S^n)$$

$$\text{where } BVTT^n = \{bvtt_1^n, bvtt_2^n, \dots, bvtt_T^n\}$$

$$Y^n = \{y_1^n, y_2^n, \dots, y_T^n\}$$

$$S^n = \{s_1^n, s_2^n, \dots, s_T^n\}$$
(2)

Figure 4 shows the proposed architecture of the ANN.⁴ At the input layer, the independent variables enter the network. At the top, there are the generic covariates (green). Typical generic covariates encountered in VTT studies are mode, purpose, age, income, distance, etc. Below the generic covariates are the variables associated with choice tasks 1 to T (red). These include the BVTTs, the choices y and experimental covariates s (sizes and signs). Below the variables for choice tasks 1 to T is an extra set of input nodes for choice task R (blue). Choice task R is a replication of one choice task, randomly picked from the set choice tasks 1 to T . These input nodes come in handy later when the ANN is used for simulation (they make it possible to use all $T+1$ observations instead of only T observations in the simulation). Finally, at the bottom are the variables associated with hold-out choice task $T+1$ (yellow). These are essentially the ‘knobs’ of the model that can be used for simulation. The output of the network is the probability for choosing the fast and expensive alternative in choice task $T+1$. The input layer and the output layer are connected via so-called hidden layers by arrows, which contain the weights w that need to be learned (see Bishop 1995 for an extensive overview of ANNs and their characteristics). At each node, the inputs are summed and an activation function is applied. For the output node (purple) this operation is exemplified in Fig. 4. The value V is computed in a linear-additive fashion, based on the outputs of the nodes in the previous layer and a set of associated weights. Here, V has the same function as the difference in utility in a conventional (binary) discrete choice model, and also enters a sigmoid function (aka logit function, in choice modeller’s parlance) to yield the choice probability. One or multiple hidden layer can be used. In our analyses in Sections “Results” and “Application to real VTT data” we find two layers to work optimal. However, the optimal number of hidden layers and the number of nodes depend on the complexity of the DGP that needs to be learned from the data, and thus may vary across applications.

To train (‘estimate’ in statistics parlance) the network in Fig. 4, we need to prepare the data. To do so, for each decision maker in the data we randomly draw T explanatory choice tasks from the $T+1$ choice tasks that are available in the data for each decision maker. These T choice tasks are used as independent variables to predict the remaining choice. To

⁴ Note that while the graph representation of ANNs looks in some ways similar to that of a cross-nested logit model, ANNs and cross-nested logit models are disparate, conceptually as well as mathematically.

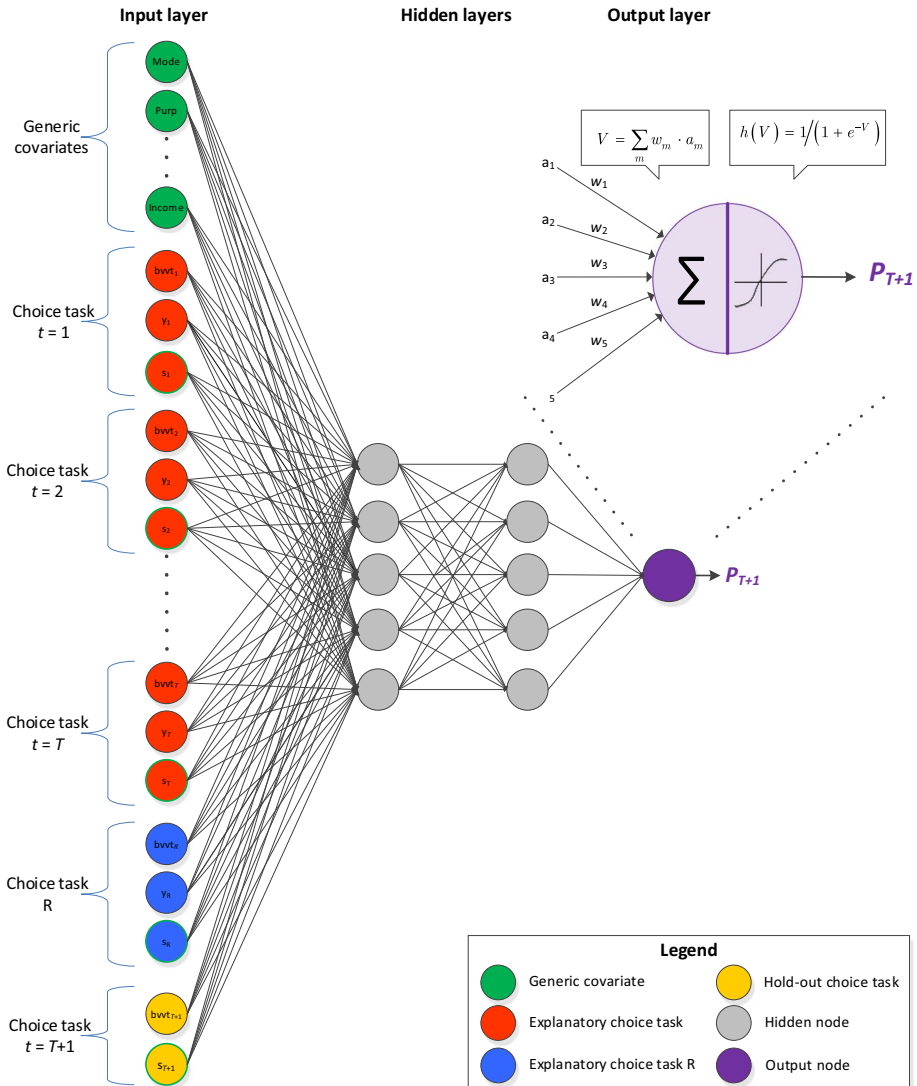


Fig. 4 ANN architecture

avoid that the network undesirably learns a particular structure in the data, rather than the explanatory power of the variables it is crucial that the order in the set of T explanatory choice tasks is randomised.⁵ Furthermore, for each decision maker we create K observations, by using different randomisations of the order in the set of explanatory choice tasks. The idea behind this is that the weights associated with the choice tasks attain (roughly)

⁵ Unless the order of the choice tasks is randomised during the data collection. Note that by doing so the network becomes blind to potential learning effects on the side of the respondent when conducting the survey. We come back to this point in the discussion.

similar sizes. By doing so, we create a network that produces stable predictions, which is insensitive to the order in which the explanatory choice tasks are presented to the network. Note that while such ‘oversampling’ of data would be considered a sin in theory-driven research (as it will lead to underestimated standard errors), in machine learning oversampling is occasionally done (Chawla et al. 2002). In each manifestation of the K randomisations, choice task R (see Fig. 4) is a randomly selected replication of one of the T explanatory choice tasks. By selecting a random choice task, we make sure that no single choice task weights more heavily in the training and ensure that the weights of the network are generic across all choice tasks.

Training the network involves maximising its prediction performance. For this task, we minimise the cross-entropy (Shannon and Weaver 1949), since we are dealing with a classification problem. Note that minimising cross-entropy is equivalent to maximising the likelihood of the data. To train the network backward propagation is used (Rumelhart et al. 1986). There are a number of reasons why backward propagation is preferred over e.g. Maximum Likelihood Estimation (MLE) based techniques for training ANNs. The most important reasons are that backward propagation is better equipped to deal with the (1) indeterminacy and (2) non-convexity of ANNs than MLE based techniques. Furthermore, to evaluate the performance of the network, in machine learning always a hold-out sample is used. The reason is that ANNs can be sensitive to overfitting when it is trained on in-sample data only. More details on training neural networks can be found in many textbooks, including Bishop (1995).

2. Simulate

After having trained the ANN (i.e. learned g) we use the ANN to simulate choice probabilities for each decision-maker in the data. Specifically, we simulate P_{T+1}^n while letting $bvtt_{T+1}^n$ run from 0 to a maximum BVTT value, set by the analyst using a finite step size. For simulation, we can use all $T+1$ choice observations of a decision maker as explanatory choice tasks. This is possible because we created the extra choice task R in the network (see step 1). Thus, this ‘trick’ allows using all available information on a decision maker’s preference for predicting his or her response to a given probed BVTT in the simulation in an elegant way.

Finally, it is important to note that in the simulation the analyst can also manipulate the experimental covariates s in choice task $T+1$ (yellow input node at the bottom-left of Fig. 4). The analyst can use this to assess the effects of e.g. sizes and signs on the VTT. The latter is particularly useful when the analyst would like to investigate the gap between WTA and WTP or desires to obtain a reference free VTT (which requires having quadrant specific VTTs). In contrast, the analyst should not manipulate the generic covariates (green input nodes at the top-left of Fig. 4), e.g. to simulate the effect of a change in income levels. The reason for this is that the generic covariates (e.g. income, gender, etc.) and choices in the explanatory choice tasks are intrinsically correlated. As such, changing a generic covariate (say income level) only partially captures its effect on the VTT as the choices in the explanatory choice tasks do not change when changing a generic covariate, while they actually ‘should’.

3. Recovery of the VTT of individual decision makers

The next step is to infer from these simulated probabilities the VTT for each decision maker. The core idea in this method is that g approximates f_n for all n , and thus can be

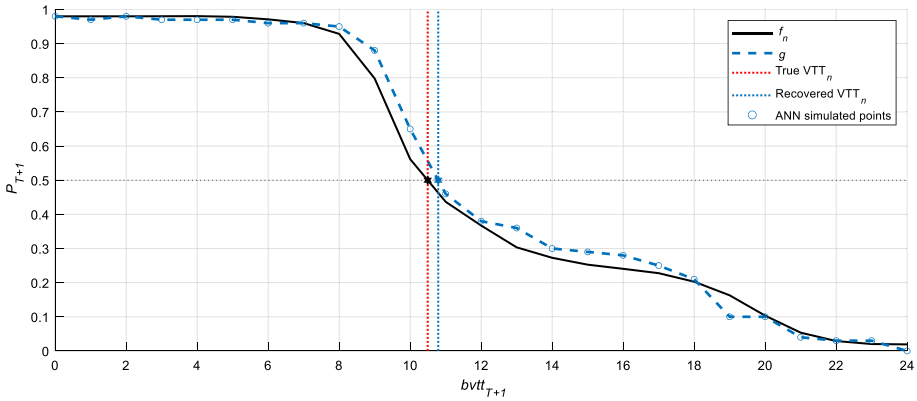


Fig. 5 Simulated choice probabilities

used to infer the VTT for each individual based on the notion of indifference. To infer a VTT of a decision maker we need to find the BVTT that makes the decision maker indifferent between the fast and expensive and the slow and cheap alternatives. Figure 5 illustrates this idea conceptually. The x -axis shows the BVTT in the $T + 1$ th choice task, and the y -axis shows the probability that decision-maker n chooses the fast and expensive alternative in this choice task. The true choice probability generating function for this decision-maker f_n is depicted by the black line. Note that f_n does not necessarily have the conventional logit S -shape: for illustrational purposes, here it is asymmetric and it does not asymptotically go to $P = 1$ and $P = 0$ for very low and high BVTTs, respectively. The black star indicates the true indifference point, which hence reveals the true VTT for this decision maker. The ANN’s approximation g is depicted in blue. The blue star indicates the indifference point according to the ANN, which is conceived as a proxy for the VTT. The true and recovered VTT do not complete coincide. The (horizontal) difference between the true and the recovered VTT can be understood as the estimation error.

Practically, we have several options to determine the BVTT which yields $P = 0.5$, based on the simulated data points. A simple and effective approach to do this is by determining the last simulated point above $P = 0.5$ and the first simulated point below $P = 0.5$, and then make a linear interpolation between those two points and to solve for the BVTT which makes the individual indifferent.

4. Repeat steps 2 and 3

We repeat steps 2 and 3 numerous times (e.g. 20 times). In each repetition we shuffle the order of the explanatory choice tasks. This step is not strictly obligatory, but it helps to improve the stability of the outcomes. In particular, it takes out the effect of the order in which the explanatory choice tasks are presented to the network. Hence, for each decision maker his/her VTT is computed numerous times. After that, we compute each decision maker’s VTT by taking the mean across all repetitions.

5. Construct the VTT distribution

Having an estimate of the VTT for each decision maker, we can construct an empirical distribution of the VTT. Also, from the constructed empirical distribution we can readily compute the mean and standard deviation of the VTT.

6. VTT for use in applied appraisal

For appraisal typically not a single mean VTT is needed, but rather one or more tables that show the mean VTT for specific trip categories; usually combinations of modes and purposes. We can create such tables from the VTTs derived under step (5). In case the sample is representative for the target population, we can suffice with splitting the sample into categories and compute the mean VTT for each category. However, in most real life situations the sample is skewed as compared to the target population. Therefore, the in-sample mean VTTs usually cannot immediately be used to produce tables for appraisal. To correct for the skew, the analyst needs to reweight the sample and compute weighted mean VTTs (for each category). The most straightforward approach to account for an unrepresentative sample is by means of computing the VTT for each combination of the covariates, say of income and distance, and reweight the contribution of each combination to the overall VTT according to its under or overrepresentation (as compared to the ideal population shares). This matrix-based approach is taken in several VTT studies (e.g. the UK 2003 VTT study and the Norwegian 2009 study) and especially works well with a limited number of covariates. In case of many covariates to account for, this approach can be unwieldy. An alternative approach which works better in case of many covariates is based on sample enumeration (Batley et al. 2017). This approach applies weights at the level of the respondent in the data and is used in e.g. the latest Dutch and UK VTT studies. Both approaches can be used in combination with the ANN-based method. However, given that our method directly provides VTTs at the individual level, the sample enumeration based approach is the most natural choice to account for an unrepresentative sample.

Consistency with random utility maximisation

Examining Random Utility Maximisation (RUM) consistency of the proposed method is important considering the use of the VTT derived by this method in utilitarian economic appraisal methods, such as the Cost–Benefit Analysis. The RUM modelling paradigm is an extremely general approach to modelling behaviour; in the core it postulates that individual behaviour at each moment is consistent with utility maximisation. RUM consistency has extensively been studied in the choice modelling literature since the inception of RUM models in the 1970s (Daly and Zachery 1978; McFadden 1981; Fosgerau et al. 2013). Central to RUM consistency tests are the notions of transitivity and regularity (Hess et al. 2018). A RUM consistent model satisfies both transitivity and regularity. Transitivity means that if alternative A is preferred to alternative B and alternative B is preferred to alternative C, then alternative A must be preferred to C. Regularity entails that the probability of choosing any given alternative from a set should not increase if the offered set is expanded to include additional alternatives.

The proposed method is not RUM consistent *by design*, but it can be RUM consistent. In Appendix A we show that a necessary condition for our method to be RUM consistent

is that the VTT is uniquely determined. That means that the indifference point must be crossed once and only once. Given the flexible, data-driven nature of the ANN we cannot enforce this to happen in our model (nor would we want that). When the ANN is trained on RUM consistent data, the learned behaviour (and thus the model) will turn out to be RUM consistent, and vice versa. In light of this, we believe it is good practice when using this method to test (for each respondent) whether $P=0.5$ is crossed once and only once. Using the ANN in this way can shed light on the extent to which people behave consistent with RUM. In case unique VTTs are found for all respondents, the learned behaviour is consistent with RUM, and hence the derived VTT can be used in utilitarian economic appraisal methods. However, in case for a number of the respondents (or for some input space) the VTT is not uniquely determined, it is not entirely clear how to use the derived VTTs. Is the ANN as a whole then not consistent with RUM? Or, should only the VTTs derived for those respondents be discarded? These questions are beyond the scope of this paper, but justify further research.

ANN development

In Sect. 2.3 we presented the ANN without going into much detail on its architecture or on underlying design choices. In this subsection we discuss these in more detail. However, it should be noted here that in contrast to theory-driven research, in data-driven research methodological design choices are not guided by theory. Rather, in lack of a (behavioural) theory providing guidance, design choices in data-driven methods typically involve trial-and-error and building forth on previous works that have been tested on standardised data sets.

To develop an ANN capable of learning function f_n (for all n) we have tested numerous different architectures, including fully and semi-connected networks, different numbers of hidden layers, the presence or absence of bias nodes, and we have tried several different activation functions. The two-hidden layer architecture presented in Fig. 4 with ten nodes at each hidden layer is found to work particularly well for our data.^{6,7} The proposed architecture is a so-called Feed-forward Multilayer Perceptron (MLP). This is one of the most widely used ANNs architectures and is available in virtually all off-the-shelf machine learning software packages. For the transfer functions in the network we find good results using a tan-sigmoid function at the nodes of the hidden layers, and a sigmoid at the nodes of the output layer. Using a sigmoid function at the output layer ensures that the predicted choice probabilities across the two alternatives add up to 1. Hence, one way to look at the ANN is to see it as a juiced-up logistic regression model, where the juicing-up comes from the flexibility provided by the hidden layers.

The fact that off-the-shelf software can be used is a nice feature of this method, as it makes the method accessible for a wide research community. Admittedly, from a methodological perspective our network consumes more weights than is strictly needed, in the sense that in the input layer there are $T+1$ weights for the $bvtt$, y and s , while just one set of weights to be used across all the $T+1$ choice tasks would suffice and hence would yield a more parsimonious network. However, while it is possible to create an architecture with shared weights across inputs variables, this would substantially hinder other researchers

⁶ The network consumes 491 weights in total.

⁷ Note that while no bias nodes are depicted in Fig. 4, the network contains 2×10 bias nodes connected to the hidden layers as they are found to improve the classification performance.

from using this method as most off-the-shelf software does not allow weight sharing, meaning that the analyst needs to write customised codes.

Monte Carlo experiments

This section aims to assess how well the developed ANN-based method is able to recover the underlying true VTT distributions. To do so, we create a series of synthetic data sets, having different DGPs and try to recover the shape of the distribution as well as its mean and its standard deviation using the ANN-based method.

Data generating process

For the Monte Carlo experiments we use the experimental design of the empirical data set that we aim to analyse in the next section, namely the Norwegian 2009 VTT data set, see Ramjerdi et al. (2010) for details on the experimental design. By doing so, we are able to assess how well we can retrieve the underlying VTT distribution with the proposed method in the context of the VTT trade-off points that are embedded in this data set. After cleaning, this data set consists of 5832 valid respondents. For each respondent, 9 binary choices are observed. Based on the experience with the Danish VTT, in the Norwegian 2009 VTT study special care was taken to avoid a large share of non-traders at the high end side. Furthermore, while the currency in the SC experiment was Norwegian Kronor, for reasons of communication we converted all costs into euros (using the exchange rate at the time the experiment was conducted).

So, rather than using the actual observed choices, in this section we replace them with synthetically generated choices. To create synthetic choices, we assume that decision makers make their choices using the Random Valuation (RV) model (Cameron and James 1987) (Eq. 3).^{8,9} The RV model postulates that decision makers choose the faster but more expensive alternative if the decision maker's VTT is higher than the probed BVTT. If the decision maker's VTT is lower than the BVTT, the cheaper but slower alternative is chosen. Like conventional linear-additive Random Utility Maximisation (RUM) models, the RV model consists of an additive error term to accommodate for randomness ε . In the RV model, μ represents the scale factor, which is estimated jointly with the VTT. Note that RV models are consistent with RUM; they can also be casted as a RUM model with a specific form of heteroscedasticity in the error term, see e.g. Börjesson and Eliasson (2014).

$$\begin{aligned} U_{1n} &= \mu \cdot BVTT + \varepsilon_{1n} \quad \text{where } \varepsilon_{in} \sim \text{iid Extreme Value type I} \\ U_{2n} &= \mu \cdot VTT_n + \varepsilon_{2n} \end{aligned} \quad (3)$$

The synthetic data sets are created such that the VTT varies across decision makers according to a specific distribution. Specifically, we have created three data sets in which the VTT takes a normal, a lognormal and a bimodal normal distribution. The normal and lognormal distributions are chosen for two reasons. Firstly, they are frequently used in VTT and, more generally, in choice modelling studies. Secondly, they differ from one another in terms

⁸ Note that we also tested data with linear-additive RUM DGPs. These gave similar results.

⁹ See Ojeda-Cabral and Chorus (2016) and Ojeda-Cabral et al. (2016) for the theoretical relationship between RUM and RV models.

Table 1 Parametrisation of the DGP in synthetic data sets

Data set no	1	2	3
VTT distr	Normal	Lognormal	Bimodal normal
WTP	$VoT_n^{WTP} \sim N(20, 6)$	$VoT_n^{WTP} \sim LN(2.6, 0.5)$	Mixture of 2 normals $VoT_1 \sim N(12, 3.33)$ $VoT_2 \sim N(28, 3.33)$
EL / EG	$VoT_n^{WTP} + 5$	$VoT_n^{WTP} + 5$	$VoT_n^{WTP} + 5$
WTA	$VoT_n^{WTP} + 10$	$VoT_n^{WTP} + 10$	$VoT_n^{WTP} + 10$
μ	- 0.30	- 0.40	- 0.50
ρ^2	0.75	0.75	0.84
Hit rate	0.93	0.93	0.95

Table 2 Training performance on validation data

Data set no	1	2	3
VTT distr	Normal	Lognormal	Bimodal normal
Cross-entropy	0.17	0.17	0.13
ρ^2	0.76	0.75	0.81
Hit rate	0.93	0.94	0.95

of skewness. Therefore, analysing these two distributions can shed light on the extent to which the ANN-based method can accurately capture skew. The bimodal normal distribution is specifically chosen because of its challenging shape. This type of distribution would be quite difficult to recover using conventional parametric methods. As such, this distribution may give insights on how capable the ANN-based method is to recover challengingly shaped VTT distributions.

In line with empirical findings in many recent VTT studies, we created the data such that the VTT of the synthetic decision makers is not a single fixed value, but rather a function of experimental covariates. Specifically, the VTT of the synthetic decision makers is a function of the quadrant in which the choice task is presented (i.e. WTP, EL, EG, or WTA domain). Table 1 shows the parametrisations. To compute the VTT for the different quadrants, we drew a value for the VTT for the WTP quadrant from the associated distribution, and calculated the EL, EG and WTA VTTs by shifting the draw five or ten euros up. By incorporating experimental covariates in the synthetic data sets, we are able to investigate whether the proposed method is able to capture the effect of such covariates (in casu: sign effects) in real data (if present). Furthermore, we use three different scale parameters μ . Thereby, we can test whether method responds well to differences in scale.

Training results

The ANN is implemented in MATLAB2017.¹⁰ To train the network, we find good results using a scaled conjugate gradient algorithm. The training sequence takes about 2 min using a desktop PC (using 4 CPUs). For training, the data were split as follows: 70% of

¹⁰ Code is available upon request from the first author.

Table 3 Monte Carlo results (based on full data)

Data set no VTT distr	1				2				3			
	Normal				Lognormal				Bimodal normal			
	Mean		SD		Mean		SD		Mean		SD	
	DGP	ANN	DGP	ANN	DGP	ANN	DGP	ANN	DGP	ANN	DGP	ANN
WTP	19.98	20.32	5.00	4.34	15.28	15.31	8.06	6.57	20.06	21.13	8.66	7.81
WTA	29.98	30.04	5.00	4.28	25.28	24.22	8.06	6.20	30.06	29.35	8.66	8.11

the data are used for training, 15% for validation and 15% for testing. The observations were randomly allocated to these subsets. Furthermore, we use $K=20$ randomisations (see Sect. 2.3).

Table 2 shows performance indicators for the three data sets. To ease comparison we report besides cross-entropy, also the ρ^2 and hit rate (all based on the performance on the hold-out sample). Table 2 shows that the ANN-based method is able to learn the underlying DGP well in all three data sets. Ideally, the obtained ρ^2 's should approach, but not exceed, those of the true models (Table 1). As can be seen, this is the case. For data set 2 the ANN marginally exceeds the true ρ^2 . But, it exceeds it not to the extent that it is worrisome (e.g. suggests overfitting).

Results

Recovery of VTT distributions

Table 3 shows the results of the Monte Carlo analysis. Specifically, it reports the true and the recovered means and standard deviations of the VTT for the WTP and WTA quadrants. Note that for the sake of exposition, in this section we focus only on the WTP and WTA results. The results for EL and EG are fully in accordance with those of WTP and WTA. Table 3 convincingly shows that ANN-based method is able to accurately recover the mean VTTs for all distributions. In our view, this is quite an achievement, considering that no information on the shape of the VTT distribution has been given to the network. Also the standard deviations are rather well recovered, although the results seem to suggest that they are somewhat underestimated. One possible explanation for this underestimation of the standard deviation is that the ANN has during training only seen few choice patterns of VTTs from the tail of the distribution. Therefore, the ANN may not have been able to adequately learn to predict the responses of those extreme VTTs. In essence, this issue relates to training on unbalanced data sets. This is a frequently encounter issue in data-driven classifiers (Chawla et al. 2002; Prieto et al. 2016). What is different in this context however is that the extent to which the data are unbalanced is inherently unknown, since the VTT is not observed in the data. Therefore, unbalances in the data cannot be accounted for up-front in a direct way. Further research may explore whether techniques used in machine learning to balance data sets improve recovery of the standard deviations.

To see to what extent the ANN-based method is able to recover the shapes of the distributions, Figs. 6, 7 and 8 show histograms and kernel density plots for respectively, the

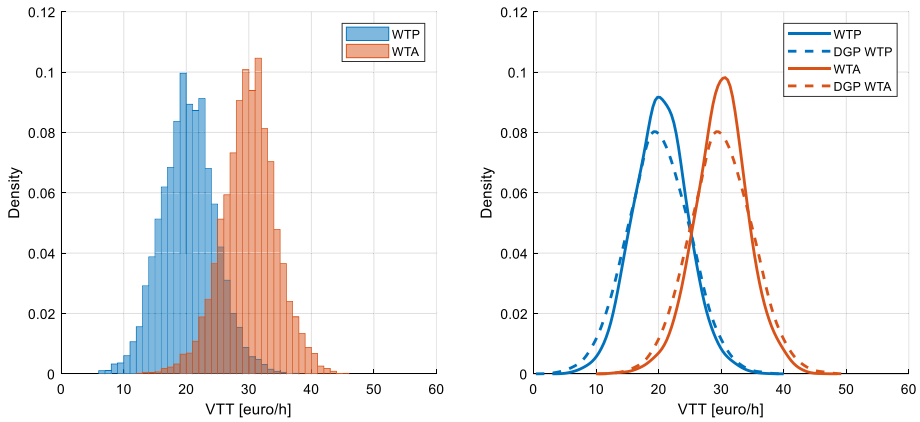


Fig. 6 Normal distribution: Histograms and kernel plots for DGP and ANN VTT

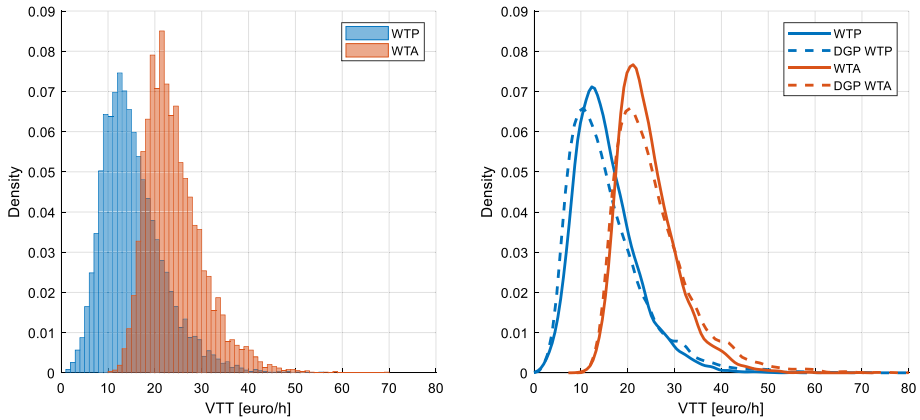


Fig. 7 Lognormal distribution: Histograms and kernel plots for DGP and ANN VTT

normal, lognormal and bimodal normal distribution. The histograms show the VTT distributions for WTP (blue) and WTA (orange). The kernel density plots show the densities of the empirical as well as of the true VTTs (for both quadrants).

A number of observations can be made. Firstly, the histograms reveal that the shapes are well recovered. In line with the true DGPs, Fig. 6 (normal distribution) shows a nice symmetric distribution, while Fig. 7 (lognormal distribution) shows a clearly positively skewed distribution. This shows that the method is sensitive for differences in skew. In fact, Fig. 8 shows that even the challenging shape of the bimodal normal distribution has been recovered. Secondly, the sign effect has accurately been picked up in all three distributions. In line with the true DGP, the WTP and WTA distributions are roughly identical, with the WTA distribution shifted €10/h to the right. Thirdly, the density plots reveal that the shapes of the true VTT distributions and the recovered distributions are close to one another. This is especially true for the normal and lognormal distributions. All in all, the

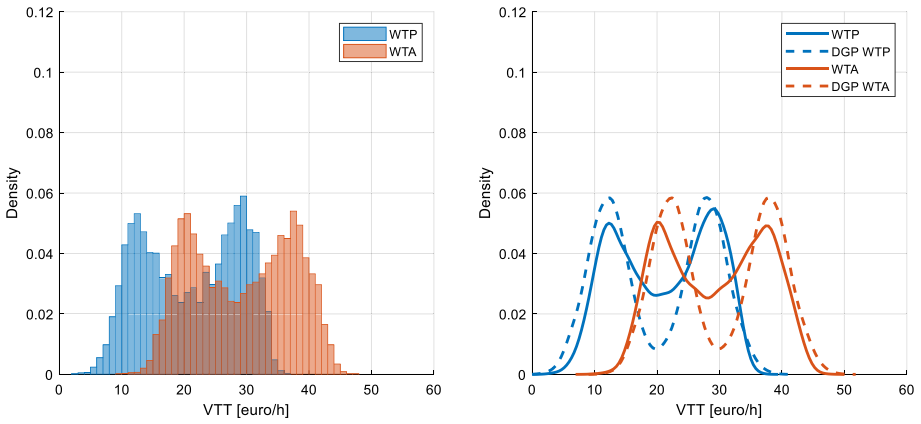


Fig. 8 Bi-modal normal distribution: Histograms and kernel plots for DGP and ANN VTT

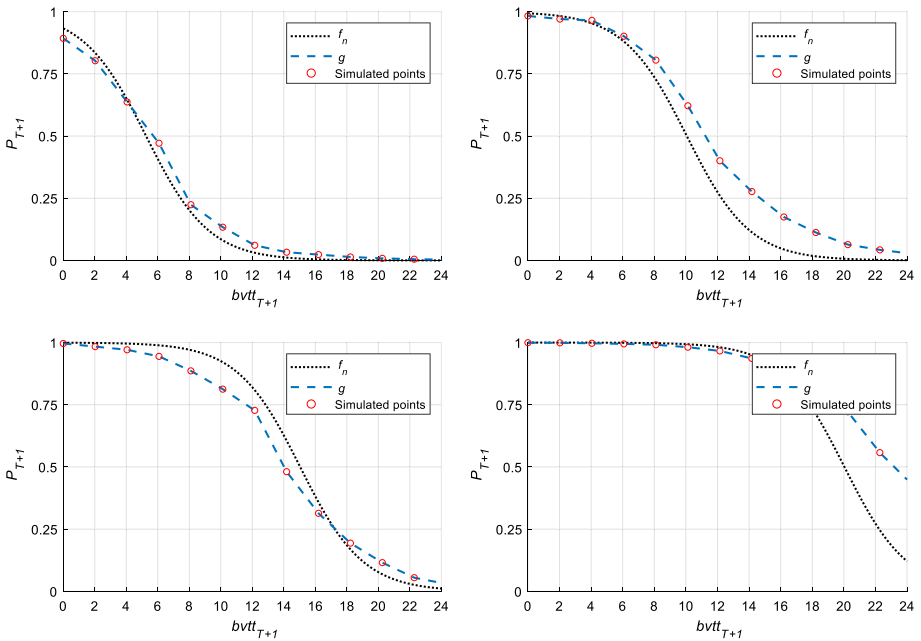


Fig. 9 Approximations of f_n (for four respondents)

results convincingly show that the method is able to accurately recover the shape of the VTT distributions.

Recovery of f_n

The previous subsection has demonstrated that the shape and moments of a VTT distribution can be estimated using the proposed method. This suggests, but does not prove, that the ANN

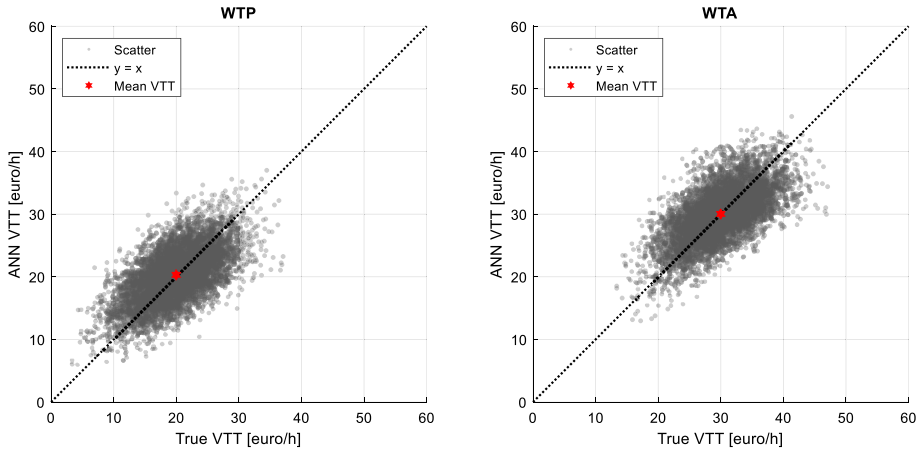


Fig. 10 Normal distribution: true vs recovered VTTs

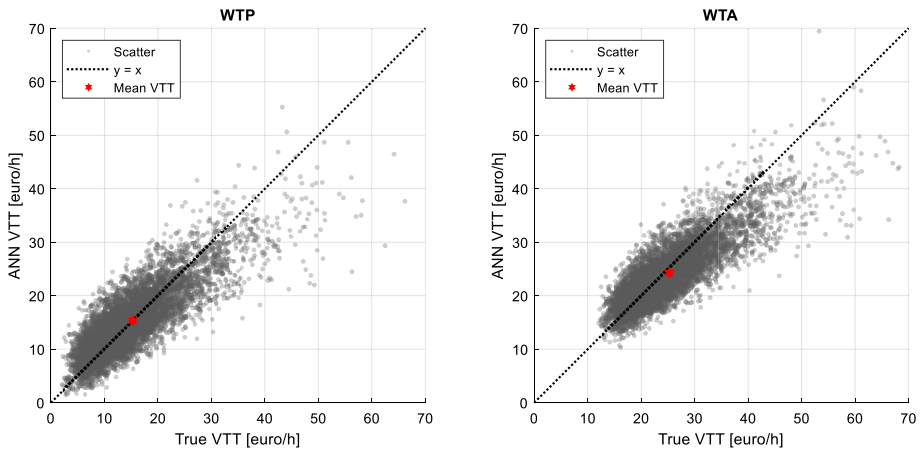


Fig. 11 Lognormal distribution: true vs recovered VTTs

has indeed been able to approximate individual level choice models f_n . To assess this, we plot the true individual choice probabilities f_n and the choice probabilities learned by the ANN in one plot for several individuals and see their similarities and differences. To do so, instead of randomly picking individuals, we select four individuals with true VTT of 5, 10, 15 and 20 euro per hour. Figure 9 shows the four choice probability plots (taken from data set no. 2). We see that although f_n and g do not coincide, the ANN does seem to approximate the individual choice models. Although these are just four snapshots, they support the notion that the BVTT at the indifference point can be used to obtain individual level estimates of the VTT.

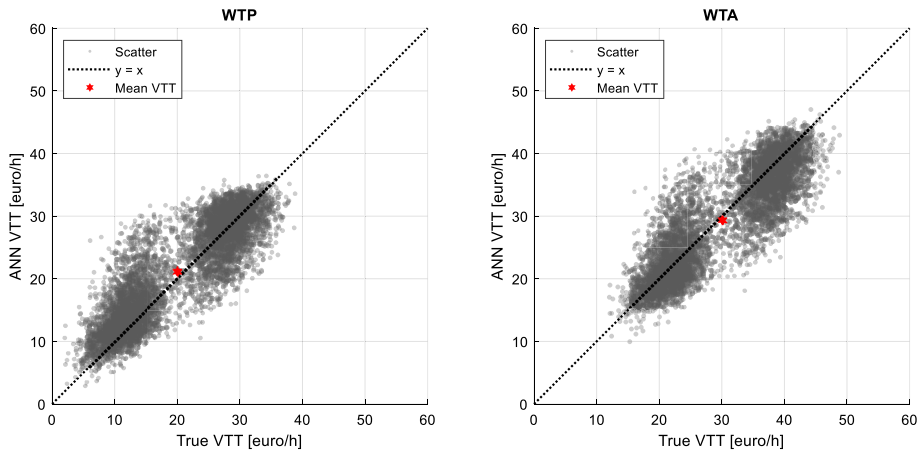


Fig. 12 Bimodal normal distribution: true vs recovered VTTs

Recovery of individual level VTTs

To assess the accurateness at which individual level VTTs are recovered, we scatter the true and recovered VTTs. While prediction accuracy at the individual level is not a common yardstick for a method developed to determine VTTs, it is still interesting to investigate this, at least from a methodological perspective. In addition, in other domains of application to which this method may spread prediction accuracy at the individual level may be highly relevant (e.g. in the context of online marketing or personalised travel advice). Figures 10, 11 and 12 show scatter plots of the true VTT (x -axis) versus the recovered VTT (y -axis) for the normal, lognormal and bimodal normal distributions (both for WTP and WTA quadrants). In case the VTTs would be perfectly recovered all points would lie on the $y=x$ line (depicted in the plots by the dashed black line). Hence, the spread around the $y=x$ line provides insights on the prediction accuracy at the individual level.

The scatter plots for all three distributions provide a consistent view. Firstly, they show that the true VTT and the recovered VTT are strongly positively correlated. Secondly, they show that individual level VTT predictions are subject to some variance, but that variance is rather homogeneously spread around the $y=x$ axis. The observation that VTT predictions is subject to some variance is fully in line with expectations. After all, the generated choices in the data are subject to random noise (see Eq. 3). Therefore, it is entirely possible that a decision maker with a low (high) VTT chooses the fast and expensive (slow and cheap) alternative a few times—with the intuitive result that the ANN overestimates (underestimates) that decision maker's VTT. The more important observation here is that the variance is *homogeneously* spread around the $y=x$ axis. This suggests the method does not systematically over- or underestimates VTTs at the individual level.

Table 4 Confusion plot (based on validation and test data)

	Target 1 (fast and expensive)	Target 2 (slow and cheap)	Σ
Output class 1 (fast and expensive)	26.7%	6.9%	79.4% <i>(positive predictive value)</i>
Output class 2 (slow and cheap)	8.3%	58.1%	87.5% <i>(negative predictive value)</i>
Σ	76.3% <i>(sensitivity)</i>	89.4% <i>(specificity)</i>	84.8% (overall accuracy)

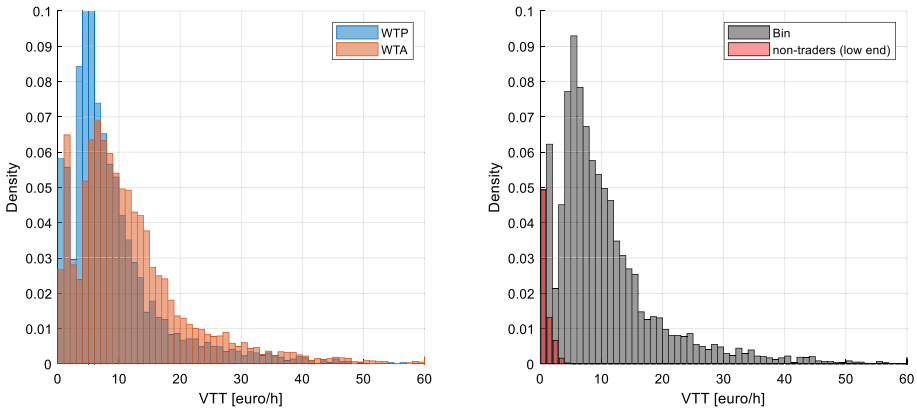


Fig. 13 WTP and WTA VTT distribution (left), reference free VTT distribution (right)

Application to real VTT data

Data and training

As mentioned before, in this study we use the Norwegian 2009 VTT data set (Ramjerdi et al. 2010).¹¹ The choice experiments are carried out by a self-administered internet-based questionnaire. Respondents of 18 years or older were recruited from an internet panel. In total, 47,000 persons were contacted, of which 9,280 completed the survey. The data were collected between June 11th and July 2nd 2009. In this study we used the data collected concerning non-business trips using car and public transport modes.

To train the network on these empirical data, we took the same approach as with the synthetic data: 70% of the data were used for training, 15% for validation and 15% for testing. The observations were randomly allocated to these subsets. We use $K=20$ randomisations (see Sect. 2.3). The trained ANN acquires a cross-entropy of 0.36 (which boils down to a ρ^2 of 0.49 based on hold-out data). Table 4 shows the confusion plot. The cells on the

¹¹ We choose to analyse this data set for a number of reasons: it contains a relatively large number of respondents, it has nine choice tasks per respondents, and the BVTs are spread across a large range.

Table 5 Mean and standard deviation of VTT

	WTP	WTA	Ref. free
Mean VTT [€/h]	10.12	13.71	11.75
Std deviation VTT	12.47	15.26	13.68
Max VTT [€/h]	140.1	118.8	123.3

diagonal show the percentage of the choices that are correctly predicted. The off-diagonal cells show the percentage of choices that are mispredicted. The confusion plot shows that overall about 85% of the choices are correctly predicted (based on highest probability).

Results

To obtain the VTT distribution, we use the network to simulate choice probabilities and search for the BVTTs that yield $P=0.5$. We do this 20 times¹² for each respondent (i.e., steps 2 to 4, see Sect. 2.3). Figure 13 shows the resulting distribution of the VTT for WTP and WTA (left) as well as a reference free VTT (right). Inspired by the work of De Borger and Fosgerau (2008), the reference free VTT is computed by taking the geometric averages across the WTP and WTA VTT at the level of the individual VTTs.¹³ The mean and the standard deviations of these distributions can be found in Table 5. Note that for eight respondents, it has not been possible to obtain a VTT estimate. For these respondents, the ANN predicts choice probabilities below 0.5, even for BVTTs close to zero, suggesting a zero or even a negative VTT. In the remainder of our analyses these eight respondents are given a VTT of zero. For another 14 respondents, the simulated choice probability crossed the $P=0.5$ point more than once. For these respondents (predominantly non-traders), we rather pragmatically used the first instance at which the $P=0.5$ point is crossed to compute the VTT. However, further research is needed to explore how to deal with these respondents when output of the method is used in utilitarian appraisal methods.

Based on Fig. 13 and Table 5 we can make a number of important observations. Firstly, Fig. 13 shows that the shape of the VTT distribution is positively skewed. The lognormal-like shape is behaviourally intuitive and has occasionally been found in previous VTT studies. However, when fitting the lognormal distribution onto these data, we find that it does not fit the data well: in particular, it cannot accommodate for the spike at around $VTT=€2/h$ and the drop at $VTT=€16/h$. Close inspection of the bins around $VTT=€2/h$ reveal that they are predominantly populated with those respondents that always choose the slow and cheap alternative (for clarity, non-traders are depicted in red in the right-hand side plot). The bimodal shape of this distribution essentially emphasises the need for flexible methods to uncover the distribution of the VTT. Secondly, in line with behavioural intuition and the findings of previous VTT studies we see that the mean WTA VTT is higher

¹² We find that after 20 times the results are stable.

¹³ Note that De Borger and Fosgerau (2008) derive that the geometric average yields a reference free VTT in the context of their behavioural framework. Our data-driven method lacks such a behavioural framework. Therefore, it is not fully clear what is the best way to obtain a reference free VTT. Nonetheless, for the purpose of this paper using the geometric average as an approximation of the reference free VTT is sufficient. In any case, the main results of this study will not change by using a different approach to obtain a reference free VTT (e.g. taking the arithmetic mean across the WTP and WTA VTT).

Table 6 Comparison of unweighted VTT estimates

	ANN-based method ^a		Norwegian VTT approach	
	Short distance	Long distance	Short distance	Long distance
Car	9.01 (0.22)	19.36 (0.50)	8.36 (0.06)	22.64 (0.28)
Public transport	6.61 (0.29)	N/A	4.80 (0.08)	N/A
Bus	N/A	12.19 (0.42)	N/A	13.72 (0.33)
Train	N/A	15.04 (0.56)	N/A	16.88 (0.33)

^aStandard errors in brackets below

than the mean WTP VTT. The difference between the two domains is, on average, €3.6/h. Thirdly, it can be seen that the tail of the WTA VTT is fatter than the tail of the WTP distribution. This is also reflected by the larger standard deviations for the WTA VTT.

Fourthly, the right-hand side tail is fully recovered, in the sense that for all respondents, including non-traders, a VTT has been recovered. To understand how the tail of the distribution is recovered by the ANN one has to keep in mind that the ANN merely learns associations between variables, just like a regression model. Thus, there is no behavioural model underlying the ANN that postulates strict relations between the explanatory variables and the dependent variable. Rather, it uses the information of all other respondents to make VTT estimates for the non-traders as well. However, it goes without saying that the tail of the distribution can only be recovered if the data allow for it. This is a prerequisite that goes for all methods. But, for this method it is particularly important since, due to the black box nature of ANNs, the analyst cannot easily judge whether the tail has been identified, or not, by e.g. looking at the ANN's weights. When using this method, the analyst needs to bear in mind that ANNs generally perform poor in extrapolation, i.e. making predictions outside the domain where they have been trained.

VTT for appraisal

Given the methodological scope of this paper, we have no intention to derive new VTT values to be used for appraisal. However, to see the method works out when used in the context of appraisal, this subsection derives mean VTTs and compares them to mean VTTs that are obtained by replicating the approach taken in the official Norwegian VTT study. A direct comparison with the official VTTs is not possible as this would require a full-fledged analysis, involving reweighting the sample to correct for skew in income, age and distance and accounting for size effects (which we have not considered in this study). As this goes beyond the scope of this paper, here we compare the unweighted VTTs instead. Note that reweighting or accounting for size effects is unlikely to substantially affect the main differences that we find in this comparison.

To obtain mean VTTs for the ANN-based method for specific categories (e.g. mode-distance combinations), we split the sample into the categories and immediately compute the mean VTT for each category. To obtain mean VTTs from replicating the Norwegian approach we have re-estimated the model using the exact same sample and explanatory

variables that we have used for the ANN-based method and using the exact same RV specification. Furthermore, rather than estimating separate choice models on segments of the data belonging (to each category), as was done in the official Norwegian VTT study, for a meaningful comparison with the ANN method we have estimated a single model in which the modes and distance categories enter the model as covariates. Finally, the Norwegian VTT study tested lognormal and semi-nonparametric distributions (as proposed in Fosgerau and Bierlaire 2007), depending on the mode and purpose. We also tested both sorts of distributions and found that on the full data set the semi-nonparametric method hardly added explanatory power over the lognormal distribution (which is much simpler in use). This result is in line with results reported by Börjesson et al. (2012), whom also tested both sorts of distributions. Therefore, henceforth we only discuss the results from the RV model with the lognormal distribution. Estimation results for this model can be found in Table 8 (column “RV model 3”). After estimation, we have computed mean VTT estimates by simulating the distributions for each respondent in the data—like is done in the Norwegian VTT study (Ramjerdi et al. 2010).

Table 6 presents the results. Two important inferences can be made. First, the VTTs obtained using the ANN-based method are consistent with the VTTs obtained from replication of the approach taken in the Norwegian VTT study. That is, the lowest VTT is consistently found to be for public transport trips, and the highest VTT is consistently found to be for long-distance car trips. Second, Table 6 shows that the ANN-based method systematically predicts higher VTTs for short distance trips and lower VTTs for long-distance trips, as compared to the Norwegian VTT approach. These differences are in theory large enough to shift outcomes in CBA policy practice.

Cross-validation

This section aims to cross-validate the ANN-based method in terms of four aspects: the shape and mean of the recovered VTT distribution (Sect. 5.1), the impact of covariates on the VTT (Sect. 5.2), the individual level VTTs (Sect. 5.3) and the goodness-of-fit (Sect. 5.4). In this section it should be kept in mind that there is no ground truth. Rather, all methods should be seen in light of their own pros and cons.

Shape and mean of the VTT distribution

To cross-validate the shape and mean of the recovered VTT distribution by the ANN-based method, we compare with state-of-the-art (parametric) choice models as well as with three (semi) nonparametric methods that have been used in recent VTT studies. The parametric models that we use in this cross-validation study are RV models, with two types of distributions, namely the lognormal and the log-uniform distributions. The lognormal distribution has been used in the most recent Swedish VTT study; the log-uniform has been used in the most recent UK VTT study. The estimation results of these models can be found in Table 8. Note that we also have estimated conventional RUM models, but the RV models are found to outperform their random utility counterparts. Therefore, we report only on the RV models. Regarding the nonparametric methods, the first nonparametric method that we consider is called local-logit. This method is developed by Fan et al. (1995), pioneered in the VTT research literature by Fosgerau (2007) and further extended by (Koster and Koster 2015). The local-logit method essentially involves estimation of logit models

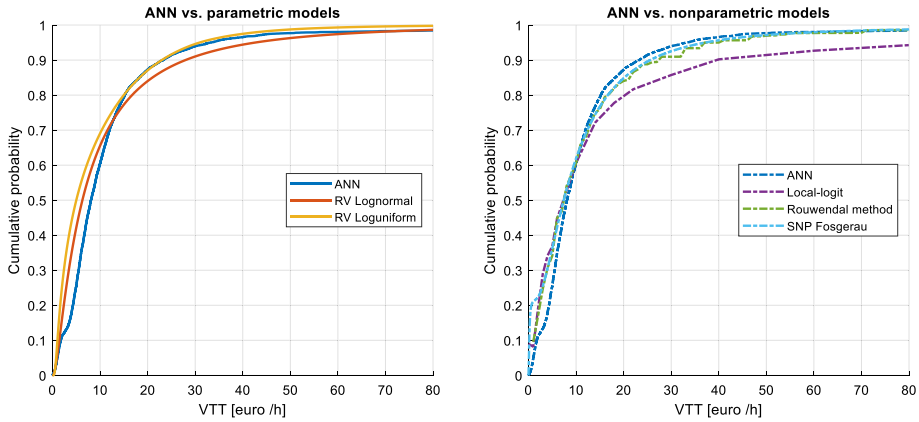


Fig. 14 Cross-validation of shape

Table 7 Mean, median and standard deviations of recovered VTT distributions

	ANN	RV lognormal	RV log-uniform	Rouwendal method	Local-logit	SNP ^a
Mean	11.75	12.13	9.34	12.51	12.16 ^b	12.34
Median	8.09	6.30	5.01	7.44	7.33	7.40
Std deviation	13.68	17.57	11.41	15.22	15.24 ^b	15.64

^aCensored at VTT = €200/h

^bUnreliable due to a large unrecovered tail of the distribution

at ‘each’ value of the BVTT using a kernel with some shape and bandwidth. In our application we use a triangular shaped kernel with a bandwidth of 10 euro. The second non-parametric method is developed by Rouwendal et al. (2010). Henceforth, we refer this method as ‘The Rouwendal method’. This method assumes that everybody has a unique VTT and makes consistent choices accordingly. But, at each choice there is a fixed probability that the decision maker makes a mistake and hence chooses the alternative that is inconsistent with his/her VTT. More details on this method are given in Appendix B. The third nonparametric method is put forward by Fosgerau and Bierlaire (2007). This is actually a semi-nonparametric method which approximates the VTT distribution using series approximations. We apply the method—which we henceforth refer to as ‘SNP’—to the RV model that we also used in the parametric case. Estimation results of this model can be found in Table 8.

The left-hand side plot in Fig. 14 shows the Cumulative Density Function (CDF) of the VTT recovered using the ANN-based method (blue) and the parametric RV models. The right-hand side plot in Fig. 14 shows, besides the CDF of the ANN VTT (blue), the CDFs created using the local-logit (orange), the Rouwendal method (green) and the SNP method (turquoise). A number of findings emerge from Fig. 14. A first general observation is that all methods roughly recover the same shape of the VTT distribution, except for the local-logit. But, there are non-trivial differences between the shapes too. Looking at the parametric methods, we see that between $VTT = \text{€}3/\text{h}$ and $VTT = \text{€}10/\text{h}$, the VTT

Table 8 Effects of covariates on the VTT

VTT distribution	Regression ANN VTTs		RV model 1		RV model 2		RV model 3		RV model 4		RV model 5	
	N/A	p val	N/A	p val	Normal	p val	Lognormal	p val	Loguniform	p val	SNP4	p val
	Est		Est		Est		Est		Est		Est	
(Intercept)	5.55	0.00										
<i>Mode</i>												
Car	0.00		0.00		0.00		0.00		0.00		0.00	
Public transport	-2.01	0.00	-3.70	0.00	-3.82	0.00	-0.50	0.00	-0.51	0.00	-0.49	0.00
Bus	-2.94	0.00	-2.25	0.00	-3.37	0.00	-0.21	0.00	-0.37	0.00	-0.23	0.00
Train	-1.69	0.01	-1.64	0.00	-2.32	0.00	-0.13	0.03	-0.28	0.00	-0.11	0.02
<i>Gender</i>												
Male	0.00		0.00		0.00		0.00		0.00		0.00	
Female	-0.65	0.06	-0.38	0.08	-0.39	0.32	-0.04	0.19	-0.12	0.00	-0.02	0.49
<i>Age</i>												
18-20	0.00		0.00		0.00		0.00		0.00		0.00	
21-35	1.89	0.14	3.17	0.00	3.22	0.02	-0.09	0.48	0.29	0.04	0.27	0.02
36-50	1.23	0.34	1.60	0.05	1.65	0.25	0.22	0.00	0.21	0.14	0.18	0.13
51-64	-0.66	0.62	-0.88	0.27	-0.86	0.55	-0.11	0.00	0.04	0.75	0.04	0.74
65+	-2.71	0.05	-3.49	0.00	-3.78	0.01	-0.35	0.00	-0.26	0.07	-0.17	0.19
<i>Purpose (short dist only)</i>												
Return home	0.00		0.00		0.00		0.00		0.00		0.00	
Commute (outward)	0.28	0.60	0.81	0.01	0.80	0.17	0.09	0.09	0.05	0.25	0.09	0.09
Purchasing (outward)	-0.26	0.70	-1.50	0.00	-1.48	0.05	-0.15	0.02	-0.07	0.25	-0.19	0.01
Other (outward)	0.19	0.73	0.43	0.19	0.38	0.52	0.04	0.44	0.01	0.81	0.09	0.11
<i>Short/long distance</i>												
Short	0.00		0.00		0.00		0.00		0.00		0.00	

Table 8 (continued)

VTT distribution	Regression ANN		RV model 1		RV model 2		RV model 3		RV model 4		RV model 5	
	N/A	p val	N/A		Normal		Lognormal		Loguniform		SNP4	
			Est	p val	Est	p val	Est	p val	Est	p val	Est	p val
Long	6.23	0.00	8.21	0.00	9.16	0.00	0.74	0.00	0.67	0.00	0.62	0.00
<i>Personal gross income</i>												
Under 300,000 NOK/year	0.00		0.00		0.00		0.00		0.00		0.00	
300,001–400,000 NOK/year	1.50	0.00	2.69	0.00	2.80	0.00	0.23	0.00	0.20	0.00	0.20	0.00
400,001–500,000 NOK/year	2.71	0.00	4.50	0.00	4.59	0.00	0.39	0.00	0.28	0.00	0.34	0.00
+500,001 NOK/year	6.51	0.00	8.47	0.00	8.86	0.00	0.71	0.00	0.61	0.00	0.58	0.00
Do not know/no answer	1.96	0.01	1.76	0.00	1.82	0.04	0.13	0.10	0.13	0.04	0.11	0.13
<i>Current trip characteristics</i>												
Travel time	-0.011	0.00	-0.009	0.00	-0.007	0.01	0.000	0.27	0.000	0.27	0.000	0.16
Travel cost	0.128	0.00	0.151	0.00	0.125	0.00	0.007	0.00	0.006	0.00	0.006	0.00
<i>Sign effect</i>												
WTA	0.00		0.00		0.00		0.00		0.00		0.00	
EL	-1.63	0.00	-1.66	0.00	-1.56	0.00	-0.14	0.00	-0.13	0.00	-0.13	0.00
EG	-1.73	0.00	-1.76	0.00	-1.78	0.00	-0.13	0.00	-0.13	0.00	-0.13	0.00
WTP	-3.91	0.00	-3.98	0.00	-3.88	0.00	-0.34	0.00	-0.31	0.00	-0.33	0.00
<i>SNP parameters</i>												
SNP1											0.21	0.02
SNP2											-0.13	0.07
SNP3											-0.68	0.00
SNP4											0.41	0.00
<i>RV scale and distribution parameters</i>												
μ (scale)	0.11	0.00	0.16	0.00	0.19	0.00	0.18	0.00	0.19	0.00	0.19	0.0

Table 8 (continued)

VTT distribution	Regression ANN VTTs		RV model 1		RV model 2		RV model 3		RV model 4		RV model 5	
	N/A	N/A	N/A	Normal	Lognormal	Loguniform	SNP4					
	Est	p val	Est	p val	Est	p val	Est	p val	Est	p val		
δ (location)	2.16	0.01	4.27	0.00	1.41	0.00	1.44	0.00	1.19			0.00
σ (standard deviation)			11.90	0.00	0.93	0.00	1.80	0.00	1.70			0.00
Number of respondents	5832		5832		5832		5832		5832			5832
No. parameters	19		24		24		24		28			28
Adjusted R-squared:	0.18											
Final LL	-26105.2		-22450.6		-22530.7		-22844.0		-22448.9			-22448.9
BIC	52460		45162		45322		45949		45202			45202
ρ^2	0.28		0.38		0.38		0.37		0.38			0.38

distribution recovered by the ANN is shifted by about 2 euros to the left. Furthermore, we see that in the tail the CDFs of the ANN and of the lognormal neatly coincide (but they do not before). The tail of the log-uniform seems to be substantially underestimated, at least as compared to the CDFs recovered using the other methods. Looking at the nonparametric methods, we see that the CDF of the Rouwendal method coincides with that of the ANN well, except between €2/h and €8/h. The CDF of the SNP method coincides well with that of the ANN for VTTs of 85/h and higher. The local-logit CDF deviates most from the other CDFs, in particular below $VTT = €30/h$. Possibly, this is caused by its inability to account for the panel nature of the data and its inability to disentangle unobserved heterogeneity from irreducible noise in the data. After all, the local-logit method only considers choices from several respondents around the same BVTT, without considering the other choices made by these (or other) respondents.

Table 7 summarises key statistics of the recovered VTT distributions for the methods that we have used. The overview shows that the mean recovered by the ANN-based method is within a €1 distance from those of all other methods, except the RV log-uniform. From the viewpoint of cost–benefit analysis,—with the exception of the VTT derived using the log-uniform distribution—for this data set it seems unlikely these relatively small differences in the recovered means would critically affect assessment results. The median VTT recovered by the ANN is higher than those of the parametric methods. This is presumably due to the limited flexibility of the latter methods to account for the substantial number of respondents having a very low VTT (13% of the respondents always choose the slow and cheap alternative), while still covering the VTT distribution over a large range. Altogether, it can be concluded that the shape, mean and median recovered by the ANN seem plausible.

The impact of covariates on the VTT

Next, we cross-validate the ANN-based method by looking at the predicted impacts of covariates on the VTT. Doing so could shed light on the relationships the ANN has learned. To do this, similar to Koster and Koster (2015), we regress socio-demographic variables and travel characteristics on the individual level VTTs recovered by the ANN-based method (dependent variable). Table 8 shows the regression results alongside with the estimation results of five RV models with the same covariates. RV model 1 does not accommodate for unobserved heterogeneity of the VTT (i.e. no distribution of the VTT), while RV models 2–4 assume the VTT is respectively normal, lognormal, and log-uniform distributed. RV model 5 is an RV model with a semi nonparametric distribution as proposed by Fosgerau and Bierlaire (2007). The parameters of RV models 1 and 2 allow for direct comparison with those of the ANN regression since they have a one-to-one relation with the location of the (mean) VTT. In contrast, the parameters of RV models 3 to 5 can not immediately be compared with those of the ANN regression, as the effect on the mean of these parameters is both a function of the location and the scale for the distribution. However, we can use the estimated parameters of these models to compute the (expected) VTTs for each individual, conditional on the covariates in the data, which in turn can be scattered against the regressed ANN VTTs. This is done in Fig. 15 for RV models 2 to 4. Importantly, the *regressed* ANN-based VTTs are used in these plots; not the VTTs as directly obtained from the ANN-based method.

First, we look at Table 8. A number of observations can be made. Firstly, looking at the ANN regression results, we see that all parameters have the expected signs and that

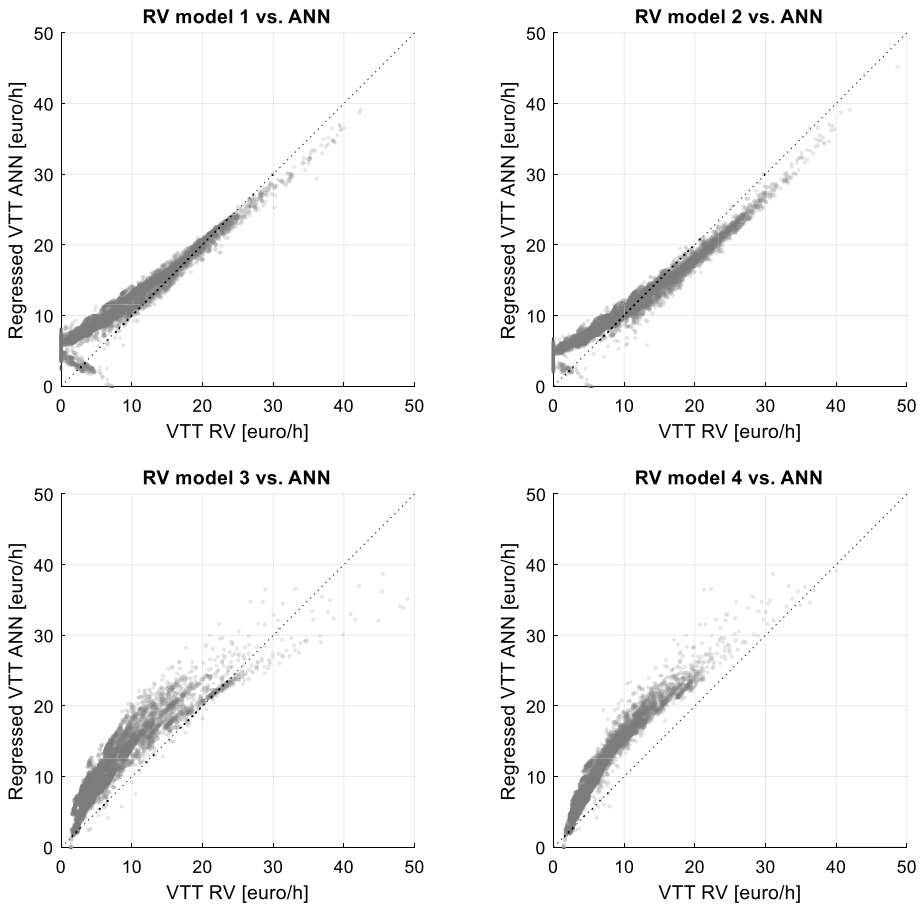


Fig. 15 Scatter plot of regressed VTTs: RV normal vs. ANN

most parameters are significant. For instance, male, long-distance trips and high levels of income all correlate positively with the VTT. Secondly, comparing the regression results with those of the RV models 1 and 2, we see that all signs are consistent across models. Thirdly, one-to-one comparison of the ANN regression parameters and RV parameters of models 1 and 2 reveals that also the relative strength of the effects are largely consistent across these models. Fourth, the RV models show that, all else being equal, the gap between the WTP and the WTA is about €3.95 per hour. This gap is close to what is found in the ANN regression and also close to what we found in Sect. 4.2. This also implies that the relatively small WTP-WTA gap compared to other studies, is intrinsic to this data set and not artefact of the ANN method.

Aside from comparing the covariates across models, it is also interesting to briefly discuss the model fits of the RV models. The BIC value shows that RV model 2 achieves the best statistical performance, taking the number of parameters into account. However, this model predicts negative VTTs for a substantial share of travellers—which seems behaviourally unrealistic. When we discard RV model 2, RV model 5 performs best in the

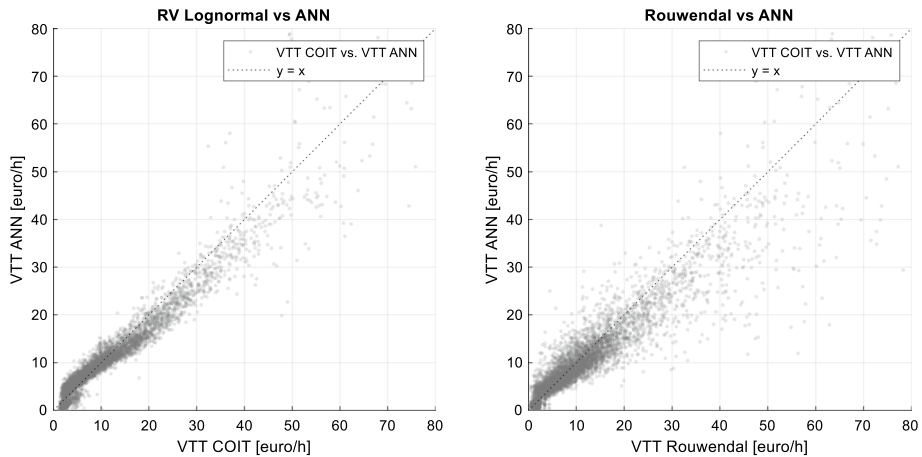


Fig. 16 Scatter plot of individual VTT: COIT vs ANN

statistical sense. But, looking at the rho squares across the RV models, we can also conclude that the imposed shape does not have a major impact on the model fit.

Now we turn to the scatter plots in Fig. 15. All scatter plots show a strong positive correlation between the VTTs as computed by applying the RV models (x -axis) and the regressed VTT of the ANN-based method (y -axis). The fact that many points scatter around $y=x$, indicates that the strength of the effects of the covariates is roughly captured equally strong in the ANN as in the RV models. All-in all, it can be concluded that the ANN-based method is able to capture the effects of covariates.

Individual level VTTs

Also within the discrete choice modelling literature methods have been devised to obtain individual level parameters (Allenby and Rossi 1998; Revelt and Train 1999; Train 2003). These methods allow estimation of individual VTTs—just like the ANN-based method does. Therefore, as a third cross-validation of the ANN-based method we apply one such a method. Specifically, we apply a method called Conditioning Of Individual Tastes [COIT] (Revelt and Train 1999). There are a number of implementations of this method, which differ from one another in subtle ways. In this study, we take the most simple approach, which considers the distribution of the population when determining the individual-level parameters as a given, and computes the expected locations of the VTT for each decision maker given this population distribution. The COIT method cannot only be applied to discrete choice models, but also to the Rouwendal method. While applying COIT to the Rouwendal method has—to the best of the authors' knowledge—never been done before, doing so provides yet another cross-validation for the individual level VTTs.

Figure 16 shows two scatter plots. The left-hand side plot shows the VTT recovered using COIT (x -axis) for RV model 3 versus the VTT recovered using the ANN-based method (y -axis); the right-hand side plot shows the VTT recovered using Rouwendal (x -axis) versus the VTT recovered using the ANN-based method (y -axis). The two plots show that the VTT points nicely scatter around the $y=x$ line. This visual observation is supported by pearson-correlation coefficients that are close to one (respectively 0.88 and

Table 9 Goodness-of-fit

	ANN	COIT RV Lognormal	COIT Rouwendal
Observations	$4 \times 1458 \times 9 = 52,488$	$5832 \times 9 = 52,488$	$5832 \times 9 = 52,488$
Null-LL	- 36,382	- 36,382	- 36,382
Final LL	- 19,108.6	- 21,690.2	- 20,338.4
ρ^2	0.48	0.40	0.44
Hit rate	0.84	0.84	0.92

0.93). This result confirms that the ANN has been able to accurately recover the individual level VTTs.

Goodness-of-fit

Finally, we compare the goodness-of-fit. Direct comparison of the goodness-of-fit across the ANN-based method and discrete choice models (Table 8) is not possible. The reason is that the ANN-based method predicts the choice probability in the hold-out choice task, *conditional* on T observed choices. In contrast, a typical choice model predicts *unconditional* choice probabilities. That is, the predicted choice probabilities are independent of other choices made by a decision maker. Therefore, statistical tests for goodness-of-fit do not immediately apply in a meaningful way.

However, we can compare the model fit of the ANN-based method with those of the COIT method. To do so, we computed the conditional VTTs based on 8 choice tasks and predict the 9th hold-out choice. This is done nine times, such that each of the nine choice tasks is predicted once, based on the other eight choices. This approach is taken for both the ANN-based method and the COIT method, except that we use four hold-out folds for the ANN based method. That is, we train the ANN based on three-fourth of the data (i.e. three folds), and predict for each decision maker in the hold out fold, the hold-out choice based on the remaining eight choices (nine times such that each of the nine choice tasks is predicted once, based on the other eight choices). By using hold-out folds for this analysis, we ensure that the results we present are not inflated due to potential overfitting of the ANN.

Table 9 shows goodness-of-fit statistics. It shows that the ANN-based method outperforms the other methods by a considerable margin in terms of log-likelihood. This result tells us that the ANN-based method has learned the underlying DGP better than the other two methods. In terms of hit rate, the ANN-based method performs on par with the RV model. The Rouwendal method achieves the best hit rate.

Conclusions and discussion

This study proposes a novel ANN-based method to study the VTT distribution. This method brings together and blends theory-driven and data-driven modelling paradigms. Specifically, in this method an ANN is used to *approximate* disaggregate choice behaviour. By doing so, the predictions of the ANN can be interpreted in light of a theory-based behavioural framework. Key advantages of this method are: (1) that it is highly flexible,

in the sense that it does not impose strong assumptions regarding the specification of the utility function, the VTT distribution, or the structure of the error terms, (2) that it can incorporate covariates, (3) that it makes use of the panel structure of the data, and (4) that it does yield a distribution right of the maximum probed BVTT. By doing so, it overcomes limitations associated with nonparametric methods that are put forward in the VTT literature (Börjesson and Eliasson 2014). Furthermore, in relation to other studies using ANNs to mode choice behaviour (c.f. Golshani et al. 2018; Sifringer et al. 2020), our approach is new in that it explicitly accounts for the panel structure often encountered in SC data. In this study we have extensively tested and cross-validated the proposed method. Based on the encouraging results of this study, we believe that there is a place for ANN-based methods in future VTT studies. In particular, we believe our method could complement theory-driven approaches for VTT inference, in particular by shedding light on the shape of the VTT distribution. Aside from our findings regarding the ANN-based method, our extensive cross-validation demonstrates the added value of using a range of parametric and nonparametric methods to investigate the VTT distribution. It helps drawing more robust conclusions, e.g. regarding the shape of the distribution.

The proposed method also has a number of drawbacks. Firstly, despite that the proposed method blends a data-driven method with theory-driven behavioural notions, it does not have an as strong theoretical base as a fully theory-driven method has. More specifically, in this study we have shown that the method can be RUM consistent, and thus has the potential to be used in economic appraisal methods grounded in utility theory, such as the CBA. A necessary condition for RUM consistency in our method is that the VTT has been uniquely determined. However, in our empirical analysis on data from the Norwegian VTT study this condition was not met for 14 respondents (out of 5832)—mostly non-traders. The practical consequences of this for recovering distribution of the VTT are perhaps minor. But, it raises deeper theoretical questions relating to its use in economic appraisal methods. For instance, does finding a handful of violations of RUM consistency in an empirical application theoretically invalidate the method as a whole for use in economic appraisal? And, what are sufficient conditions for RUM consistency for data-driven methods in general? Can we develop ex-post tests for establishing those? In absence of answers to these questions, at the very minimum this method can be used as a complementary method to learn about the shape of the distribution of the VTT, after which parametric RUM consistent discrete choice models are estimated to derive VTTs for RUM consistent appraisal. Relatedly, the issue brought to light by the non-traders in our empirical analysis also revives the longstanding question in choice behaviour modelling on how to deal with non-traders (Lancsar and Louviere 2006; Hess et al. 2010). Dealing with non-traders is especially relevant in light of the growing use of highly flexible methods in the choice modelling field. Whereas in traditional discrete choice models (e.g. a standard mixed logit model) non-traders are ‘brushed away’ as they are forced to fit a predefined mixing distribution, in more flexible models, like the one proposed in this paper, non-traders surface. This means that analysts will increasingly be confronted with issues caused by non-trading behaviour in their models, and will increasingly have to make decisions on how to treat them. Secondly, ANNs are black boxes; e.g. they cannot be diagnosed by looking at their weights. Therefore, it is difficult to understand *how* the ANN makes its predictions. One promising approach to gain a deeper understanding on the ANN in this method—which is explored by Van Cranenburgh and Kouwenhoven (2019a)—is by replacing the ANN with a conventional logistic regression model. Such a logistic regression model can be seen as a simplified, but more tractable proxy of the ANN. Therefore, this approach could provide a deeper understanding on how the ANN works in this context and shed light on how the

VTT distribution is recovered. Thirdly, as ANNs are not uniquely identifiable (Ran and Hu 2017) there is some instability of the results. Every time the network is trained (from a set of random starting weights) it will find a ‘different’ solution. Although in our case the predictions of the network turn out to be quite stable, it inherently leads to some instability of the outcomes. A possible way to deal with this is by training the network multiple times, and averaging out the variances.

The method proposed in this study provides ample scope for further research. A first direction for further research involves acquiring a better understanding regarding the data requirements for this method to work well. For instance, how many respondents are at least needed for this method? A commonly used rule-of-thumb in the machine learning field is that the number of observations needs to be at least ten times more than the number of estimable weights. However, a recent study on this topic in the context of choice data suggests a more conservative factor of 50 times more observations than weights (Alwosheel et al. 2018). Likewise, what is the ‘minimum’ number of choice tasks per respondent that is needed? In our study we found good results with nine choice tasks per respondents. But, will the method also work with just five choice tasks per respondent, or will it work even better with fifteen choice tasks? A second, related, direction for further research concerns the experimental design of the SC experiment. Current SC experiments are optimised for estimation of discrete choice models (Rose and Bliemer 2009). However, data from these experiments may actually be suboptimal for this ANN-based method. A question that remains to be answered therefore is how to design experiments optimised for this method? A third direction for further research concerns the generalisation of the method to work with choice tasks having three or more attributes. While it is clear that it becomes more difficult to recover a VTT from a choice task consisting of three or more alternatives using this method, there are—as far as we can see tell—no fundamental reasons why the method would be confined to data from two-attribute experiments only. A fourth interesting direction is investigating whether it is possible to also capture and incorporate learning and ordering effects. Some empirical studies suggest that respondents are subject to learning effects and ordering anomalies (Day and Pinto Prades 2010). A fifth research direction concerns non-traders. As for many VTT methods, non-trading behaviour could be a concern to this method. While in the data that we analysed only very few respondents (2%) always chose the fast and expensive alternative, it seems plausible that when non-traders are abundant this could jeopardise the method to work well. Further research may delve into best ways to deal with non-traders in the context of this method. A sixth direction for further research is application of this method to other VTT data sets, as well as applying the method to other areas of application, such as inference of the distribution of the value of reliability.

Author contributions Both authors have contributed to the conceptual development of the ANN-based method to uncover the VTT distribution. The ANN implementation and Monte Carlo analysis have been conducted by Sander van Cranenburgh. Both authors have contributed to the empirical study and cross-validation study. The manuscript is written by Sander van Cranenburgh. Marco Kouwenhoven has commented on previous versions of the manuscript. Both authors have read and approved the final manuscript.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not

permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Appendix A: RUM consistency

Central to RUM consistency tests are the notions of transitivity and regularity (Hess et al. 2018). A RUM consistent model satisfies both. Below we show that the method is not RUM consistent by design, but can be RUM consistent. And, we show that a necessary condition in for our method to be RUM consistent is that the VTT is uniquely determined.

Transitivity

Transitivity means that if alternative A is (weakly) preferred to alternative B and alternative B is (weakly) preferred to alternative C, then alternative A must be (weakly) preferred to C. In the context of our method, where we work with binary choice tasks and BVTTs, testing for transitivity works out as follows:

Suppose there are three alternatives {A, B, C} consisting of Travel Time (TT) and Travel Cost (TC). For simplicity of communication, we order the alternatives {A, B, C} such that $TT_A > TT_B > TT_C$, (and hence $TC_A < TC_B < TC_C$). Transitivity pertains that if $A \gg B \wedge B \gg C$, then $A \gg C$ should hold.¹⁴

$A \gg B \wedge B \gg C$ implies that $BVTT_{AB} > VTT$ and $BVTT_{BC} > VTT$. Or, in words: in the choice set {A,B} alternative A is chosen (i.e. the slow and cheap alternative). Therefore, $BVTT_{AB}$ must have been larger than the individual’s VTT. Likewise, in the choice set {B,C} alternative B is chosen. Therefore, the $BVTT_{BC}$ must have been larger than the individual’s VTT.

$A \gg C$ holds true if and only if $BVTT_{AC} > VTT$. Therefore, we thus must proof $BVTT_{AC} > VTT$. To do so, let’s express $BVTT_{AC}$ in terms of differences between in TC and TT between A and B and B and C:

$$\begin{aligned}
 TT_B &= TT_A + \Delta TT_{AB}, \quad TC_B = TC_A + \Delta TC_{AB} \\
 TT_C &= TT_B + \Delta TT_{BC}, \quad TC_C = TC_B + \Delta TC_{BC}
 \end{aligned}
 \tag{A.1}$$

$$BVTT_{AC} = -\frac{(TC_C - TC_A)}{(TT_C - TT_A)} = -\frac{(\Delta TC_{AB} + \Delta TC_{BC})}{(\Delta TT_{AB} + \Delta TT_{BC})}$$

This can be rewritten as:

¹⁴ $A \gg B$ is short for alternative A is (weakly) preferred to alternative B.

$$\begin{aligned}
 BVTT_{AC} &= -\frac{\Delta TT_{AB} \frac{\Delta TC_{AB}}{\Delta TT_{AB}} + \Delta TT_{BC} \frac{\Delta TC_{BC}}{\Delta TT_{BC}}}{\Delta TT_{AB} + \Delta TT_{BC}} \\
 &= \frac{\Delta TT_{AB} \cdot BVTT_{AB} + \Delta TT_{BC} \cdot BVTT_{BC}}{\Delta TT_{AB} + \Delta TT_{BC}}
 \end{aligned}
 \tag{A.2}$$

Since $BVTT_{AB} > VTT$ and $BVTT_{AC} > VTT$ we have

$$BVTT_{AC} > \frac{\Delta TT_{AB} VTT + \Delta TT_{BC} VTT}{\Delta TT_{AB} + \Delta TT_{BC}}
 \tag{A.3}$$

and hence, $BVTT_{AC} > VTT$.

For our method to unambiguously satisfy this final statement, the VTT has to be uniquely defined. Otherwise, we cannot execute the substitution to obtain equation A.3. This means that a necessary condition is that the indifference point $P=0.5$ is crossed once and only once for each individual. This is however not something that is (or, can be) imposed by the method. Therefore, we conclude that the method is not RUM consistent by design, but it can be RUM consistent.

Regularity

This notion entails that the probability of choosing any given alternative from a set should not increase if the offered set is expanded to include additional alternatives (Marschak, 1960). The regularity condition is, for instance, violated by models capturing context dependencies, such as Random Regret Minimisation models. Regularity as a test for RUM consistency is however not applicable for our method. Our method only operates in a binary setting. In particular, the BVTT does not exist in choice tasks with more than two alternatives.

Appendix B: Rouwendal method for binary choice data

Rouwendal et al. (2010) develop a nonparametric method to estimate the VTT and the Values-of-Statistical-Life (VOSL) from stated choice data consisting of three attributes: cost, time and safety. This method relies on two assumptions regarding choice behaviour. First, the VTT and VOSL are—for each decision maker—constant across all his or her choices. Second, for every choice there is some fixed probability the choice is consistent with the decision maker’s underlying VTT. Note that this fixed probability approach deviates from standard RUM models, in the sense that the probability of making an inconsistent choice is independent from the differences in utility across the available alternatives. In contrast, in RUM models, larger (smaller) differences in utility translate into larger (smaller) differences in choice probabilities across alternatives.

For this study we adapt the Rouwendal method. Specifically, we made a minor modification such that it works with two-attribute VTT data. Furthermore, note that in this study we aim to recover the most probable distribution of the VTT, while Rouwendal et al. (2010) focus on computing the lower and upper bound of the cumulative distributions.

Method

Suppose we have data from a classic binary VTT SC experiment, which consists of K choice observations per decision maker. In this format there is always a fast and expensive alternative and a slow and cheap alternative, and in each choice task there is an implicit price of time, referred to as the Boundary VTT (BVTT).

In case a decision maker n makes choices that are fully consistent with his own underlying VTT, then we can immediately derive a lower and upper bound for his VTT. His VTT will be higher than the highest BVTT for which the fast and expensive alternative is chosen and will be lower than the lowest BVTT for which the slow and cheap alternative is chosen.^{15,16} However, real life decisions are subject to noise. Therefore, it is assumed that a decision maker has a probability q that he makes choices that are consistent with his VTT (and hence, a probability $1-q$ that he makes choices that are inconsistent with his VTT). The probability of observing a series of K choices for decision maker n $Y^n = \{y_1^n, y_2^n, \dots, y_K^n\}$ conditional on his VTT being v is:

$$p(Y^n|v) = q^{\tau_n} \cdot (1 - q)^{K - \tau_n}$$

where τ_n is the number of choices that are consistent given v . The unconditional probability for decision maker n to make choice series Y^n can be computed by summing (or integrating) over all possible values of v :

$$p(Y^n) = \sum_v f(v) \cdot q^{\tau_n} \cdot (1 - q)^{K - \tau_n}$$

where $f(v)$ is the probability density function of the VTT. This function can be recovered, jointly with q , by maximising the likelihood of the observed series of choices of all decision makers. In our study, we divided the range of possible values of v (from €0 to €111.11) into 220 bins: 100 bins with a width of €0.22 (i.e. 2 NOK), 80 bins of €0.56 (5 NOK) and 40 bins of €1,11 (10 NOK).

References

- Abrantes, P.A.L., Wardman, M.R.: Meta-analysis of UK values of travel time: an update. *Transp. Res. A Policy Practice* **45**(1), 1–17 (2011)
- Allenby, G.M., Rossi, P.E.: Marketing models of consumer heterogeneity. *J. Econom.* **89**(1–2), 57–78 (1998)
- Alwosheel, A., van Cranenburgh, S., Chorus, C.G.: Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. *J. Choice Model* **28**, 167–182 (2018)
- Alwosheel, A., van Cranenburgh, S., Chorus, C.G.: ‘Computer says no’ is not enough: using prototypical examples to diagnose artificial neural networks for discrete choice analysis. *J. Choice Model.* **33**, 100186 (2019)
- Batley, R., Bates, J., Bliemer, M., Börjesson, M., Bourdon, J., Cabral, M.O., Chintakayala, P.K., Choudhury, C., Daly, A., Dekker, T.: New appraisal values of travel time saving and reliability in Great Britain. *Transportation* **46**, 1–39 (2017)

¹⁵ The VTT is between 0 and the lowest BVTT if always the slow and cheap alternative is chosen; the VTT is above the highest BVTT (with no upper limit) if always the fast and expensive alternative is chosen.

¹⁶ In the data from the Norwegian 2009-VTT study, 62% of the decision makers made series of choices consistent with an underlying VTT.

- Bishop, C.M.: *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford (1995)
- Börjesson, M., Eliasson, J.: Experiences from the Swedish value of time study. *Transp. Res. A Policy Practice* **59**, 144–158 (2014)
- Börjesson, M., Fosgerau, M., Algiers, S.: Catching the tail: empirical identification of the distribution of the value of travel time. *Transp. Res. A Policy Practice* **46**(2), 378–391 (2012)
- Cameron, T.A., James, M.D.: Efficient estimation methods for "closed-ended" contingent valuation surveys. *Rev. Econ. Stat.* **69**, 269–276 (1987)
- Cantarella, G.E., de Luca, S.: Multilayer feedforward networks for transportation mode choice analysis: an analysis and a comparison with random utility models. *Transp. Res. C Emerg. Technol.* **13**(2), 121–155 (2005)
- Castelvecchi, D.: Can we open the black box of AI? *Nat. News* **538**(7623), 20 (2016)
- Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002)
- Chen, C., Ma, J., Susilo, Y., Liu, Y., Wang, M.: The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transp. Res. C Emerg. Technol.* **68**, 285–299 (2016)
- Cybenko, G.: Approximation by superpositions of a sigmoidal function. *Math. Control Signals Syst. (MCSS)* **2**(4), 303–314 (1989)
- Daly, A., Tsang, F., Rohr, C.: The value of small time savings for non-business travel. *J. Transp. Econ. Policy (JTEP)* **48**(2), 205–218 (2014)
- Daly, A., Zachery, S.: Improved multiple choice models. In: Hensher, D. (ed.) *Determinants of Travel Choice*. Saxon House, Sussex (1978)
- Day, B., Pinto Prades, J.-L.: Ordering anomalies in choice experiments. *J. Environ. Econ. Manag.* **59**(3), 271–285 (2010)
- De Borger, B., Fosgerau, M.: The trade-off between money and travel time: a test of the theory of reference-dependent preferences. *J. Urban Econ.* **64**(1), 101–115 (2008)
- Erdem, T., Srinivasan, K., Amaldoss, W., Bajari, P., Che, H., Ho, T., Hutchinson, W., Katz, M., Keane, M., Meyer, R.: Theory-driven choice models. *Mark. Lett.* **16**(3–4), 225–237 (2005)
- Fan, J., Heckman, N.E., Wand, M.P.: Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Am. Stat. Assoc.* **90**(429), 141–150 (1995)
- Fosgerau, M.: Investigating the distribution of the value of travel time savings. *Transp. Res. B Methodol.* **40**(8), 688–707 (2006)
- Fosgerau, M.: Using nonparametrics to specify a model to measure the value of travel time. *Transp. Res. A Policy Practice* **41**(9), 842–856 (2007)
- Fosgerau, M., Bierlaire, M.: A practical test for the choice of mixing distribution in discrete choice models. *Transp. Res. B Methodol.* **41**(7), 784–794 (2007)
- Fosgerau, M., Bierlaire, M.: Discrete choice models with multiplicative error terms. *Transp. Res. B Methodol.* **43**(5), 494–505 (2009)
- Fosgerau, M., Hjorth, K. & Lyk-Jensen, S. V. (2007). *The Danish value of time study: Final Report*.
- Fosgerau, M., McFadden, D., Bierlaire, M.: Choice probability generating functions. *J. Choice Modell.* **8**, 1–18 (2013)
- Golshani, N., Shabanpour, R., Mahmoudifard, S.M., Derrible, S., Mohammadian, A.: Modeling travel mode and timing decisions: comparison of artificial neural networks and copula-based joint model. *Travel Behav. Soc.* **10**, 21–32 (2018)
- HCG (1998) *The second Netherlands' value of time study—final report*. Report 6098–1 The Hague
- Hess, S., Daly, A., Batley, R.: Revisiting consistency with random utility maximisation: theory and implications for practical work. *Theor. Decis.* **84**(2), 181–204 (2018)
- Hess, S., Daly, A., Dekker, T., Cabral, M.O., Batley, R.: A framework for capturing heterogeneity, heteroskedasticity, non-linearity, reference dependence and design artefacts in value of time research. *Transp. Res. B Methodol.* **96**, 126–149 (2017)
- Hess, S., Rose, J.M., Polak, J.: Non-trading, lexicographic and inconsistent behaviour in stated choice data. *Transp. Res. D Transp. Environ.* **15**(7), 405–417 (2010)
- Karlaftis, M.G., Vlahogianni, E.I.: Statistical methods versus neural networks in transportation research: differences, similarities and some insights. *Transp. Res. C Emerg. Technol.* **19**(3), 387–399 (2011)
- Koster, P.R., Koster, H.R.A.: Commuters' preferences for fast and reliable travel: a semi-parametric estimation approach. *Transp. Res. B Methodol.* **81**, 289–301 (2015)
- Kouwenhoven, M., de Jong, G.C., Koster, P., van den Berg, V.A.C., Verhoef, E.T., Bates, J., Warffemius, P.M.J.: New values of time and reliability in passenger transport in The Netherlands. *Res. Transp. Econ.* **47**, 37–49 (2014)
- Lancsar, E., Louviere, J.: Deleting 'irrational' responses from discrete choice experiments: a case of investigating or imposing preferences? *Health Econ.* **15**(8), 797–811 (2006)

- Lee, D., Derrible, S., Pereira, F.C.: Comparison of four types of artificial neural network and a multinomial logit model for travel mode choice modeling. *Transp. Res. Record.* (2018). <https://doi.org/10.1177/0361198118796971>
- Luce, R.D.: *Individual Choice Behavior: A Theoretical Analysis*. Dover Publications, Mineola (2014)
- McFadden, D.: *Econometric models of probabilistic choice*. In: *Structural analysis of discrete data with econometric applications* (1981)
- McFadden, D.L.: *Conditional logic analysis of qualitative choice behavior*. In: Zarembka, P. (ed.) *Frontiers in Econometrics*, pp. 105–142. Academic Press, New York (1974)
- Mohammadian, A., Miller, E.: Nested logit models and artificial neural networks for predicting household automobile choices: comparison of performance. *Transp. Res. Rec. J. Transp. Res. Board* **1807**, 92–100 (2002)
- Ojeda-Cabral, M., Batley, R., Hess, S.: The value of travel time: random utility versus random valuation. *Transportmet. A Transp. Sci.* **12**(3), 230–248 (2016)
- Ojeda-Cabral, M., Chorus, C.G.: Value of travel time changes: theory and simulation to understand the connection between random valuation and random utility methods. *Transp. Policy* **48**, 139–145 (2016)
- Ojeda-Cabral, M., Hess, S., Batley, R.: Understanding valuation of travel time changes: are preferences different under different stated choice design settings? *Transportation* **45**(1), 1–21 (2018)
- Omrani, H., Charif, O., Gerber, P., Awasthi, A., Trigano, P.: Prediction of individual travel mode with evidential neural network model. *Transp. Res. Rec. J. Transp. Res. Board* **2399**, 1–8 (2013)
- Paliwal, M., Kumar, U.A.: Neural networks and statistical techniques: a review of applications. *Expert Syst. Appl.* **36**(1), 2–17 (2009)
- Pereira, F.C., Rodrigues, F., Ben-Akiva, M.: Using data from the web to predict public transport arrivals under special events scenarios. *J. Intell. Transp. Syst.* **19**(3), 273–288 (2015)
- Prieto, A., Prieto, B., Ortigosa, E.M., Ros, E., Pelayo, F., Ortega, J., Rojas, I.: Neural networks: an overview of early research, current frameworks and new challenges. *Neurocomputing* **214**, 242–268 (2016)
- Ramjerdi, F., Flügel, S., Samstad, H., Killi, M. (2010) Value of time, safety and environment in passenger transport-Time, TØI report 1053-B/2010. Institute of Transport Economics (TØI), Oslo
- Ramjerdi, F., Lindqvist Dillén, J.: Gap between willingness-to-pay (WTP) and willingness-to-accept (WTA) measures of value of travel time: evidence from Norway and Sweden. *Transp. Rev.* **27**(5), 637–651 (2007)
- Ran, Z.-Y., Hu, B.-G.: Parameter identifiability in statistical machine learning: a review. *Neural Comput.* **29**(5), 1151–1203 (2017)
- Revelt, D., Train, K.: *Customer-Specific Taste Parameters and Mixed Logit*. University of California, Berkeley (1999)
- Rose, J.M., Bliemer, M.C.J.: Constructing efficient stated choice experimental designs. *Transp. Rev. A Trans. Transdiscip. J.* **29**(5), 587–617 (2009)
- Rouwendaal, J., de Blaeij, A., Rietveld, P., Verhoef, E.: The information content of a stated choice experiment: a new method and its application to the value of a statistical life. *Transp. Res. B Methodol.* **44**(1), 136–151 (2010)
- Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. *Nature* **323**(6088), 533–536 (1986)
- Shannon, C.E., Weaver, W.: *The Mathematical Theory of Communication*. University of Illinois Press, Urbana (1949)
- Sifringer, B., Lurkin, V., Alahi, A.: Enhancing discrete choice models with representation learning. *Transp. Res. Part B: Methodol.* **140**, 236–261 (2020)
- Small, K.A.: Valuation of travel time. *Econ. Transp.* **1**(1), 2–14 (2012)
- Train, K.E.: *Discrete Choice Methods with Simulation*. Cambridge University Press, New York (2003)
- Van Cranenburgh, S., Alwosheel, A.: An artificial neural network based approach to investigate travellers' decision rules. *Transp. Res. C Emerg. Technol.* **98**, 152–166 (2019)
- Van Cranenburgh, S., Kouwenhoven, M.: A logistic regression based method to uncover the Value-of-Travel-Time distribution. Working paper (2019a)
- Van Cranenburgh, S., Kouwenhoven, M.: Using artificial neural networks for recovering the value-of-travel-time distribution. In: *International Work-Conference on Artificial Neural Networks*. Springer (2019b)
- Wang, S., Mo, B., Zhao, J.: Deep neural networks for choice analysis: architecture design with alternative-specific utility functions. *Transp. Res. Part C: Emerg. Technol.* **112**, 234–251 (2020)
- Wardman, M., Chintakayala, V.P.K., de Jong, G.: Values of travel time in Europe: review and meta-analysis. *Transp. Res. A Policy Practice* **94**, 93–111 (2016)
- Wong, M., Farooq, B., Bilodeau, G.-A.: Discriminative conditional restricted Boltzmann machine for discrete choice and latent variable modelling. *J. Choice Model.* **29**, 152–168 (2017)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Sander van Cranenburgh is assistant professor in Choice modelling at the Engineering Systems and Services department of Delft University of Technology. His research focusses on discrete choice and machine learning models to study human behaviour. Sander is director of the CiTyAI Lab – which aims to bring together AI and domain knowledge on cities and transport. He holds a Ph.D. from Delft University of Technology.

Marco Kouwenhoven is researcher at the Engineering Systems and Services department of Delft University of Technology. Besides, he is Research Leader at Significance, an independent research agency focused on mobility and transport. Marco's work focuses on transport modelling and the design and analysis of stated choice experiments. He holds a Ph.D. in astronomy from Utrecht University.