Transportation Research Part C 171 (2025) 105014

Contents lists available at ScienceDirect

Transportation Research Part C

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

Incorporating domain knowledge in deep neural networks for discrete choice models[☆]

Shadi Haj-Yahia [*], Omar Mansour, Tomer Toledo

Faculty of Civil and Environmental Engineering, Technion – Israel Institute of Technology, Haifa 32000 Israel

ARTICLE INFO

Keywords:
Choice analysis
Deep neural networks
Discrete choice models
Domain knowledge
Interpretability
Machine learning

ABSTRACT

This paper explores the integration of domain knowledge into deep neural network (DNN) models to support the interpretability of travel demand predictions in the context of discrete choice models (DCMs). Traditional DCMs, formed as random utility models (RUM), are widely employed in travel demand analysis as a powerful theoretical econometric framework. But, they are often limited by subjective and simplified utility function specifications, which may not capture complex behaviors. This led to a growing interest in data-driven approaches. Due to their flexible architecture, DNNs offer a promising alternative for learning unobserved non-linear relationships in DCMs. But they are often criticized for their "black box" nature and potential deviations from established economic theory.

This paper proposes a framework that incorporates domain knowledge constraints into DNNs, guiding the models toward behaviorally realistic outcomes while retaining predictive flexibility. The framework's effectiveness is demonstrated through a synthetic dataset and an empirical study using the Swissmetro dataset. The synthetic study confirms that domain knowledge constraints enhance consistency and economic plausibility, while the Swissmetro application shows that constrained models avoid implausible outcomes, such as negative values of time, and provide stable market share predictions. The proposed approach is independent of the model structure, making it applicable on different model architectures. The methodology was applied on both standard DNN and an alternative-specific utility DNN (ASU-DNN). Although constrained models exhibit a slight reduction in predictive fit, they generalize better to unseen data and produce interpretable results. This study offers a pathway for combining the flexibility of machine learning with domain expertise for DCMs, across diverse model architectures and datasets.

1. Introduction

Travel demand predictions are essential in transportation systems modeling. Travel demand has traditionally been modeled by discrete choice models (DCMs), which are widely used to understand individuals' decision-making processes (Ben-Akiva and Lerman, 1985). Most DCMs are formed as random utility models (RUMs), which assume that individuals apply utility maximization decision protocols (McFadden, 1974). Various RUM-based models have been proposed, incorporating different assumptions about error structures (e.g., Ben-Akiva, 1973; McFadden, 1978; Revelt and Train, 1998; Train, 2009). Within these models, utility functions

[☆] This article is part of a special issue entitled: 'Data Analytics and ML' published in Transportation Research Part C.

[*] Corresponding author.

E-mail addresses: shadi8@campus.technion.ac.il (S. Haj-Yahia), omar@technion.ac.il (O. Mansour), toledo@technion.ac.il (T. Toledo).

https://doi.org/10.1016/j.trc.2025.105014
Received 30 January 2024; Received in revised form 15 November 2024; Accepted 16 January 2025

describe how different alternative attributes are valued by individuals, which varies depending on their personal characteristics and preferences. However, specifying suitable functional forms, variable transformations and interactions in the utility model is a challenging task. It has been shown that incorrectly specified utility functional forms can bias parameter estimates and the resulting predictions (Torres et al., 2011). However, the selection of variables and their functional forms within utility functions remains a subjective modeling decision.

Recently, data-driven approaches using machine learning (ML) methods to learn DCM specifications have emerged as a promising avenue to overcome the limitations of subjective specifications. ML methods, such as neural networks, decision trees, and ensemble learning can learn nonlinear mapping functions (Bishop, 2006). Specifically, deep neural networks (DNNs) (LeCun et al., 2015) have gained popularity as a data-driven approach that has shown higher prediction accuracy in various tasks. Unlike RUMs, DNN models require essentially no a-priori knowledge about the form of the underlying relationships among variables and choices, which allows them to capture complex non-linear specifications. However, the focus of their use has been primarily on individual-level prediction rather than market-level analysis.

In DNN models, input variables are passed through multiple hidden layers before reaching the output layer. The output layer has the same dimension as the number of alternatives, with each output representing a score y_j of an alternative j. These scores are then transformed into choice probabilities for alternative i using a softmax function $e^{y_i}/\sum_j e^{y_j}$. Although DNNs lack the same behavioral constraints as RUMs and do not explicitly define utility functions, the similarity of the choice probabilities expression to that of multinomial logit (MNL) models promotes their use as a flexible substitute and allows for the interpretation of scores as utilities.

Studies using DNNs for DCM mostly focused on prediction accuracy (Chang et al., 2019; Mahajan et al., 2020; Van Cranenburgh and Alwosheel, 2019). Hillel et al., (2021) systematically reviewed ML models for mode choice applications, while Wang et al., (2021) performed a large-scale comparison of ML classifiers and DCMs. In both studies, DNNs achieved the highest predictive performance, which aligns with previous studies (Hagenauer and Helbich, 2017; Hillel et al., 2021; Omrani, 2015). Temporal transferability of ML models has been investigated by Chapleau et al. (2019). They applied an ML algorithm to two large-scale household travel surveys. Their model was trained on one survey, and then applied on the other, which was conducted five years later. The results showed that the model is stable and may be transformed between time periods, when no major changes occur (e.g., in the available alternatives or infrastructure).

DNN models are often criticized for their “black box” nature, which presents challenges in interpreting the extracted relationships and limits their usefulness in providing insights into the factors that affect choice behavior (Van Cranenburgh et al., 2021). Unlike RUMs, DNN models do not explicitly reveal how input variables affect the output, and as a result, the captured relationships may deviate from prior expectations. For instance, the effects of different variables on choices may not align with a-priori expectations, such as having negative sensitivities of travel choices to costs and travel times, or the ranges of marginal rates of substitution among variables, such as values of time,

Post-hoc tools to derive economic behavioral information from DNNs and so enhance their interpretability have been proposed. Wang et al., 2020a numerically extract different economic quantities of interest from DNNs, including estimates of individual choice predictions and probabilities, market shares, social welfare, substitution patterns among alternatives, probability sensitivities, elasticities, and marginal rates of substitution. In their case study, at the aggregate level, most of these values are reasonable and provide more information (e.g., distributions as opposed to point estimates) compared to those obtained from MNL models. However, at the disaggregated level, some results, such as values of time, are counter intuitive. Friedman (2001) used partial dependence plots (PDPs), which calculate choice probabilities for every possible value of a variable for each observation, to derive choice sensitivities to the various variables. Zhao et al., (2020) used this method and visualized the results to evaluate the direction of effects and their magnitudes. The approach is intuitive and easy to implement but provides only qualitative insights and a clear interpretation only when features are uncorrelated. If this assumption is violated, the PDP calculation will include data points that are very unlikely or even infeasible. Interpretability of DNN models may also be supported by use of utility-like architectures that resemble RUMs. Wang et al., (2020b) developed an alternative-specific utility deep neural network (ASU-DNN) model. Their model maintains separate utility functions for each alternative, which are trained using separate neural networks using only the variables relevant to that alternative. Thus, the utility scores for each alternative depend only on its own attributes. Systematic heterogeneity is captured by inclusion of socio-demographic variables that first enter a separate DNN and then integrated into each of the utility functions to obtain the final outputs. This model was more readily interpreted compared to fully connected DNNs and achieved comparable or even better fit to the training and testing data. Based on this, Hernández et al. (2024) proposed a neural network with alternative-specific and shared weights, where the cost-dependent utility function has the same form across different choices, making it consistent with both RUM theory and the interchangeable nature of money, leading to economically-consistent outcomes. However, these works do not prevent learning of unreasonable effects of explanatory variables on choices.

Another approach to using DNN for DCM involves combining them with RUMs. Han et al. (2022) proposed TasteNet, a novel method to integrate a neural network with MNL as a neural-embedded DCM, where taste parameters in the utility function are learned through the neural network. Their work was the first to address imposing constraints on the taste parameters generated by the neural network according to prior knowledge (e.g., a negative coefficient of travel time) to regularize the network and obtain interpretable results. However, the final model is still constrained in a linear-in-parameter utility specification which may restrict its flexibility compared to a DNN. Sifringer et al. (2020) introduced a systematic utility function that incorporates both an interpretable and a learned part. The interpretable part is specified explicitly by the modeler, as is common with RUMs, while the learned part utilizes a DNN to extract additional utility terms from the data. The model improved the estimation log-likelihood value compared to a traditional MNL. However, a potential limitation of this approach is that the relative magnitude of the DNN term is unbounded, and its

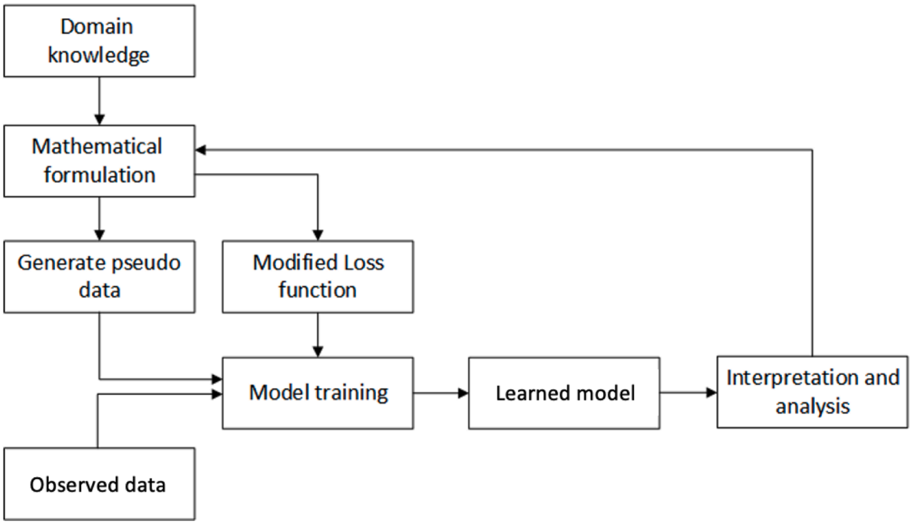


Fig. 1. Overall framework for incorporating domain knowledge into a DNN model.

relationship with the explanatory variables remains unknown. The modeler's choice of variables to include in the utility function remains subjective.

The above-mentioned methods primarily focus on interpretability and do not impose specific restrictions on the relationships between explanatory variables and choices as in the work of TasteNet (Alwosheel et al., 2019, 2021). Consequently, they may fail to ensure that the resulting non-linear relationships are controllable and consistent with established knowledge.

In practice, the estimation process of a DCM extends beyond a purely data-driven approach. When estimating a model, the focus of the modeler is not limited to achieving a high empirical fit (as indicated by the maximum likelihood objective) and statistical significance alone. Modelers also employ informal tests to ensure consistency with the domain knowledge, which includes insights that are accepted in the specific fields of study. Integrating them in models, especially data-driven ones, ensures that the outcomes also align with established theories or previous empirical findings. This may contribute to the model's interpretability and prediction performance, especially in scenarios where data is sparse or noisy.

In the context of mode choice, for example, domain knowledge may involve the directions of effects of certain factors have on choices, such as the negative impact of travel time and cost or the positive effects of accessibility, service frequency, and comfort on utilities and choice probabilities. Domain knowledge may also inform on the magnitudes and ranges of these effects, elasticities, or marginal rates of substitutions (e.g., values of time). This knowledge serves as a guiding principle for the modeler during the estimation process.

Incorporating domain knowledge into DNN models constrains them to learn patterns within accepted understandings. Domain knowledge can act as a form of regularization that helps mitigate overfitting in ML models. By incorporating domain-specific knowledge and constraints into the model, the learning process is guided towards more generalizable solutions. It also increases the trustworthiness of the predictions that are more aligned with expert knowledge (Rudin et al., 2022). Moreover, consistency with domain knowledge is crucial when the model is used to evaluate new policies in new scenarios, which require prediction beyond the fitting region. In such cases, extrapolation is required, which is another challenge area for DNN models (Aboutaleb et al., 2021). Domain knowledge guides the learning process based on established relationships or behavioral principles within the domain and thus supports making these extrapolations. Hence, the model development and training are required to balance between fit to the data and alignment with domain knowledge. Kim and Bansal (2024) have recently addressed this issue in parallel with our work. They use lattice network to ensure monotonicity of utility functions relative to selected attributes, represented as hard constraints. This special DNN architecture may face computational complexity as the lattice size and inequality constraints increase. In addition, ensuring hard monotonicity may lead to oversimplification, or constrain the model to neutralize the attributes that cause violations, as lattices are structurally rigid thereby restricting the hypothesis space significantly (Runje and Shankaranarayana, 2023).

This study aims to enhance the consistency of DNN models with domain knowledge, particularly in the context of mode choice. The proposed framework incorporates domain knowledge through the introduction of constraints to the model. These constraints are expressed mathematically and represent the behavioral realism that the model should follow. For instance, increasing travel time of an alternative should negatively affect its choice probability. As a result, a new objective function is formulated to consider both empirical fit and domain knowledge violation penalties. The proposed approach is independent of the model structure, making its implementation feasible with different model architectures.

The rest of the paper is organized as follows: the next section describes the proposed methodology for incorporating domain knowledge with different DNN architectures. The following section presents two case studies, synthetic and empirical, implementing the proposed framework. Next, the case studies' results are presented and discussed. Conclusions and discussions on potential enhancements and extensions to the proposed methodology are presented in the final section.

2. Methodology

This work aims to incorporate domain knowledge within DNNs in the context of DCMs. In the following, domain knowledge refers to the sign sensitivity of choice probabilities to the attributes of interest. For example, it is expected that increased travel time will negatively impact the perceived utility of a travel mode alternative and hence its choice probability. The proposed methodological framework involves the formulation of domain knowledge constraints, generation of synthetic pseudo-data points, and modification of the loss function to consider penalties for domain knowledge violations.

The model framework is presented in Fig. 1. A set of domain knowledge items (e.g., signs of choice sensitivities) are specified and formalized as inequality constraints on the resulting DNN model. Evaluating these constraints analytically within DNN models is challenging. To facilitate the evaluation, synthetic pseudo-data points are generated and the model predictions for these points are evaluated. The loss function is modified to incorporate terms that measure the violations of the constraints at these points to the basic data-driven loss function (e.g., negative log-likelihood). Thus, the model training process uses both observed data and the domain knowledge constraints. Interpretation of the learned models may reveal inconsistencies with additional domain knowledge items and leads to an iterative process of inclusion of new constraints and model training.

The formulation of domain knowledge constraints, generation of pseudo-data points, the modified loss function formulation, and the training process which utilizes the backpropagation algorithm for optimizing the model considering both fit and domain knowledge alignment are described as follows:

2.1. Domain knowledge constraints

Domain knowledge relates to the expected directional effects (signs) of sensitivities of choice probabilities with respect to attributes of interests. These sensitivities are calculated as the derivative of choice probabilities with respect to these attributes, as follows:

$$\frac{\partial P\left(y_j = 1|X\right)}{\partial x_{j,k}}\delta_{j,k} \geq 0 \forall (j,k) \in (J^* K^*), \forall x \tag{1}$$

Where $P\left(y_j = 1|X\right)$ is the probability of choosing a discrete alternative $j \in \{1, .., J\}$, given a vector $X = \left[x_{1,1}, \cdots, x_{j,k}, \cdots, x_{J,K}\right] \in \mathbb{R}^{K\times J}$ of K attributes for each alternative. $\delta_{j,k}$ defines the expected sign of sensitivity. It takes value 1 if it is positive, and -1 if it is negative. $J^* K^*$ is the set of combinations (j,k) of alternatives and attributes for which domain knowledge is considered.

The constraints in Equation (1) are on derivative functions and should hold for any value of X. However, with DNNs, an explicit form of this derivative function is not readily available. Instead, the constraint can be evaluated on a set of points $X_r, r \in \{1, .., R\}$. In principle, this can be done by the backpropagation algorithm. However, this approach may suffer from exploding and vanishing gradients (Wright et al., 2022). An alternative approach adopted in this study is to numerically calculate derivatives by finite differences applied on the DNN model predictions. By approximating the derivative through small changes in the input, the finite differences method helps to stabilize the penalty term. When the finite difference is sufficiently large, it helps mitigate the issue of extreme derivative values. The derivative is approximated using forward difference as follows:

$$\frac{\partial P\left(y_j = 1|X_r\right)}{\partial x_{j,k}} = \frac{P\left(y_j = 1|X_r^{j,k}\right) - P\left(y_j = 1|X_r\right)}{\varepsilon} \tag{2}$$

Where $X_r^{j,k} = \left[x_{r,1,1}, \cdots, x_{r,j,k} + \varepsilon, \cdots, x_{r,J,K}\right]$ is a perturbation of X_r in the variable k of alternative j. ε is a small perturbation value.

2.2. Loss function

The loss function used in the DNN training is adjusted to penalize violations to the domain knowledge constraints. It is composed of two parts: prediction and domain knowledge losses.

The prediction loss is defined as the negative log-likelihood of the sample to minimized:

$$\mathscr{L}_{NLL} = -\sum_{n=1}^{N}\sum_{j=1}^{J} y_{j,n} \cdot \log\left(P\left(y_{j,n} = 1|X_n\right)\right) \tag{3}$$

Where $y_{j,n} = 1$, if alterative j was chosen by individual (observation) n, and 0 otherwise.

The domain knowledge loss measures the violation of the constraints on the set of R points:

$$\mathscr{L}_{j,k}^{P} = \sum_{r=1}^{R} \max\left(0, -\frac{\partial P\left(y_j = 1|X_r\right)}{\partial x_{j,k}}\delta_{j,k}\right) \tag{4}$$

Where $\mathscr{L}_{j,k}^{P}$ is the loss penalty term for alternative j and variable k.

The total loss to be minimized is given by:

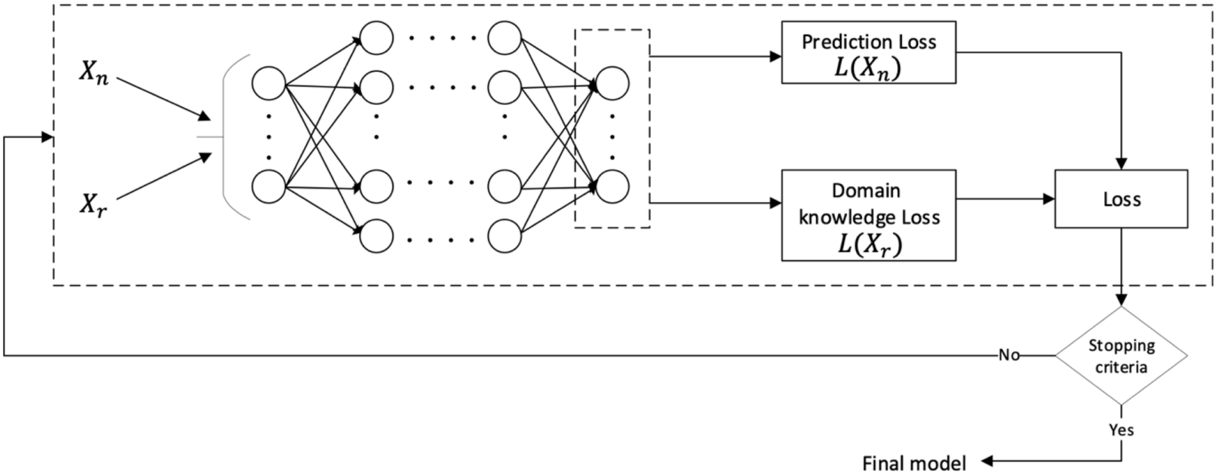


Fig. 2. Model training process.

$$\mathscr{L}_{total} = \mathscr{L}_{NLL} + \sum_{(j,k)\in J^*K^*} \lambda_{j,k} \cdot \mathscr{L}^{P}_{j,k} \tag{5}$$

Where $\lambda_{j,k}$ is the weight of each constraint violation penalty.

With this loss function, the training process is presented in Fig. 2. It involves two sets of data points: the observed data samples, X_n and the points used to evaluate the constraints X_r. They are both fed into the DNN. Each of them produces a different loss function: The prediction loss and the domain knowledge loss, respectively. Added together they make the total loss, which is iteratively used to train the DNN to convergence.

2.3. Generation of pseudo-data points

The natural candidates for the set of points X_r for the constraint evaluation are the observations in the training sample, which represent current world states. As noted above, perturbations of these points are generated for the numerical calculations of derivatives. However, it may also be useful to add synthetic pseudo-data points, which are generated specifically to evaluate the constraints violations. These pseudo-data points are only used to evaluate the derivative constraints, and therefore do not need to have associated choice labels. Their inclusion in the training process can provide several advantages:

1. Fills gaps in input attribute coverage: The range and distribution of the available training data may not be representative, with some regions within this range having few or no observations (i.e., are underrepresented). As a result, the DNN may overfit data-dense regions and inadvertently learn to ignore regions with sparse data. The use of pseudo-data generated in sparse data regions can help mitigate this risk by guiding at least the direction of the learned relationships between attributes and choice probabilities in these regions. It should be noted that in traditional DCMs, this problem is less pronounced because the modeler explicitly specifies the functional forms of utilities.
2. Enables extrapolation: Prediction beyond the observed data range is challenging for DNNs, given their tendency to interpolate based on dense data. By generating pseudo-data outside the observed range, the model is provided with additional information to learn from, that guides to predict consistently with domain knowledge even beyond the fitting region.
3. Increases dataset size: This helps the DNN model learn complex patterns, especially when observed data is scarce.

To generate pseudo-data points, there are three choices to make: the generation strategy, number of generated points, and choice of perturbation value (ε). These choices affect the evaluation of domain knowledge constraints and the tradeoff between the prediction and domain knowledge losses.

Pseudo-data generation strategy: Pseudo-data points may be drawn systematically or randomly from distributions that consider the densities of available observations in various regions of the attributes and the regions of interest for prediction. In this work, the points are generated on a grid of equally spaced points. This systematic approach allows full coverage of the attribute space and allows for evaluation of constraints across the entire domain.

Number of pseudo-data points: There is a tradeoff between domain knowledge loss and prediction loss. A larger number of pseudo points means that the importance of the domain knowledge loss increases relative to prediction loss (with similar violation penalty weights). In the generation strategy used, the number of pseudo-data points is dictated by the spacing between points.

Choice of perturbation value (ε): In finite difference approximation, the perturbation value ε is typically set to a very small value to closely resemble the analytical derivative (as shown in Fig. 3(a)). However, in DNNs, the derivative might be too large or too small. Additionally, in DCMs, modelers are often more interested in arc elasticities, which involves substantial changes in attributes like

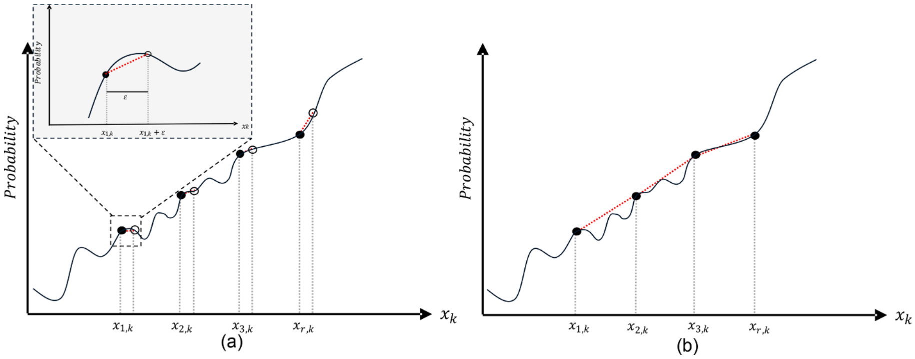


Fig. 3. Finite difference approximation to first order derivative of a choice probability of w.r.t attribute k based on pseudo-data points. (a) Presents finite approximation of the derivative with finite difference ε, and (b) presents the case where finite difference ε is the distance between the adjacent uniformly generated pseudo points.

travel time or cost, rather than very small ones. This practical consideration suggests using a perturbation value that is large enough to reflect meaningful changes in these attributes, while still being small enough to provide a reliable approximation of the derivative.

To generate data uniformly with equal distance, this distance is set to the perturbation value ε. Therefore, the resolution of the constraint is determined by the value of perturbation ε. The smaller the perturbation ε (i.e., closer approximation to numerical derivative), the larger the generated dataset size. By doing so, the finite difference approach is calculated by discretizing the continuous domain into a discrete grid, creating a piecewise linear interpolation function, providing a global description of the probability as a function of the variable rather than a series of local approximations that do not cover the entire range, as illustrated in Fig. 3(b).

3. Case studies

The proposed method is applied on a synthetic and an empirical datasets to evaluate the effectiveness of domain knowledge constraints in neural network models. By integrating known behavioral insights and domain-specific rules, we aim to assess how these constraints influence model performance, interpretability, and generalization capabilities. The synthetic case study employs Monte Carlo simulations to test a domain knowledge constrained DNN model against a standard DNN, focusing on the model's ability to capture known patterns and parameter relationships. The empirical study then extends this evaluation to real-world data using the Swissmetro dataset, examining predictive performance, behavioral consistency, and implications for transportation policy modeling.

3.1. *Synthetic case study*

Synthetic datasets generated through Monte Carlo simulations are widely used to evaluate validity and robustness of DCMs in recovering known parameters. This case study compares the performance of the proposed domain knowledge constrained DNN model (C-DNN) against a standard DNN. The focus is on the ability of the models to capture behavioral patterns accurately while adhering to domain-specific knowledge.

3.1.1. *Data generation*

The data generating process for this experiment is adopted from Kim and Bansal (2024) that extended the method originally proposed by Han et al. (2022), to introduce interactions and inherent non-linearities to the model to better reflect systematic taste heterogeneity.

The dataset is for a binary mode choice between train and metro. It includes four alternative-specific attributes: travel cost (*Cost*), travel time (*Time*), waiting time (*Wait*) and crowding (*Crowd*), and three decision-maker characteristics: income (*Inc*), full-time employment status (*Full*), and flexibility in commuting (*Flex*).

Interactions among the travel time and waiting time attributes and decision-maker characteristics represent systematic taste heterogeneity. Travel cost is modeled non-linearly to capture diminishing marginal utilities. Finally, the attributes of crowding and travel time are also interacted in the model.

The resulting systematic utility function is defined as:

$$
\begin{aligned}
&V_{jn} = -0.1 - 8 \cdot \sqrt{Cost_{jn}} - 2 \cdot Crowd_{jn} + \\
&\begin{pmatrix} -0.1 - 0.02 \cdot Crowd_{jn} - 0.5 \cdot Inc_n - 0.1 \cdot Full_n + 0.05 \cdot Flex_n \\ -0.2 \cdot Inc_n \cdot Full_n + 0.05 \cdot Inc_n \cdot Flex_n + 0.1 \cdot Full_n \cdot Flex_n \end{pmatrix} \cdot Time_{jn} + \\
&\begin{pmatrix} -0.2 - 0.8 \cdot Inc_n - 0.3 \cdot Full_n + 0.1 \cdot Flex_n \\ -0.3 \cdot Inc_n \cdot Full_n + 0.08 \cdot Inc_n \cdot Flex_n + -0.3 \cdot Full_n \cdot Flex_n \end{pmatrix} \cdot Wait_{jn}
\end{aligned}
\tag{6}
$$

Table 1
Hyperparameters search space for the synthetic Swissmetro case study.

Hyperparameter	Values
Number of hidden layers	1, 2, **3**, 4, 5
Hidden layer size	6, 8, 12, 16, 24, 32, **64**, 128
Activation function	**ReLU**, Sigmoid, Tanh
Learning rate	0.0001, 0.001, **0.01**, 0.1
Penalty weight λ	10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}, 10^{5}

Where, V_{jn} is the systematic utility of alternative j to decision-maker n.

For each synthetic observation, values of the independent variables were randomly drawn from known distributions, as detailed in Table A1 in the Appendix. The alternative utilities were computed by Equation (6). Then, choice probabilities were calculated assuming a binary logit model and used to simulate mode choices. 100 datasets with 10,000 observations in each were generated. Each dataset was randomly split into training, validation, and testing sets in a 60:20:20 ratio.

3.1.2. *Experimental design*

3.1.2.1. Models. The proposed Constrained DNN (C-DNN), which incorporates domain knowledge, was tested against a standard unconstrained DNN model. The domain knowledge constraints imposed on the C-DNN ensure that the sensitivities of choice probabilities to travel time, waiting time, and costs are non-positive:

$$\frac{\partial P\left(y_j = 1|x\right)}{\partial time_j} \leq 0, \forall j, \forall x \tag{7}$$

$$\frac{\partial P\left(y_j = 1|x\right)}{\partial wait_j} \leq 0, \forall j, \forall x \tag{8}$$

$$\frac{\partial P\left(y_j = 1|x\right)}{\partial cost_j} \leq 0, \forall j, \forall x \tag{9}$$

In this binary choice model, cross-sensitivities will by design have opposite signs to the own-sensitivities.

3.1.3. *Hyperparameter tuning*

The two models were both trained with the same hyperparameters and choices of activation functions to ensure a fair comparison and that differences in performance are due only to the domain knowledge constraints. The C-DNN model includes a penalty weight λ. To investigate its impact on fit and domain knowledge consistency, the model was trained with a wide range of values for this parameter. The search space for the hyperparameters and activation functions is shown in Table 1. The selected values are marked in bold.

3.1.4. *Evaluation metrics*

The model performance was evaluated with respect to goodness of fit and interpretability.

Goodness of fit metrics, which assess how well the model aligns with the data, were calculated both at the level of individual observations and at the market level. At the individual observation level, the average negative log-likelihood (ANLL) and classification prediction accuracy were used. For market level predictions, which are often of more interest to transportation modelers, the root mean square error (RMSE) of the model's predicted market shares compared to the observed ones were calculated. The metrics were calculated as follows.

Average negative log-likelihood (ANLL):

$$ANLL = -\frac{1}{\mathrm{N}}\sum_{n=1}^{N}\sum_{j=1}^{J} y_{j,n}\cdot\log\left(P\left(y_{j,n} = 1|X_n\right)\right) \tag{10}$$

Accuracy:

$$Accuracy = \frac{Number\ of\ correct\ predicitons}{Total\ number\ of\ predictions}\times 100[\%] \tag{11}$$

Market shares for alternative j:

Table 2
Variables in the Swissmetro dataset.

Variable	Description			
Alternative attributes		**Train**	**SM**	**Car**
Travel time	[minutes]	+	+	+
Cost	[CHF]	+	+	+
Headway	[minutes]	+	+	
Seats	Standard or Airline seats		+	
Trip characteristics				
Trip purpose	8 categories [Commute, Shopping, Business, Leisure, Return commute, Return from shopping, Return from business, Return from leisure]			
Travel class	Standard or First			
Socio-demographic characteristics				
Luggage	3 categories (None, One piece, Several pieces)			
Age	5 categories: (≤24, 25–39, 40–54, 55–65, ≥66)			
Gender	Male or Female			
Income	3 categories (≤50, 51–100, ≥101) [thousand CHF per year]			

$$MS_j = \frac{\sum_{n=1}^{N} P\left(y_{j,n} = 1 | X_n\right)}{N} \tag{12}$$

Interpretability was evaluated through the mean, standard deviation, median, minimum, maximum, and the percentage of negative WTP estimates derived from the DNN and C-DNN models against their values in the true model. WTP estimates were calculated as follows:

Value of time (VOT):

$$VOT_{jn} = \frac{\frac{\partial P\left(y_j = 1 | x_n\right)}{\partial traveltime_j}}{\frac{\partial P\left(y_j = 1 | x_n\right)}{\partial cost_j}} \tag{13}$$

Value of waiting time (VOWT):

$$VOWT_{jn} = \frac{\frac{\partial P\left(y_j = 1 | x_n\right)}{\partial waitingtime_j}}{\frac{\partial P\left(y_j = 1 | x_n\right)}{\partial cost_j}} \tag{14}$$

All the metrics were calculated separately for the training, validation, and testing datasets, for all 100 synthetic datasets.

3.2. Empirical case study

The empirical case study is applied using the Swissmetro dataset. It explores how integrating domain knowledge constraints impacts predictive performance, interpretability, and generalization on two model architectures: DNN and ASU-DNN.

3.2.1. Dataset

The openly available Swissmetro dataset (Bierlaire et al., 2001) was used. It was collected in a stated preference survey conducted in Switzerland in 1998. The survey solicited preferences among three travel modes: train, the new Swissmetro (SM) and car. The explanatory variables available in the dataset are listed in Table 2. They include level of service (LOS) attributes of the three alternatives, characteristics of the trip and socio-demographics of travelers. The categorical variables were one-hot encoded to dummy variables. Observations with unavailable alternatives, unknown attributes or outlier values were filtered, resulting in a final sample of 7,778 observations. The dataset was then split into training, validation, and testing sets in ratios of 60:20:20.

3.2.2. Experimental design

3.2.2.1. Models. Five model structures were used to assess both predictive performance and behavioral consistency:

1. Standard DNN
2. DNN with alternative specific utilities (ASU-DNN) as propose by Wang et al., (2020a).
3. Constrained (i.e., incorporating domain knowledge) DNN (C-DNN)
4. Constrained ASU-DNN (C-ASU-DNN)
5. Multinomial Logit (MNL)

Table 3
Hyperparameters search space for the empirical Swissmetro case study.

Hyperparameter	Values
Number of hidden layers	2, 4, 6, 8
Hidden layer size	6, 8, 12, 24, 32, 48, 64, 128
Activation function	ReLU, Sigmoid, Tanh
Penalty size λ	0.01, 0.1, 1, 10, 100, 10^3, 10^4, 10^5, 10^6
Perturbation size ε	0.001, 0.01, 0.1, 1, 10

The C-DNN and C-ASU-DNN models incorporated domain knowledge constraints on the signs of choice probabilities' sensitivities to travel times and costs. Own-sensitivities were constrained to be non-positive:

$$\frac{\partial P\left(y_j = 1|x\right)}{\partial traveltime_j} \leq 0, j = Train, SM, Car, \forall x \tag{15}$$

$$\frac{\partial P\left(y_j = 1|x\right)}{\partial cost_j} \leq 0, j = Train, SM, Car, \forall x \tag{16}$$

Cross-sensitivities were constrained to be non-negative:

$$\frac{\partial P(y_i = 1|x)}{\partial traveltime_j} \geq 0, i, j = Train, SM, Car, i \neq j, \forall x \tag{17}$$

$$\frac{\partial P(y_i = 1|x)}{\partial cost_j} \geq 0, i, j = Train, SM, Car, i \neq j, \forall x \tag{18}$$

The C-ASU-DNN model has separate utility functions for each alternative, which depends only on the variables related to that alternative. Therefore, only own-sensitivity constraints are required. Cross-sensitivities will by design have the opposite signs. In total there are 18 and 6 constraints, in the C-DNN and C-ASU-DNN models, respectively. They were all incorporated simultaneously in these models.

3.2.2.2. Model specifications and hyperparameter tuning. The hyperparameters of DNN models were determined through trial-and-error approach. Various combinations of optimization algorithms, learning rates, batch sizes, and regularization settings were tested to determine the best hyperparameters (See Table 3). The final models' specifications are as follows: DNN and C-DNN include two hidden layers with sizes of 48 and 64 neurons. In ASU-DNN and C-ASU-DNN models, train, SM and car neural networks consist of two hidden layers of 32 and 8 neurons. The sociodemographic network consists of two hidden layers with 24 and 6 neurons. ReLU was used for the activation function in all models.

The MNL model specification is alternative specific linear-in-parameters, with each variable entering the model independently. Interactions between variables were not considered. The model specification process was an iterative and trial-and-error procedure, aimed at achieving a model in which all parameters were statistically significant. Estimation was based only on the training set. The final model consisted of 26 statistically significant parameters (the estimation results are presented in Table A.2. in the Appendix).

3.2.2.3. Generation of Pseudo-Data for constraints. The generation of pseudo-data followed a systematic approach to ensure coverage of each constraining attribute across its range. First, for each constraining attribute (i.e., travel time and cost), pseudo samples were generated at equal intervals across its range (e.g., one-minute intervals). Next, to fill the remaining attributes in each sample, an interpolation method was used. It involved averaging the values of companion attributes from observed data samples that are adjacent in value to the pseudo sample for the corresponding attribute. This is done to approximately reflect the empirical relationships among variables within the observed data. The resulting pseudo-data sample thus combines a specified value from the constraining attribute with estimates for the remaining attributes.

3.2.3. Evaluation metrics

The performance of the five models was assessed using the same metrics defined in the synthetic study: ANLL, accuracy and RMSE of market shares (Equations 10, 11 and 12). These metrics were calculated for the training, validation, and testing datasets, offering a consistent evaluation across models and regularization levels.

3.2.3.1. Market share sensitivity analysis. To evaluate the effect of domain knowledge constraints on aggregate predictions, the estimated market shares were analyzed as functions of the explanatory variables, specifically travel times and costs. For each observation, the value of the relevant variable was varied systematically, ranging from a 50 % reduction to a 150 % increase of its observed value, in 1 % intervals, while other variables remained unchanged. This approach serves as a form of extrapolation, simulating how the model might perform on unseen scenarios beyond the scope of the original dataset.

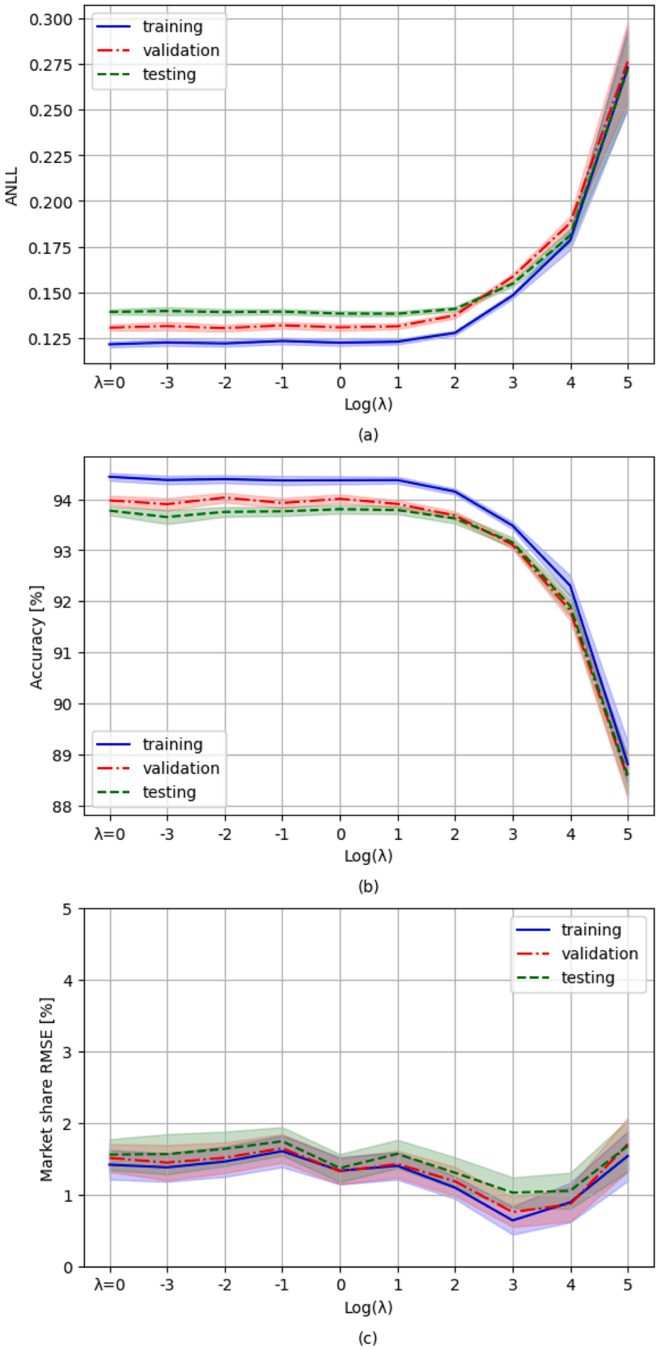
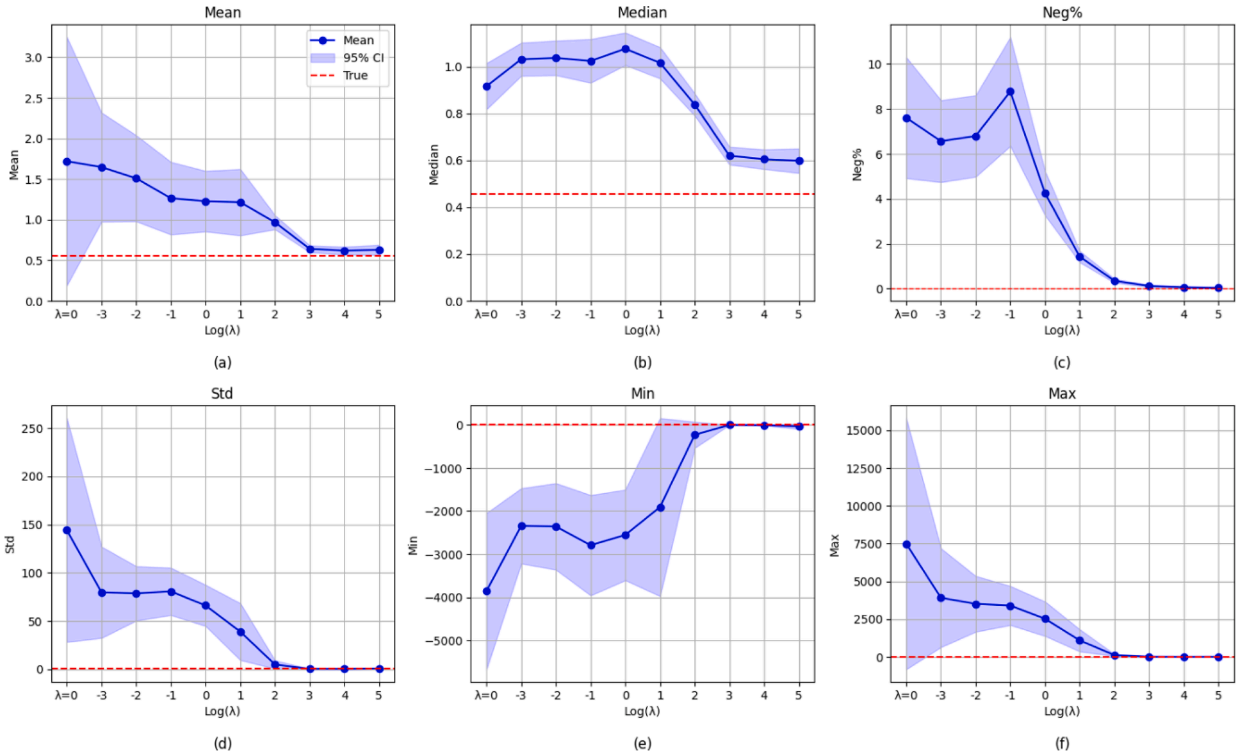


Fig. 4. Impact of the penalty weight λ on model performance.

As DCMs are often employed to forecast future policy changes (such as the introduction of new taxes, reduced travel times, or improved service levels), this method mimics the real-world application of the models in predicting future outcomes. By systematically altering key explanatory variables, we assess the robustness of the models in generalizing beyond the observed data, reflecting how they might respond to policy interventions or scenario changes.

This analysis was conducted for all six variables (two variables for each of the three travel modes), with the results averaged over all observations and plotted to illustrate the effect of constraints on market-level trends.

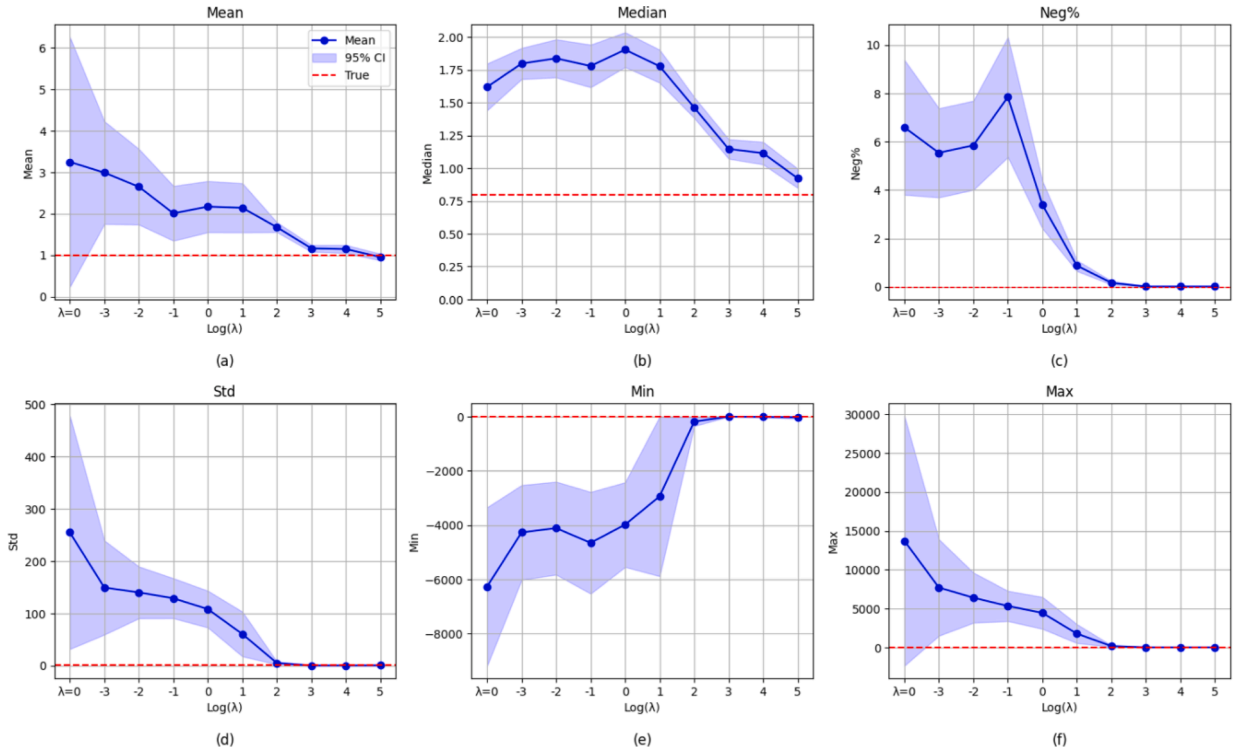
3.2.3.2. Value of time estimation. The value of time (VOT), representing travelers' willingness to pay for time savings, was also evaluated to gain insights into the behavioral consistency of each model. Unlike the synthetic case study, where the true VOT values



(a) (b) (c)

(d) (e) (f)

Fig. 5. Effect of λ on the statistics of VOT estimates.



(a) (b) (c)

(d) (e) (f)

Fig. 6. Effect of λ on the statistics of VOWT estimates.

were known by design, the Swissmetro dataset does not provide ground truth VOT estimates. Therefore, the focus of this analysis is to illustrate the variability of VOT estimates across the different models in a real-world dataset, rather than to determine a definitive correct value. The VOT for each mode and observation was calculated as:

$$VOT_{jn} = \frac{\frac{\partial P\left(y_j=1|x_n\right)}{\partial traveltime_j}}{\frac{\partial P\left(y_j=1|x_n\right)}{\partial cost_j}}, j = Train, SM, Car, \forall n \tag{19}$$

To evaluate how domain knowledge constraints influence the consistency of behavioral estimates, we report summary statistics, including the mean, median, and percentage of negative VOT values. Negative VOT estimates, which are economically implausible, highlight inconsistencies in the behavioral predictions. Distribution plots are also generated to visualize the spread of VOT estimates across models, helping to illustrate the differences in interpretation introduced by the constraints.

4. Results

This section presents the findings of the synthetic and empirical case studies, assessing model performance, interpretability, and consistency with domain knowledge. We begin by examining how various configurations of the Constrained DNN (C-DNN) model perform on synthetic data generated to incorporate realistic behavioral parameters. Key goodness-of-fit measures such as ANLL, accuracy, and RMSE of market shares are evaluated across models and penalty weight values. Interpretability is addressed by evaluating the consistency of C-DNN predictions with expected economic behavior, especially regarding WTP estimates.

Following the synthetic study, the empirical case study uses the Swissmetro dataset to evaluate goodness-of-fit and behavioral consistency of five models, including constrained and unconstrained DNNs. We assess the impact of domain knowledge constraints on model generalization, focusing on both individual-level predictions and aggregate market share trends. An in-depth sensitivity analysis of penalty weight and perturbation values reveals how model performance and consistency vary with regularization.

4.1. Synthetic case study

Fig. 4 shows the goodness of fit measures of ANLL, accuracy and market share RMSE for the C-DNN models with various λ values and the unconstrained DNN ($\lambda = 0$). Fig. 5 and Fig. 6 evaluate the consistency of the model results with the underlying domain knowledge. They present the summary statistics of estimated VOT and VOWT estimates, respectively, and compare them to the values derived from the true model, which are marked in red. In all figures, the shaded regions in the figure represent the 95 % confidence intervals, providing insights into the variability of the metrics across the generated datasets.

The model fit, captured by the ANLL and classification accuracy results, are shown in Fig. 4(a) and Fig. 4(b), respectively. They show very similar trends. All the models exhibit stable performance across training, validation, and testing datasets. For the unconstrained DNN and C-DNN models with $\lambda \leq 10^2$, the fit is almost constant and has narrow confidence intervals. However, for larger values of λ, the fit decreases sharply, accompanied by wider confidence intervals, which indicate greater variability in performance. This result suggests that stricter adherence to the domain-knowledge constraints imposed by larger λ values is excessive and reduces the model's ability to capture the underlying structure in the data. The aggregate level RMSE of market share predictions is shown in Fig. 4(c). As with the individual-level measures, the fit of the DNN and C-DNN with smaller values of λ are similar, and with relatively narrow confidence intervals. RMSE values are lowest at $\lambda = 10^3$, and increase at larger values.

These results show that the introduction of domain-knwledge constraints does not reduce its ability to fit the data, both at the individual and aggreagte levels, as long as the penalties assigned to violating these constraints are not excessive. The addition of constraints would be useful, if they would allow for better interpretability of the model predictions.

VOT and VOWT estimates calculated on the testing set are presented in Fig. 5 and Fig. 6, respectively. The statistics of both estimates behave similarly across the range of λ values. The mean estimated VOT and VOWT are shown in Fig. 5(a) and Fig. 6(a), respectively. The unconstrained DNN overestimates the true mean, while with the C-DNN model, the estimated mean VOT approaches the true value as λ increases. The estimate confidence intervals also decrease. The median estimated VOT and WOWT, shown in Fig. 5 (b) and Fig. 6(b) show similar patterns: C-DNN models produce estimates that are closer to the true ones, and with narrower confidence intervals, than the DNN model.

Negative VOT and VOWT values are considered behaviorally unrealistic. Fig. 5(c) and Fig. 6(c) show the percentage of their occurrence in the various models. They do not occur in the true model. With the DNN model, an average of 7.6 % of VOT and 6.6 % of VOWT are negative. The constraints imposed in the C-DNN are expected to prevent this result. As λ increases, the proportion of negative VOT and VOWT estimates declines, approaching zero at around $\lambda = 10^2$. The confidence intervals also narrow as λ increases, suggesting that the model not only reduces the frequency of implausible estimates but also becomes more reliable in doing so.

The remaining statistics: standard deviation, minimum and maximum, are also calculated to understand the spread of WTP estimates within each trial, and their variability among trials across the range of λ. The standard deviation of VOT and VOWT, as seen in Fig. 5(d) and Fig. 6(d), follows a decreasing pattern as λ increases, approaching the true standard deviations 0.35 and 0.66, respectively. At $\lambda = 0$, the unconstrained DNN exhibits high variability, reflected in the large standard deviations and wide confidence intervals. This suggests that, without sufficient regularization, the model produces inconsistent estimates both within and across trials. With the increase of λ, both the mean and confidence intervals of standard deviation drop, indicating that the model becomes more stable in generating consistent estimates.

Table 4
Negative log-likelihood (NLL), prediction accuracy (ACC), and RMSE of predicted market shares.

Model	Training			Validation			Testing		
	ANLL	ACC [%]	RMSE	ANLL	ACC [%]	RMSE	ANLL	ACC [%]	RMSE
DNN	**0.592**	74.2	0.008	0.644	**72.4**	0.013	**0.679**	70.0	0.011
ASU-DNN	0.600	**74.4**	**0.000**	**0.638**	**72.4**	0.007	0.686	70.8	0.013
C-DNN	0.640	71.9	0.008	0.658	71.6	**0.003**	0.688	**71.4**	**0.009**
C-ASU-DNN	0.631	73.0	0.005	0.646	71.2	0.006	0.694	70.8	**0.009**
MNL	0.708	69.3	**0.000**	0.726	67.2	0.010	0.762	66.8	0.011

The minimum and maximum WTP estimates are presented in in Fig. 5(e) and Fig. 5(f) for VOT and Fig. 6(e) and Fig. 6(f) for VOWT, respectively. They provide further insights into the impact of regularization. At low λ values, including $\lambda = 0$, the minimum and maximum estimates exhibit extreme fluctuations, with some minimum values far below zero, violating behavioral assumptions. These fluctuations are also reflected in the wide confidence intervals, suggesting considerable variability in the unconstrained model's predictions. As λ increases, these extreme values are progressively reduced, and the confidence intervals narrow, converging toward more realistic ranges that align with domain knowledge. These findings suggest that moderate levels of regularization not only guide the model toward more interpretable solutions but also reduce variability, yielding more reliable estimates.

4.2. *Empirical case study*

4.2.1. *Goodness of fit*

The results of the goodness of fit analysis, shown in Table 4, highlight important trade-offs between model flexibility, predictive performance, and generalization to unseen data. The unconstrained DNN achieves the best fit to the training data, reflected in the lowest ANLL and highest accuracy. Similarly, the ASU-DNN model performs well, though slightly below the DNN, while the constrained models (C-DNN and C-ASU-DNN) exhibit higher ANLL values and slightly lower accuracies on the training set. As expected, the MNL model shows the weakest fit to the training data, which aligns with its more restrictive structure and simpler utility functions.

However, better fit to the training data does not necessarily translate to better performance on unseen scenarios, as demonstrated by the models' behavior on the validation and testing sets. Specifically, the DNN and ASU-DNN exhibit notable drops in ANLL and accuarcy between the training and testing sets, with the DNN and ASU-DNN losing 0.087 and 0.086 in ANLL and 4.2 % and 3.6 % in accuracy, respectively. This suggests overfitting to the training data, limiting these models' ability to generalize effectively. In contrast, the constrained models show more consistent performance across datasets, with smaller differences between training and testing results. The C-DNN and C-ASU-DNN lose 0.048 and 0.063 in ANLL, respectively, with corresponding accuracy drops of 0.5 % and 2.2 %. This suggests that introducing domain knowledge constraints mitigates overfitting, resulting in better generalization to unseen data. The MNL model also shows limited performance degradation across datasets but remains inferior to all DNN-based models in both ANLL and accuracy.

The RMSE of market share predictions, emphasize these trends. While the unconstrained models outperform the constrained ones on the training set, the results reverse on the validation and testing sets, where C-DNN and C-ASU-DNN provide the most accurate market share predictions. The constrained models achieve lower RMSE values on unseen data, indicating that they generalize better at the aggregate level as well. The MNL model, achieves perfect market share predictions on the training set, however, this is a mathematical property of the model, and not an indication to the model quality (Ben-Akiva and Lerman, 1985). It performs poorly on validation and testing sets. This highlights the MNL's limited flexibility in adapting to new data, a key limitation compared to the more data-driven neural network models.

4.2.2. *Interpretability analysis*

4.2.2.1. Market share sensitivity to travel time and cost. This analysis evaluates the market share predictions of the five models under various scenarios by systematically varying travel times and costs across their ranges. The goal of this analysis is to assess how well the models adhere to domain knowledge, which suggests that an increase in the cost or travel time of an alternative should reduce its market share while increasing the shares of other competing modes. This sensitivity analysis reflects the practical use of DCMs in predicting unseen future scenarios, such as fare changes or service improvements.

The analysis involves computing market shares at 101 discrete points, corresponding to travel time and cost changes ranging from 50 % reductions to 150 % increases, in 1 % intervals. Fig. 7 –Fig. 9 present the market share responses to cost changes for each mode, while the results of travel time variations are included in the Appendix.

Fig. 7 shows the predicted market shares as a function of train cost. The DNN model (Fig. 7(a)) shows inconsistencies with domain knowledge: as train costs decrease, the share of SM also decreases, which contradicts expectations. Similarly, car shares unexpectedly decline when train costs increase. These inconsistencies are mitigated in the C-DNN model (Fig. 7(b)), which behaves in line with domain expectations. Both ASU-DNN variants (Fig. 7(c) and Fig. 7(d)), also adhere to domain knowledge and maintaining IIA substitution patterns, similar to the MNL model (Fig. 7(e)).

Similar results are seen in Fig. 8, which shows the predicted market shares as a function of SM cost. Unexpected trends are seen in the DNN predictions in Fig. 8(a): In the range of 50 % to 80 % SM cost, the SM share increases as its cost is increased, while the car

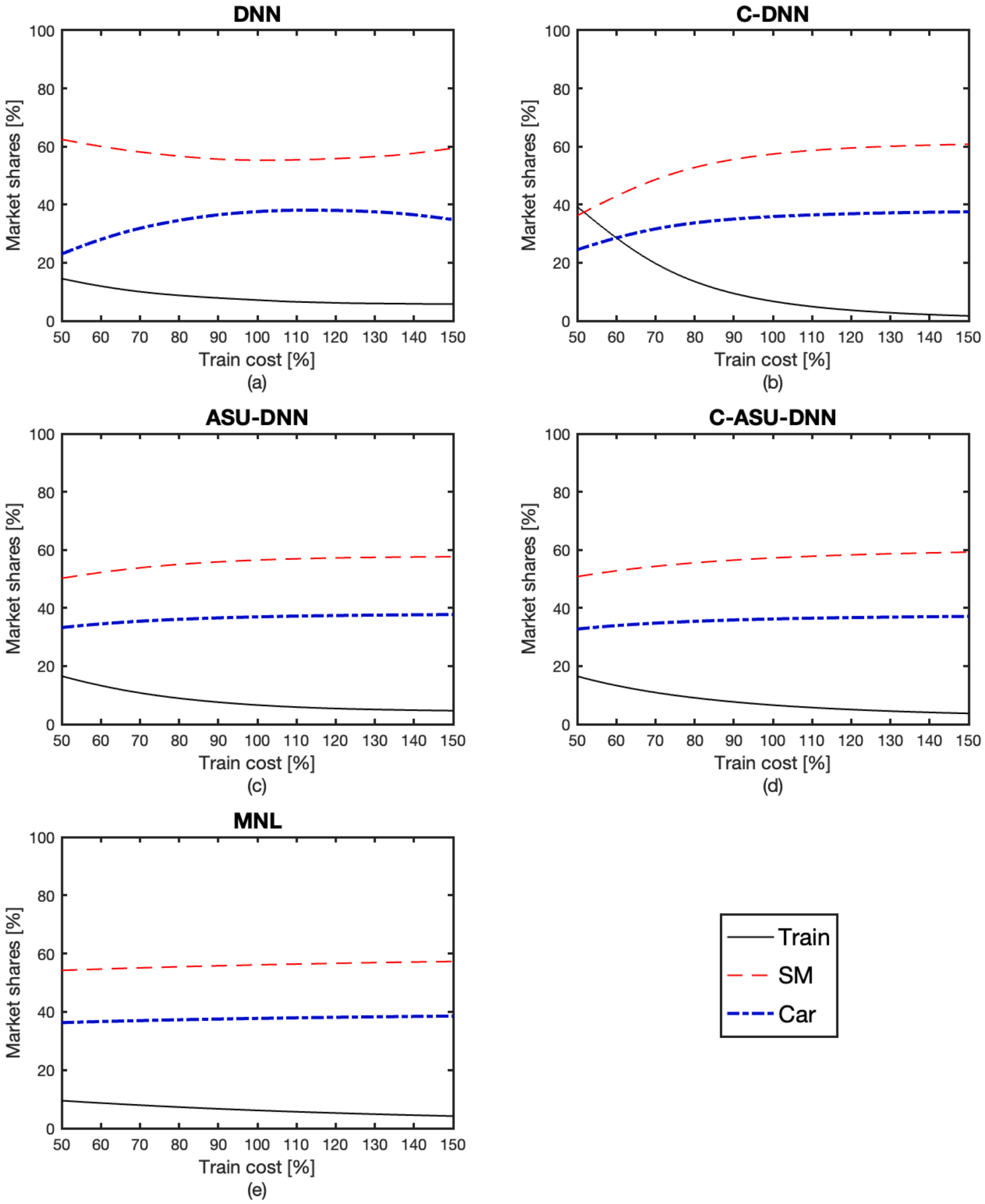


Fig. 7. Travel mode market shares as a function of train cost.

market share decreases. For the train alternative, however, market shares changes are consistent with expectations. Although very small, train shares are increasing across the whole range. This trend is absent in C-DNN (Fig. 8(b)), ASU-DNN variants (Fig. 8(c) and Fig. 8(d)), and the MNL model (Fig. 8(e)), which align with domain expectations.

Fig. 9 presents a case where the introduction of ASU-DNN is not sufficient to satisfy the expected directions of effects. It shows the

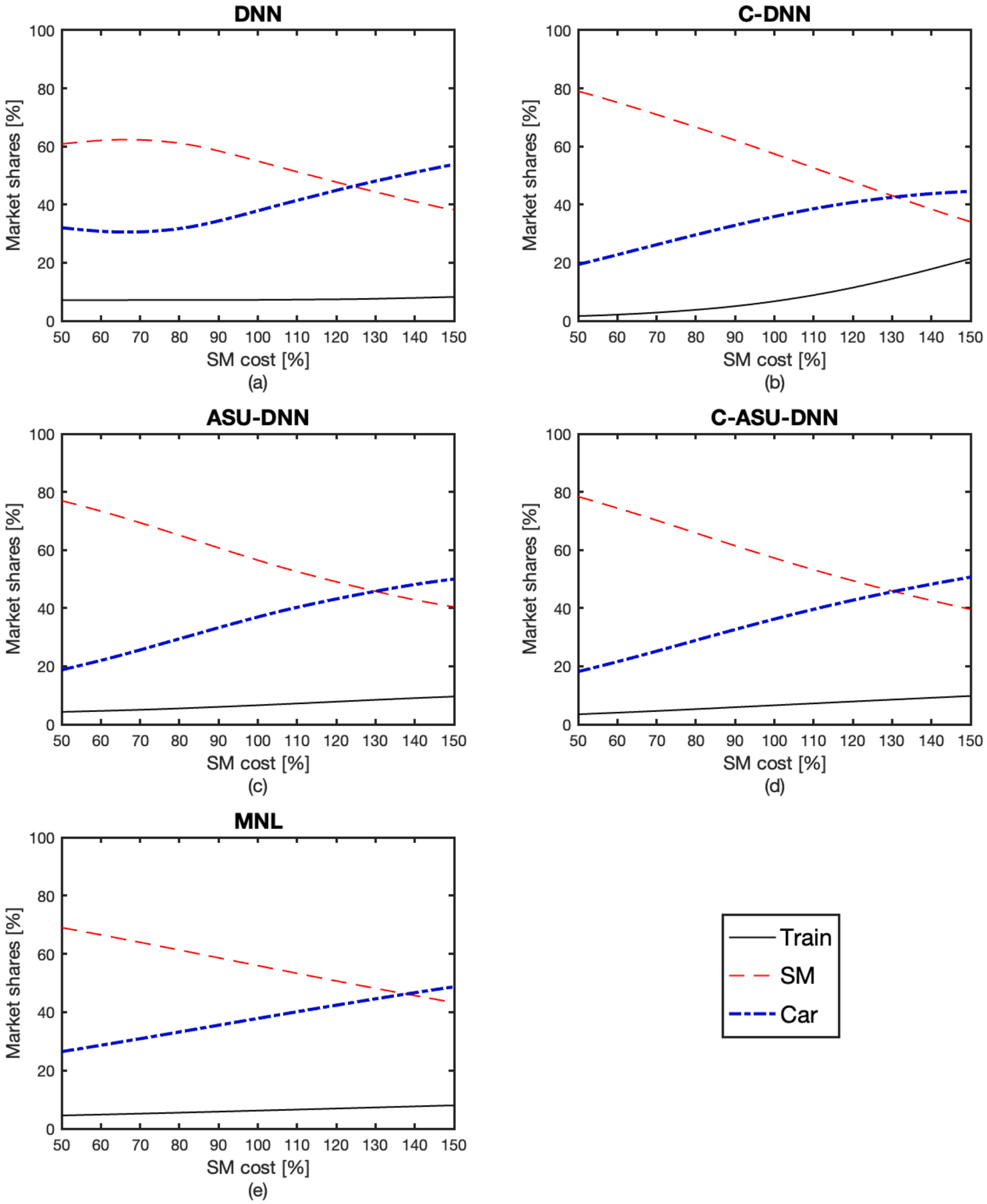


Fig. 8. Travel mode market shares as a function of SM cost.

predicted travel mode market shares as a function of car travel cost. Both the DNN (Fig. 9(a)) and ASU-DNN (Fig. 9(c)) models predict opposite directions of car and SM market shares to domain knowledge in the lower range of car costs: When the car travel costs are lower than their current values by 10 % or more, car market shares increase with higher car cost, and SM market shares decrease. This reversed direction of effect is also evident in the ASU-DNN model. Both constrained models, C-DNN (Fig. 9(b)) and C-ASU-DNN (Fig. 9

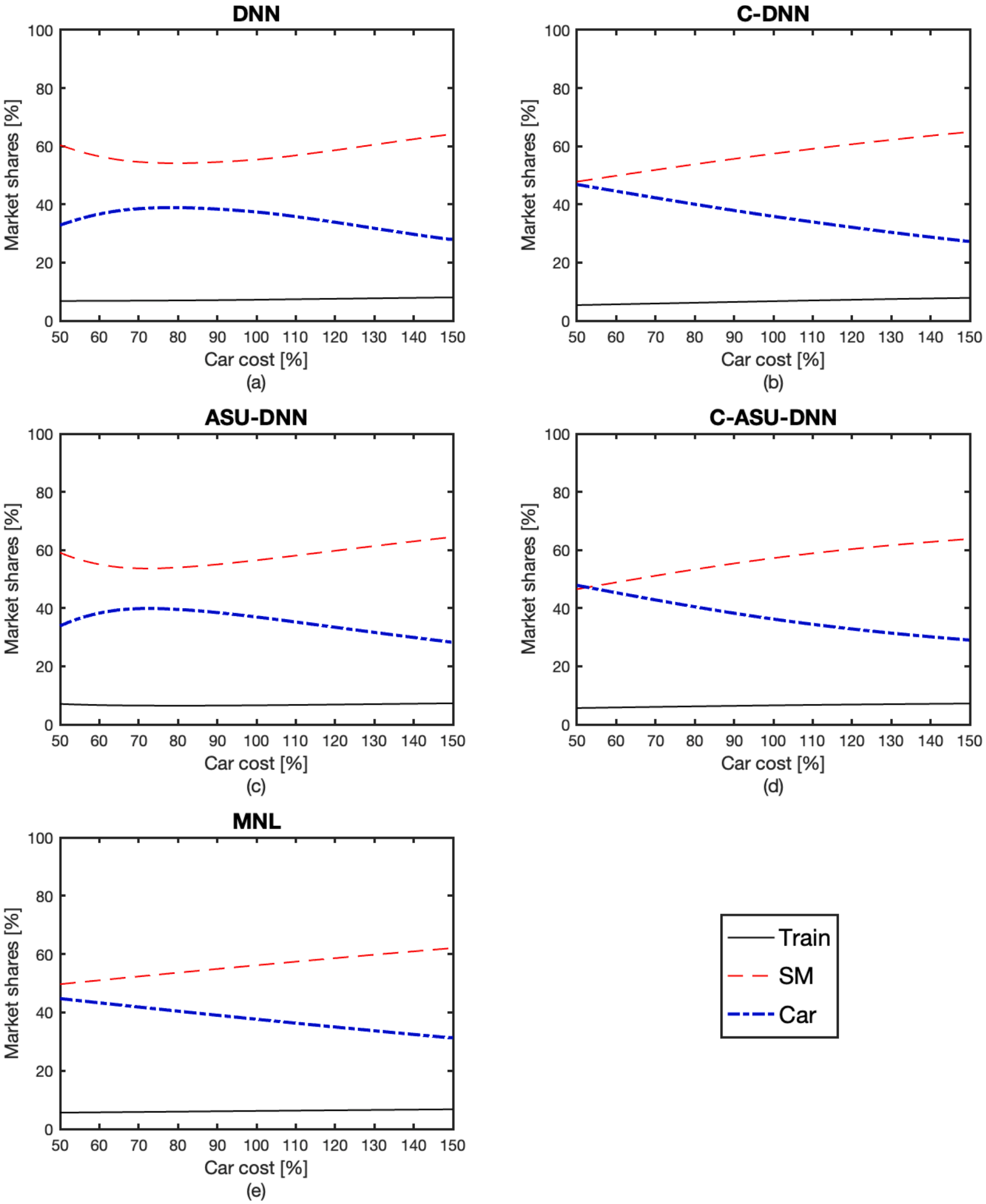


Fig. 9. Travel mode market shares as a function of car cost.

(d)), as well as the MNL model (Fig. 9(e)) exhibit the expected directions of effects in the entire range of car travel costs.

The results described above suggest that, at an aggregate level, many of the inconsistencies observed in model predictions can be mitigated by the structurally restricted ASU-DNN. This model, by design, limits the nature of the relationships it can learn, potentially preventing inconsistencies. It exhibits counter expected changes to market shares when car costs change (Fig. 9(c)) but performs as

Table 5
Percentage [%] of observations where the signs of the sensitivities of predicted mode choice probabilities with respect to changes in travel times and costs are inconsistent with domain knowledge.

	Mode	DNN	ASU-DNN	C-DNN	C-ASU-DNN	MNL
Train time	Train	15.2	0.9	0.0	0.0	0.0
	SM	54.5	0.9	8.2	0.0	0.0
	Car	29.3	0.9	7.7	0.0	0.0
Train cost	Train	2.0	3.1	0.0	0.0	0.0
	SM	29.6	3.1	8.2	0.0	0.0
	Car	47.8	3.1	7.4	0.0	0.0
SM time	Train	56.3	5.6	4.4	0.0	0.0
	SM	23.5	5.6	0.0	0.0	0.0
	Car	26.3	5.6	0.5	0.0	0.0
SM cost	Train	21.5	8.1	1.4	0.0	0.0
	SM	20.6	8.1	1.2	0.0	0.0
	Car	31.1	8.1	9.2	0.0	0.0
Car time	Train	35.9	12.5	4.4	0.2	0.0
	SM	18.2	12.5	1.9	0.2	0.0
	Car	12.2	12.5	0.0	0.2	0.0
Car cost	Train	37.1	12.2	4.6	0.0	0.0
	SM	28.8	12.2	2.2	0.0	0.0
	Car	22.5	12.2	0.1	0.0	0.0

expected when the train or SM costs change. However, when these results are examined at a more disaggregated level, further issues become evident. To see this, the derivatives of the mode choice probabilities with respect to travel time and cost variables were calculated for every observation in the sample. This was done at the current levels of the variable of interest, as well as at different values of these variables that vary in increments of 1 % between 50 % to 150 % of their current value (i.e., 101 points for each variable on each observation). The calculated derivatives are considered inconsistent with domain knowledge when their signs are opposite to expectations. These derivatives provide insight into the models' sensitivity to explanatory variables, offering a more detailed view of how well the models adhere to domain knowledge constraints across different scenarios.

Table 5 presents the percentage of observations where the predicted choice probabilities exhibit inconsistent signs with respect to travel times and costs. Inconsistent own-sensisitivities are those where an increase in travel time or cost results in a positive derivative for the corresponding alternative (shaded in grey), as described Equations (6) and (7). Inconsistent cross- sensitivities occur negative derivatives for the remaining alternatives, as described in Equations (8) and (9). In the ASU-DNN models and MNL, own- and cross-sensitivities' inconsistencies are identical due to separate utility functions based on alternative-specific variables.

The results show that the DNN model exhibits the highest percentages of inconsistencies across all sensitivities, reflecting its tendency to produce behaviorally implausible results. For example, the DNN model shows 15.2 % inconsistencies for train choice probability with respect to train time, 54.5 % for SM choice probability, and 29.3 % for car choice probability. ASU-DNN model also exhibits descent inconsistencies, particularly with respect to car travel time and cost, with 12.5 % and 12.2 % respectively.

On the other hand, domain knowledge-constrained models, show significantly lower inconsistencies compared to their counterparts. Particularly, the C-ASU-DNN model presents almost no inconsistencies, demonstrating effectiveness in maintaining consistency with domain knowledge, with 0.0 % across all sensitivities except for choice probabilities with respect to car travel time with 0.2 % violations. As expected, the MNL model shows perfect alignment to domain knowledge, as travel time and cost parameters have negative signs.

The findings from this analysis demonstrate the importance of domain knowledge constraints in ensuring consistent behavior across scenarios. While unconstrained models (DNN and ASU-DNN) provide more flexible representations, they are prone to inconsistent predictions, especially when tested on unseen scenarios. The C-DNN and C-ASU-DNN models, by contrast, exhibit robust behavioral patterns that align with domain knowledge, making them more reliable for real-world applications where predictive consistency is critical.

These results also highlight that ASU-DNN structure alone is insufficient to fully prevent behavioral inconsistencies. Instead, constraining models with domain knowledge ensures that predictions follow expected economic behavior across a range of scenarios. The analysis underscores the value of constraints not only in improving generalization, but also in maintaining interpretability, making constrained models better suited for decision-making and policy analysis.

4.2.2.2. Values of time analysis. The VOT analysis evaluates the models' predictions of travelers' willingness to pay for time savings, a critical metric in transportation modeling. Unlike the synthetic case study, where true VOT values were known and used for validation,

Table 6
Mean, median, and percentage of negatives values of time (VOT).

	Value of Time	DNN	ASU-DNN	C-DNN	C-ASU-DNN	MNL
Train	Mean [CHF/hour]	14.4	42.2	30.4	173.6	118.0
	Median [CHF/hour]	23.2	25.8	26.8	23.6	118.0
	Negative [%]	15.6	4.1	0.0	0.0	0.0
SM	Mean [CHF/hour]	23.9	61.9	52.6	251.2	84.5
	Median [CHF/hour]	20.5	23.5	39.8	41.9	84.5
	Negative [%]	31.4	8.9	0.6	0.0	0.0
Car	Mean [CHF/hour]	32.1	92.9	109.4	113.0	116.4
	Median [CHF/hour]	47.6	106.8	91.8	80.8	116.4
	Negative [%]	23.9	13.8	0.1	0.1	0.0

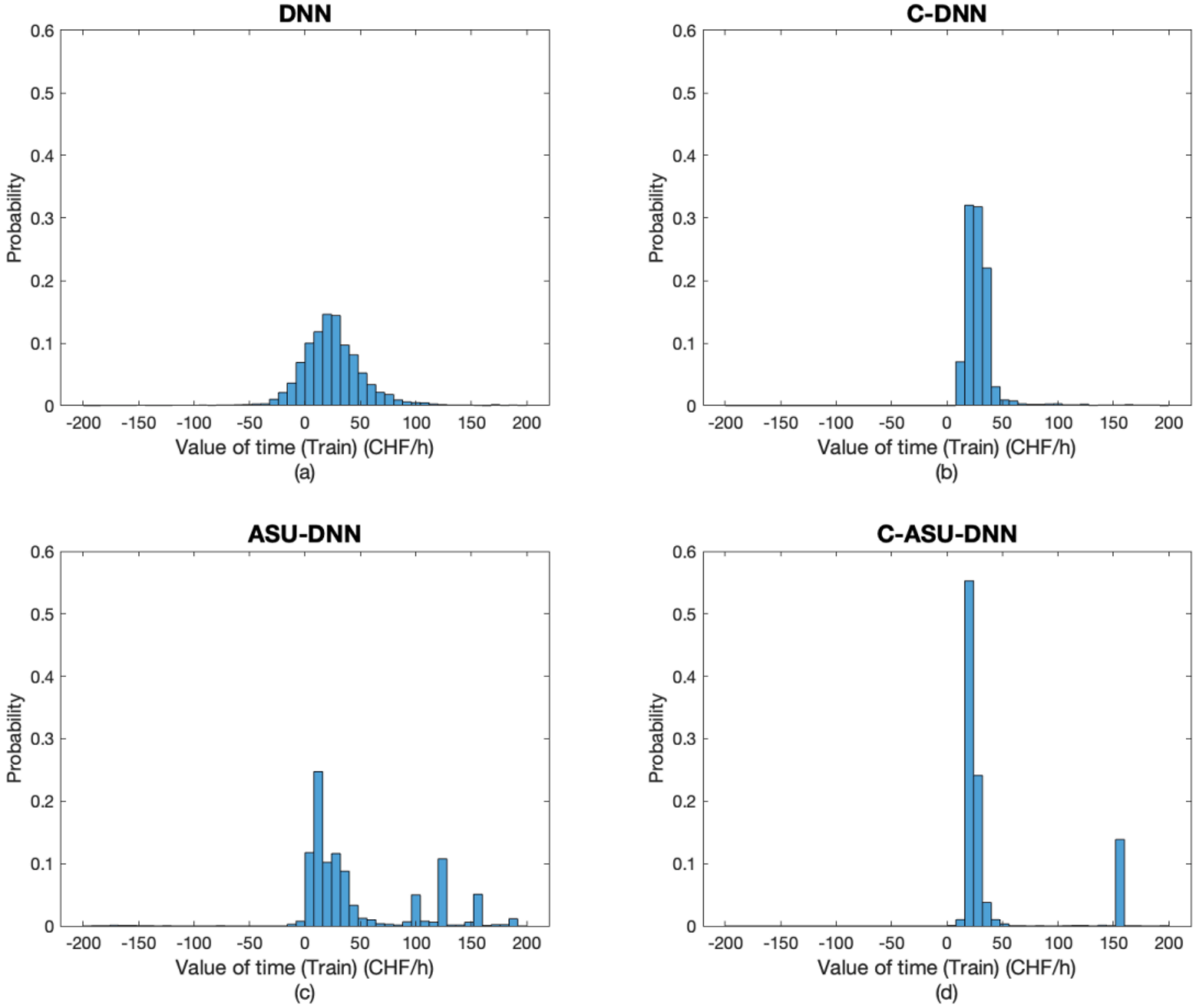


Fig. 10. Distributions of train values of time. Extreme values were cut off.

the Swissmetro data does not provide ground truth VOT estimates. Thus, the analysis focuses on the distributional properties of VOT estimates, comparing the mean, median, and percentage of negative values across the models. Specifically, the goal is to highlight the variability in VOT estimates and demonstrate how domain knowledge constraints help mitigate behavioral inconsistencies, such as negative VOTs, which indicate unrealistic predictions.

Table 6 presents the mean, median, and percentage of negative VOT estimates for each travel mode across the five models. Figs. 10-

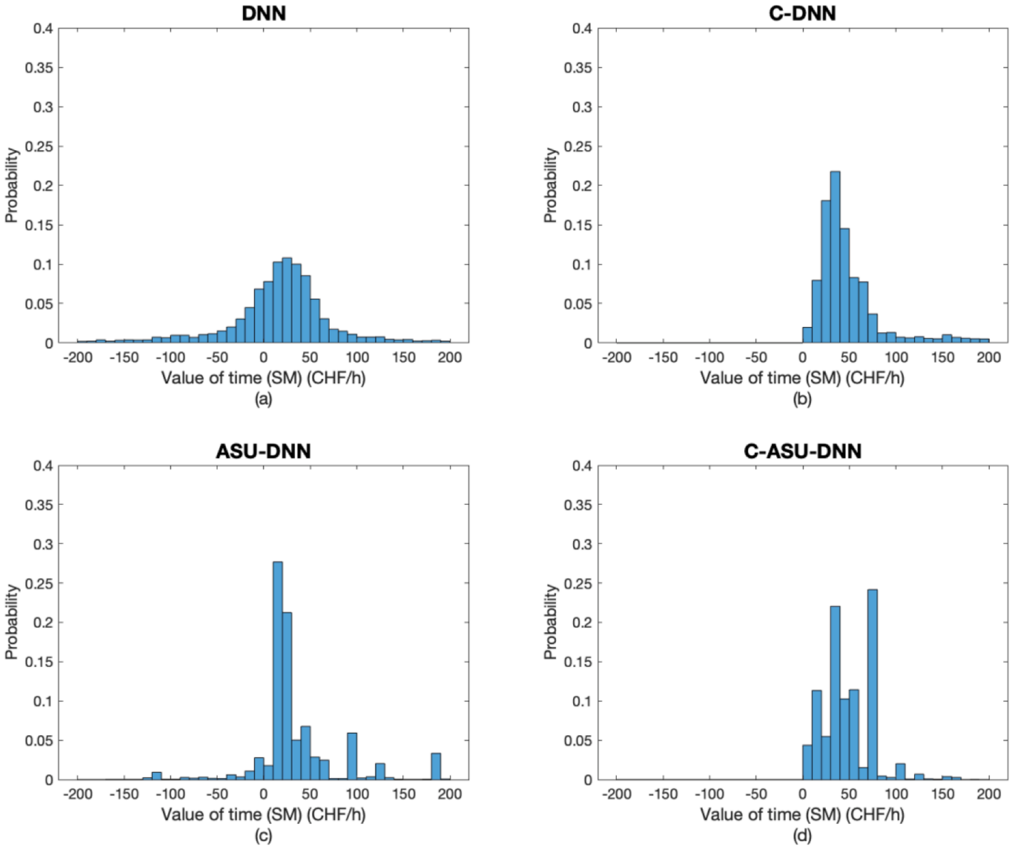


Fig. 11. Distributions of SM values of time. Extreme values were cut off.

12 show the VOT distributions across observations for the train, SM and car alternatives, respectively. The MNL model does not assume distributed VOTs but estimates a single value for each alternative.

In previous studies, Bierlaire et al. (2001) analyzed the Swissmetro data and found VOTs around 70 CHF/Hour. They did not distinguish VOTs between the various travel modes. Other studies in Switzerland reported VOTs in the range of 10–40 CHF/Hour, with higher values for car travel compared to public transportation (e.g., König et al., 2003; Hess et al., 2008).

Mean VOT values vary considerably among the different models and among alternatives within the same model. They also differ significantly from the respective median VOT values. This indicates that they are affected by large outlier values that are produced in cases where the effect of the travel cost approaches zero. Therefore, median VOT are more informative. The DNN model has the lowest median VOTs for all travel modes. This potentially results from the large percentage of negative VOTs that it produces (15.6 %, 31.4 % and 23.9 % for train, SM and car modes, respectively), which clearly contradicts expectations. The ASU-DNN model also produces substantial, although lower, percentages of negative values (4.1 %, 8.9 % and 13.8 %). VOTs for the train and SM modes are similar, but the car VOT is more than four times larger. much larger.

The two constrained models, C-DNN and C-ASU-DNN eliminate negative VOT occurrences almost entirely. This is a direct result of the introduction of constraints, whose satisfaction guarantees non-negative VOTs. The median values estimated by the two models are similar to each other and differ among the alternatives: They are lowest for the train, almost doubled for the Swissmetro and tripled for car travel.

The MNL model yields the highest VOT values, which are also higher than those reported in any previous study. This may be a result of misspecification of the utilities as linear functions, a problem that the data-driven DNNs can overcome.

The VOT values for the DNN model are approximately normally distributed with large variance for each of the travel modes. With the other models, the values are more narrowly concentrated, but their distributions are generally skewed with substantial outliers, as also indicated by the larger mean values. The ASU-DNN model provides a distribution for train VOTs where values are concentrated around 20 CHF/Hour and 125 CHF/Hour. The distribution for SM is more concentrated around its median, with significantly more negative values. The distribution for cars, however, is much narrower around two main values. This might indicate that the ASU-DNN captures lower heterogeneity in the car VOT. The C-DNN model provides narrower VOT distributions compared to DNN for all modes. The VOT distributions generated by the C-ASU-DNN are the most concentrated.

In summary, the analysis of VOTs reveals that the data-driven modeling flexibility that characterizes DNN models makes them also prone to behaviorally unrealistic predictions. These can be avoided by the introduction of relevant domain knowledge constraints. The

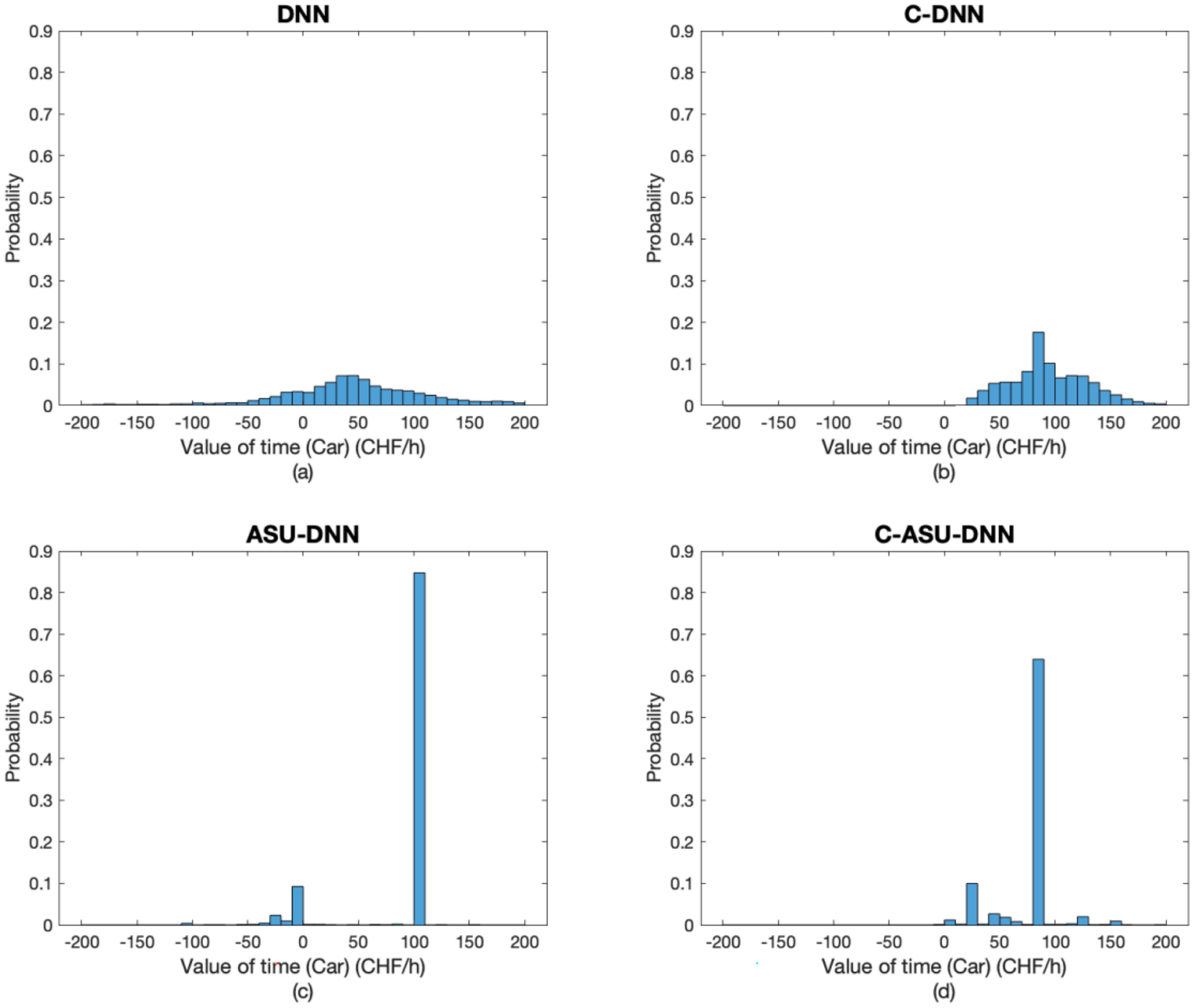


Fig. 12. Distributions of car values of time. Extreme values were cut off.

MNL model can be explicitly developed to eliminate undesired responses, but at a cost of rigid utility functional forms and potential misspecifications.

4.2.3. *Sensitivity analysis of penalty weight and perturbation value*

This section investigates the impact of the penalty weight λ and perturbation value ϵ on the performance and consistency of the C-DNN and C-ASU-DNN models. These parameters directly affect the balance between model fit and adherence to domain knowledge constraints. The analysis focuses on how variations in these parameters influence the ANLL, accuracy, RMSE of market shares, and percentage of inconsistent observations with domain knowledge across training, validation, and testing datasets.

4.2.3.1. Penalty weight. The penalty weight λ controls the strength of domain knowledge constraints in the models. In this analysis, λ values range from 10^{-2} to 10^{5}, applied identically across all constraints. The results are presented in Fig. 13 for all dataset splits: training, validation and testing, to explore if different behaviors can be observed across these sets.

As expected, there is a tradeoff between model fit and domain knowledge consistency in the C-DNN model. As the penalty coefficient λ increases, inconsistency percentages decrease across all sets, while the ANLL increases as shown in Fig. 13(g) and Fig. 13(a). Accuracy decreases for the training set, but shows slight improvement on the testing set for smaller penalty weights ranging from 10^{-1} to 10^{1} (Fig. 13(c)), suggesting that domain knowledge with small penalty weights may contribute to accuracy. Specifically, with a weight greater than 10^{4}, the improvement in domain knowledge consistency is negligible, while significant deteriorations occur in all other metrics. This indicates that with a higher weight, each inconsistent observation becomes highly important to the model to improve due to its high weight, leading the model to minimize them at the cost of overall model fit, as measured by the various metrics.

A tradeoff is also evident for the C-ASU-DNN model, but with a slightly different behavior. With penalty weight ranging from 10^{-2}

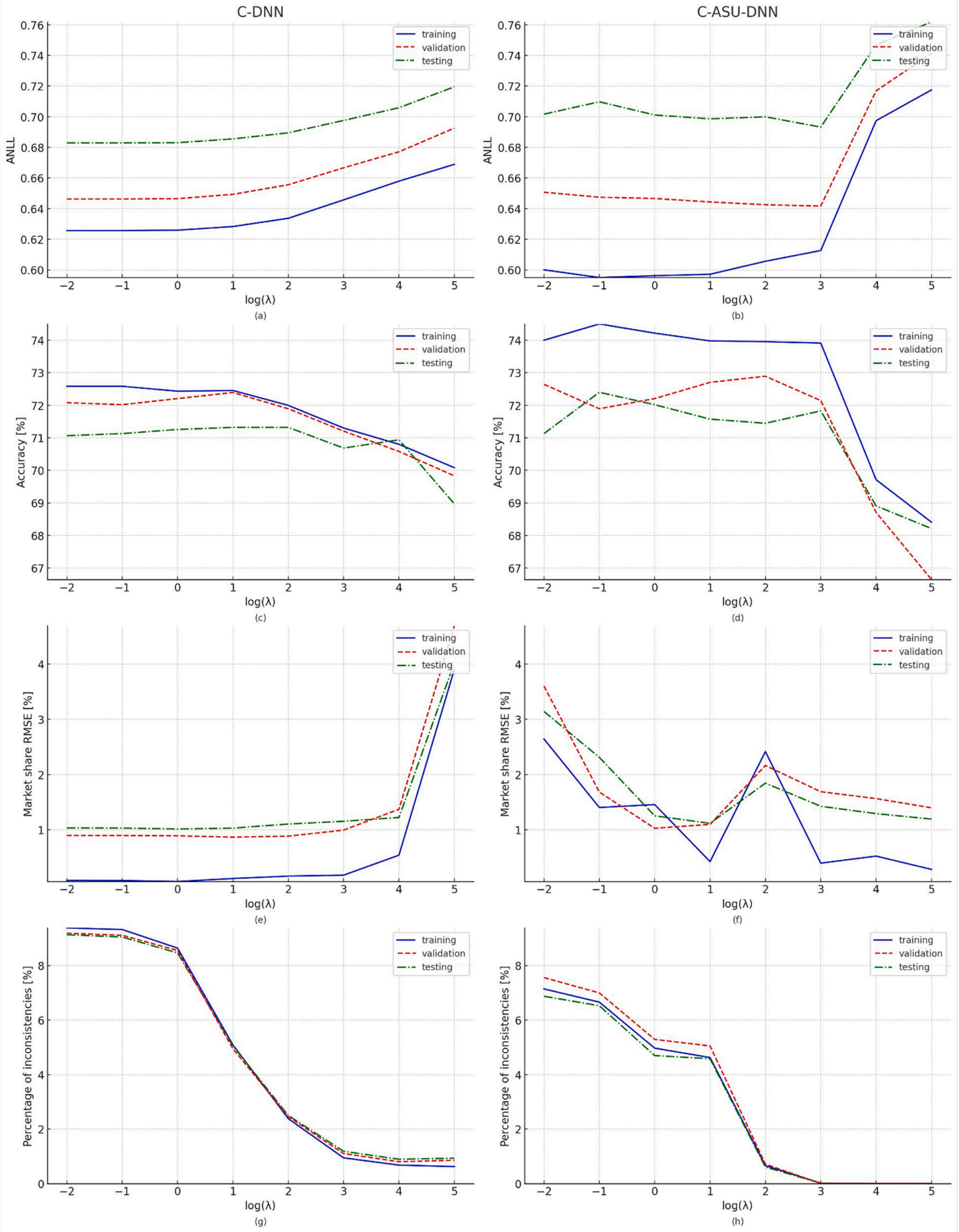


Fig. 13. Sensitivity analysis of penalty weight λ.

to 10^3, the ANLL on training set slightly deteriorates, while improves on the testing set (Fig. 13(b)), indicating a possible mitigation of overfitting. Accuracy shows slight fluctuations, but the changes are minimal (Fig. 13(d)). The RMSE of estimated market shares also improves with the penalty weight with an abrupt deterioration of 2 % at $\lambda = 10^2$, after which it continues to improve. These improvements in ANLL and RMSE of estimated market shares suggest a contribution of domain knowledge to the generalization performance of the ASU-DNN model. At penalty weights smaller than 10^3, the RMSE of estimated market shares remains consistent across all datasets. However, a larger gap appears for larger penalty weights, indicating a deterioration in the model's generalization performance, as evidenced by sharp drops in accuracy and ANLL. This behavior is related to the large penalty weight. At $\lambda = 10^3$, the percentage of inconsistent observations approaches zero, causing the model to prevent each single inconsistency at the cost of

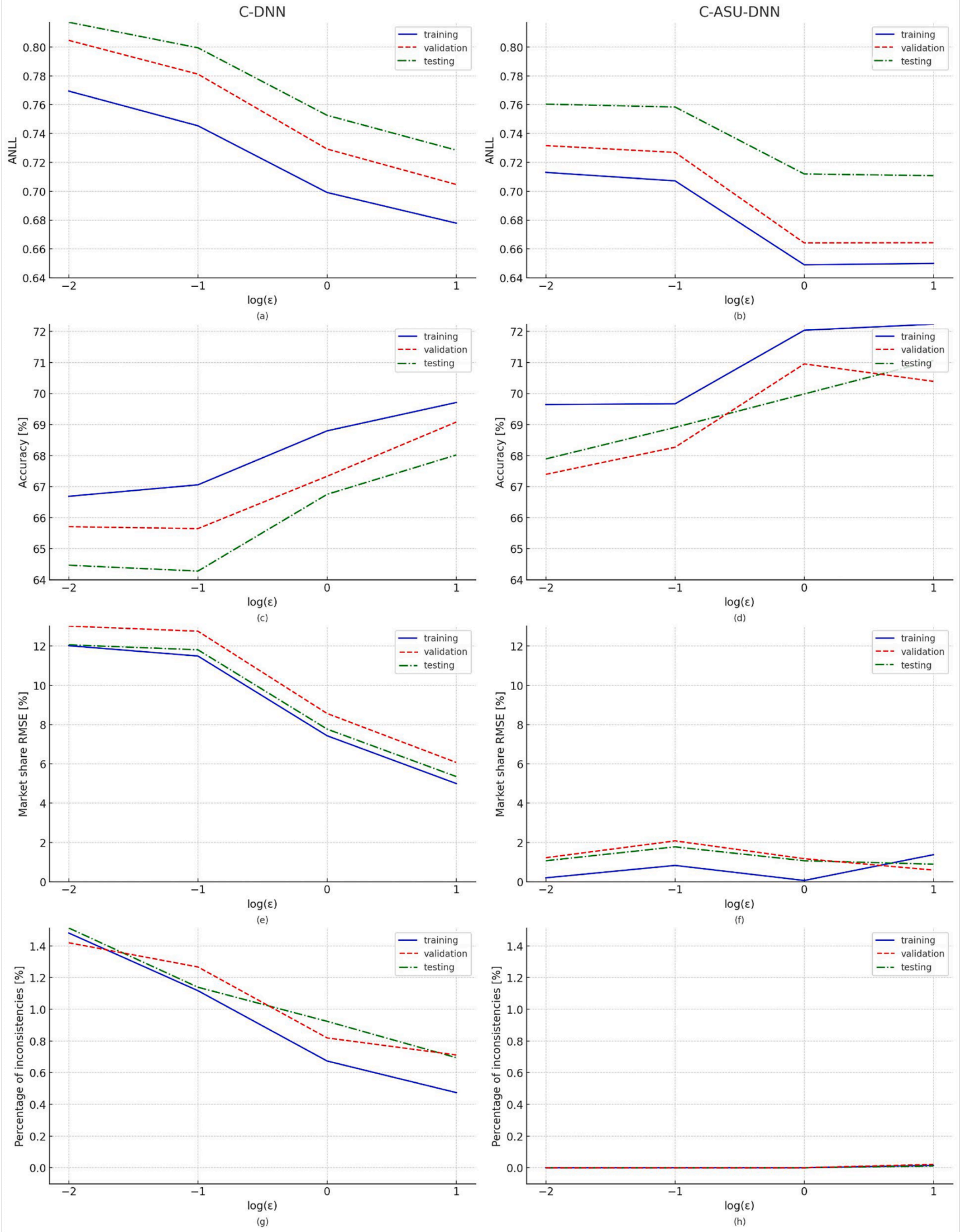


Fig. 14. Sensitivity analysis of perturbation value ε.

deteriorating other metrics.

4.2.3.2. Perturbation value $\boldsymbol{\varepsilon}$. The perturbation value (ε) determines the granularity of the local approximations used to evaluate the choice probabilities' sensitivities to time and cost variables. The investigated ε values range from 10^{-2} to 10, in variable-specific units (i.e., minutes for travel time and CHF for travel cost). The perturbation value is set identical for all time and cost variables. The analysis conducted to evaluate the impact of the perturbation value on the model's performance reveals several key insights, presented in Fig. 14.

The ANLL of both models improves with increasing perturbation values across all datasets (Fig. 14(a) and Fig. 14(b)). The accuracy also presents a general improvement trend with a slightly different behavior (Fig. 14(c) and Fig. 14(d)). For C-DNN, while accuracy increases on training set, it slightly drops on validation and testing sets at $\varepsilon = 10^{-1}$ (Fig. 14(c)). For C-ASU-DNN, however, a slight drop occurs at $\varepsilon = 10$ on testing set. The RMSE of the estimated market shares for the DNN model decreases with the perturbation value (Fig. 14(e)), while the pattern for the ASU-DNN is unclear, but remains within a 2 % bound (Fig. 14(f)).

The overall improvements in the above metrics for both models can be explained by the fact that with an increasing perturbation value, the local approximation of the derivatives becomes less accurate (due to larger finite differences) and ease in capturing the general trend of the choice probabilities sensitivities. This is also demonstrated in Fig. 14(g) and Fig. 14(h), where the percentage of inconsistent observations are decreasing. Furthermore, with greater perturbation values, less pseudo-data points are generated since their number is dictated by the perturbation value. With less pseudo-data points and same penalty weight (λ), the weight for domain knowledge constraints in the objective function becomes smaller, making the model focus more on data fit metrics.

4.3. Summary and discussion

The results from both synthetic and empirical studies highlight a critical balance between predictive performance, interpretability, and generalization in DNNs when incorporating domain knowledge constraints. Across both studies, the unconstrained DNN models achieve high accuracy and low ANLL, excelling in fit metrics. However, this improved fit comes at a cost: without constraints, the models often generate behaviorally inconsistent and counterintuitive predictions, such as extreme and negative WTP estimates. These inconsistencies highlight that high predictive accuracy alone does not guarantee realistic behavioral insights, as models may still produce outputs that deviate from economic expectations without domain-informed guidance.

Introducing constraints through penalty weight λ allows the C-DNN models to address this issue by aligning predictions more closely with behavioral theory. As λ increases to moderate values, the constrained models achieve a balance between fit and interpretability, reducing variability in WTP estimates and narrowing confidence intervals, indicating greater stability. Notably, these models maintain high predictive accuracy and low ANLL across datasets, demonstrating that moderate regularization can preserve predictive power while enhancing behavioral consistency. Conversely, when λ becomes too high, the model's flexibility diminishes, resulting in degraded fit (higher ANLL and RMSE) and generalization capability. Similarly, larger perturbation values reduce inconsistencies but may sacrifice local accuracy by emphasizing data fit over fine-grained sensitivity to time and cost variables. Thus, optimal tuning of these parameters is crucial for achieving a balance among flexibility, interpretability, and predictive power.

The interpretability of constrained models proves especially valuable in the market share sensitivity analysis, which simulates real-world policy scenarios by systematically varying travel times and costs across a range from a 50 % reduction to a 150 % increase in observed values. This method tests the models' ability to generalize under conditions beyond the original dataset and mimics scenarios relevant for policy forecasting, such as the impact of tax adjustments, travel time reductions, or service improvements. The empirical study reveals that unconstrained models sometimes produce counterintuitive results in this analysis, such as increased market shares for alternatives as their costs increase – behaviors inconsistent with economic principles. Constrained models, however, adhere to domain knowledge, ensuring plausible market share responses across varied scenarios and enhancing their reliability for forecasting policy outcomes. This adherence reinforces the importance of constraints in preventing behaviorally implausible outcomes, which purely data-driven DNN models tend to generate.

The VOT analysis further emphasizes the value of constraints. Without them, unconstrained models yield negative VOT estimates that contradict economic theory, whereas the constrained models effectively prevent these inconsistencies. While the ASU-DNN structure helped mitigate some inconsistencies, only the constrained models eliminated negative VOTs almost entirely, offering behaviorally realistic estimates. This analysis emphasizes that data-driven models, without proper constraints, may yield predictions that lack interpretability, even when they achieve good predictive performance.

These results show that introducing domain knowledge constraints enhances model interpretability and generalization without severely compromising predictive performance. The constrained models provide a viable middle ground between the flexibility of purely data-driven approaches and expected domain knowledge.

5. Conclusion and discussion

This study addresses the task of integrating domain knowledge with DNNs, in the context of travel demand predictions, bridging the gap between data-driven flexibility and behavioral consistency. While DCMs have a sound base in RUM theory, they are prone to over-simplified and subjective specifications of the utility functions. DNN models are stronger in extracting more complex relationships but are often criticized for their "black box" nature, which may lead to counter-intuitive predictions that lack consistency with domain knowledge. Analysts typically possess prior knowledge that they expect the model to capture. However, as the case studies demonstrate, this knowledge is not always captured by models that are purely data-driven.

To overcome this limitation, a framework is proposed to enhance the consistency of DNN models with domain knowledge while maintaining prediction capabilities. It involves incorporating constraints on the model to ensure that specific relationships are held, while leaving others unrestricted for data-driven learning. Only the sign of sensitivities of choice probabilities with respect to attributes of interest (i.e., positive or negative) is constrained, preserving flexibility in the model's form. The proposed framework is independent of the model structure, making it easy to implement on different architectures (e.g., DNN and ASU-DNN).

Both the synthetic and the empirical Swissmetro case studies illustrate the value of this approach, demonstrating that domain knowledge-constrained models outperform their unconstrained counterparts in terms of interpretability while maintaining robust predictive power. Although the constrained models show a slight decline in goodness of fit to the training data, they exhibit better

Table A1
Attributes' sampling distributions for systematic utility in the synthetic case study.

Individual attributes	Sampling distribution	Description
Inc	LogNormal(log(0.5),0.25) for full-time LogNormal(log(0.25),0.2) for not full-time	Household income ($ per min)
Full	Bernoulli(0.5)	Full time worker (1 = yes, 0 = no)
Flex	Bernoulli(0.5)	Flexible working hours (1 = yes, 0 = no)
Alternative attributes	**Sampling distribution**	**Description**
Cost	Uniform(0.2,100) for train and metro	Travel cost ($)
Time	Uniform(5,100) for train and metro	Travel time (min)
Wait	Uniform(5,30) for train and metro	Waiting time (min)
Crowd	Uniform(−5,5) for metro; 0 for train	Degree of crowding

Adapted from Kim and Bansal (2024).

generalization to unseen data, minimizing overfitting. Notably, these models reduce prediction errors in market shares and ensure that behaviorally implausible outcomes, such as negative VOTs, are avoided – a crucial improvement over purely data-driven approaches.

The market share sensitivity analysis further validates the approach, showing how constrained models align closely with domain knowledge across a range of travel cost and time scenarios. These predictions, performed over an extrapolated range of 50 % to 150 %, mimic real-world applications where DCMs are used to predict future scenarios, such as tax changes or travel time reductions. The constrained models consistently produced rational substitution patterns that were not observed in unconstrained versions, which ensures the practical applicability of these models for policy-relevant decisions.

The analysis of VOT further underscores the importance of domain knowledge. Unconstrained models generate significant percentages of negative VOTs, reflecting economically implausible behavior. In contrast, constrained models effectively eliminate negative VOT estimates, highlighting their ability to capture realistic behavioral patterns. The results demonstrate that while DNNs offer modeling flexibility, they also require domain knowledge to avoid producing misleading outcomes.

While, encoding numerous constraints into the neural network architecture helps guide the model towards more interpretable and theoretically consistent outcomes, excessive constraints may risk leading to tunnel vision, where the model only confirms pre-existing expectations and hinders the discovery of unexpected patterns or insights. This underscores the importance of a judicious balance in applying constraints, ensuring they enhance rather than restrict the model's learning capabilities.

The synthetic case study investigates the effect of the penalty parameter (λ) on model performance and interpretability. The results demonstrate that low λ values, including the unconstrained case, lead to less interpretable outcomes, such as inflated or inconsistent WTP estimates and a higher proportion of economically implausible negative values. As λ increases, the model becomes more interpretable, with predictions aligning better with behavioral patterns, reduced variability, and fewer negative WTP estimates. However, excessively high λ values reduce the model's flexibility, limiting its ability to capture complex patterns and leading to underfitting.

The effect of penalty weight (λ) and perturbation (ε) parameters was investigated in Swissmetro study to understand their impact on model performance and interpretability. Larger λ values restrict model flexibility, leading to higher ANLL and lower accuracy, though market share predictions remain stable until λ becomes very large. In contrast, larger ε values improve model fit and accuracy by smoothing the finite difference approximation and reducing the number of pseudo-data points, allowing the model to focus on prediction performance. These results highlight the importance of carefully tuning λ and ε to balance behavioral constraint adherence and predictive accuracy.

It is noteworthy that the domain knowledge is used as soft constraints. This allows a degree of flexibility in the model learning process. Soft constraints acknowledge the possibility that the available data may not be a perfect reflection of the expected knowledge due to potential issues such as measurement errors and unobserved attributes. Therefore, when justified by the data, the model can override the constraints and avoid the pitfall of enforcing potentially incorrect or overly rigid domain knowledge assumptions. Hard constraints (e.g., Kim and Bansal, 2024), however, may compel the model to neutralize the attributes that cause violations (e.g., neutralizing the effect of time if it fails to enforce it to be negative). Márquez-Neila et al. (2017) compared these two types of constraints and found that soft constraints perform better because satisfying hard constraints leads to sub-optimal outputs.

Therefore, the choice of λ depends on the acceptable degree of inconsistencies, such as negative VOT, where travelers may derive utility or enjoyment from travel time. For example, leisure activities during a trip (such as reading, using mobile devices, or enjoying scenic views) can make travel time feel less burdensome and even offer positive value (Salomon and Mokhtarian (2001) and Mokhtarian et al. (2001)). To capture such scenarios, λ values could be adjusted to moderately relax these constrains. This consideration will become increasingly relevant with innovations like autonomous vehicles and improved public transportation, which create more opportunities for productive or enjoyable activities while traveling, influencing both transport planning and policy development. This raises the question of whether it is acceptable to allow a limited number of negative VOTs, and adjust λ value accordingly, given these contextual shifts. Nevertheless, VOWT is perceived differently (often as more burdensome than travel time) and negative VOWT may be entirely avoided, as waiting is universally considered undesirable. This requires a larger λ value. This opens a discussion on allowing constraint-specific λ values, to assign a larger penalty weight for waiting time than that for travel time to reflect this behavioral distinction.

Future work could explore adaptive penalty tuning through Lagrangian loss-balancing algorithms (Son et al., 2023), where λ is adjusted dynamically based on model performance. Furthermore, the value of λ was set identical for all sensitivity constraints;

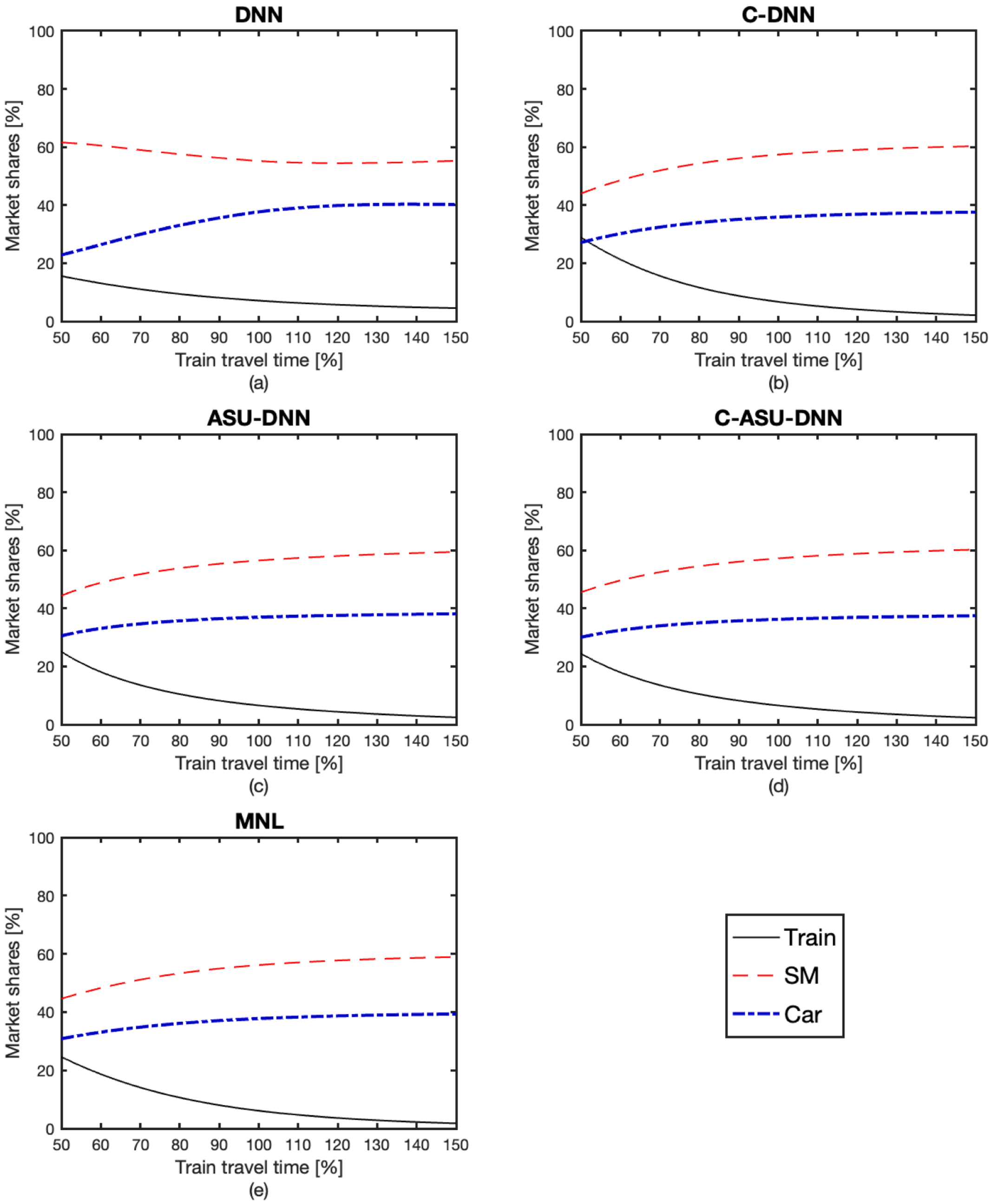


Fig. A1. Travel mode market shares as a function of train travel time.

however, this is not mandatory. Different penalty weights can be assigned to different constraints, if the model performs worse on specific constraints compared to others, or if constraints have different priorities. Similarly, the perturbation value was set identical for both time and cost variables; however, different magnitudes can be set to different variables (e.g., 10 min and 1 CHF). Additionally, future work could investigate the effects of different pseudo-data generation methods on the model's performance and interpretability.

Future work could also include additional domain knowledge, such as magnitudes of elasticities and values of time. The considered domain knowledge referred to the sensitivities' signs of choice probability, but this can be extended to other types of knowledge. Domain knowledge may refer to the magnitudes of elasticities, such as inelastic demand of public transportation modes to their cost (i.e., constraining their elasticity to be less than one) (Ben-Akiva and Lerman, 1985). In addition to monotonicity, common shape restrictions may include concavity or convexity of the utility, to demonstrate diminishing or increasing marginal utility. For example, utility function can be further constrained to be concave to model the concept of satiation (Aboutaleb, 2022). Domain knowledge might also hold information about the acceptable range of the estimated values of time, in addition to being positive. Alternatively,

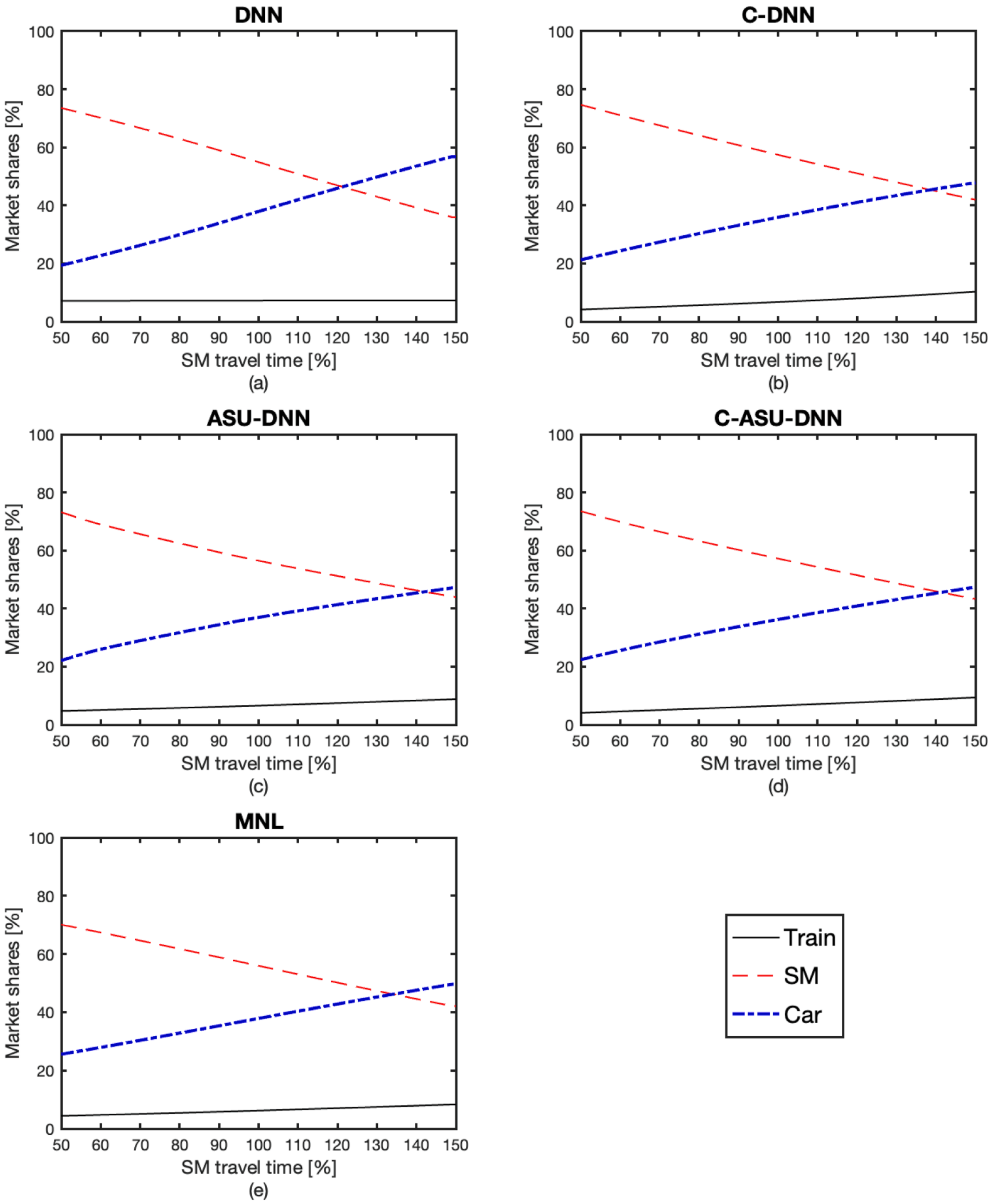


Fig. A2. Travel mode market shares as a function of SM travel time.

when the distribution of values of time is known (or assumed), an additional Kullback-Leibler (KL) divergence term can be incorporated to the loss function, which measures the difference between two probability distributions (Ahuja, 2019).

A key advantage of this framework is the high adaptability. Its independence from the core model structure makes it easy to implement on different DNN architectures. Thus, it offers potential for application beyond travel demand modeling. It is suitable for datasets with mixed data types, such as images, GPS traces, and structured data, due to its flexibility in integrating domain constraints directly into the loss function. Future work could explore how domain knowledge enhances performance in such multi-modal contexts, testing the robustness of the proposed method. In domains where interpretability and predictive performance are equally valued, such as healthcare or transportation policy, the ability to incorporate knowledge through soft constraints can offer significant advantages.

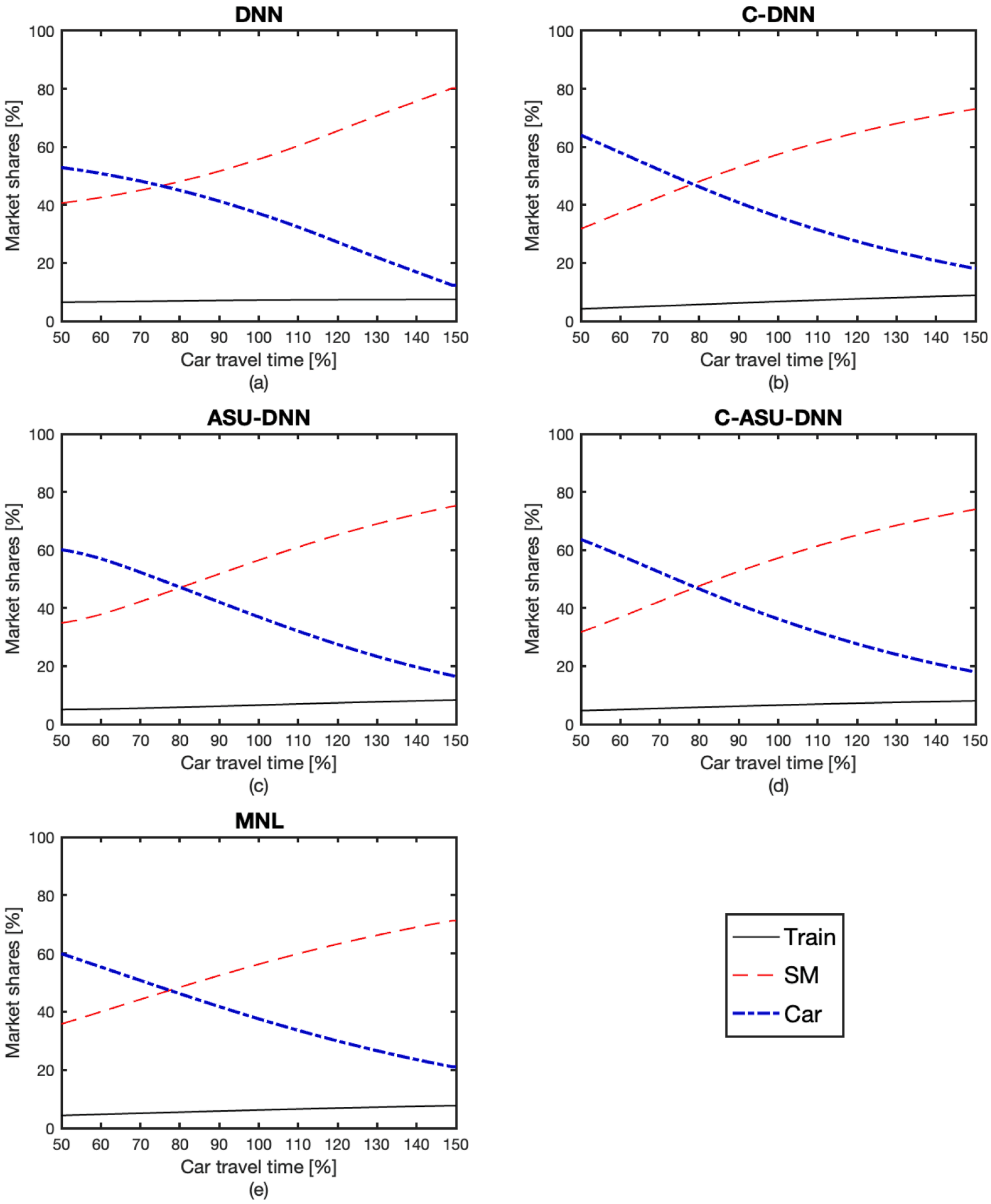


Fig. A3. Travel mode market shares as a function of car travel time.

Funding source declaration

This research was partially funded by grants from the Israeli Science Foundation (ISF), MIT-Israel Zuckerman STEM fund. Shadi Haj-Yahia is a VATAT doctoral fellow. The funding sources had no involvement in the study process.

CRediT authorship contribution statement

Shadi Haj-Yahia: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Conceptualization. **Omar Mansour:** Writing – original draft, Methodology, Investigation, Conceptualization. **Tomer Toledo:** Writing – review & editing, Supervision, Resources, Methodology, Conceptualization.

Table A2
MNL parameter estimates with statistical significance in parenthesis.

Coefficient	Train	Swissmetro	Car
ASC	1.35(<0.001)	−0.547(0.007)	
Travel time	−0.0238(<0.001)	−0.0169(<0.001)	−0.0166(<0.001)
Cost	−0.0121(<0.001)	−0.012(<0.001)	−0.00856(<0.001)
Headway	−0.00992(<0.001)		
Airline seats		0.552(0.001)	
Trip purpose			
Commute	1.05(<0.001)	1.26(<0.001)	
Shopping	1.88(<0.001)		
Business	0.647(<0.001)		
Leisure			
First class		0.169(0.043)	
Luggage			
None			
One piece			
Several pieces			
Age			
≤24		0.805(0.004)	
25–39	−0.607(<0.001)	0.681(<0.001)	
40–54	−0.713(<0.001)	0.498(0.001)	
55–65		0.413(0.008)	
≥66			
Male	−0.925(<0.001)	−0.421(<0.001)	
Income			
≤50		−0.163(0.015)	
51–100	−0.342(0.022)	−0.249(0.002)	
≥101			

Declaration of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

The similar trip purposes have been consolidated into single categories by combining the outbound and return trips (i.e., Commute and Return Commute, Shopping and Return from Shopping, Business and Return from Business, Leisure and Return from Leisure).

References

Aboutaleb, Youssef M. “Theory-constrained Data-driven Model Selection, Specification, and Estimation: Applications in Discrete Choice Models”. Diss. Massachusetts Institute of Technology, 2022.

Aboutaleb, Youssef M., Mazen Danaf, Yifei Xie, and Moshe Ben-Akiva. “Discrete choice analysis with machine learning capabilities.” arXiv preprint arXiv:2101.10261 (2021).

Ahuja, Kartik. “Estimating kullback-leibler divergence using kernel machines.” 2019 53rd Asilomar Conference on Signals, Systems, and Computers. IEEE, 2019.

Alwosheel, A., Van Cranenburgh, S., Chorus, C.G., 2019. ‘Computer says no’is not enough: Using prototypical examples to diagnose artificial neural networks for discrete choice analysis. J. Choice Modell. 33, 100186.

Alwosheel, A., van Cranenburgh, S., Chorus, C.G., 2021. Why did you predict that? Towards explainable artificial neural networks for travel demand analysis. Transp. Res. Part C Emerging Technol. 128, 103143.

Ben-Akiva, M.E., 1973. Structure of passenger travel demand models. Massachusetts Institute of Technology. PhD diss.

Ben-Akiva, M.E., Lerman, S.R., 1985. Discrete Choice Analysis: Theory and Application to Travel Demand, Vol. 9. MIT press.

Bierlaire, Michel, Kay Axhausen, and Georg Abay. “The acceptance of modal innovation: The case of Swissmetro.” In Swiss transport research conference, no. CONF. 2001.

Bishop, C.M., Nasrabadi, N.M., 2006. Pattern Recognition and Machine Learning, Vol. 4, no. 4. springer, New York.

Chang, X., Jianjun, Wu., Liu, H., Yan, X., Sun, H., Yunchao, Qu., 2019. Travel mode choice: a data fusion model using machine learning methods and evidence from travel diary survey data. Transportmetrica a: Transport Science 15 (2), 1587–1612.

Chapleau, R., Gaudette, P., Spurr, T., 2019. Application of machine learning to two large-sample household travel surveys: A characterization of travel modes. Transp. Res. Rec. 2673 (4), 173–183.

Van Cranenburgh, Sander, Shenhao Wang, Akshay Vij, Francisco Pereira, and Joan Walker. "Choice modelling in the age of machine learning-discussion paper." Journal of Choice Modelling 42 (2022): 100340.

Van Cranenburgh, Sander, and Ahmad Alwosheel. "An artificial neural network based approach to investigate travellers' decision rules." Transportation Research Part C: Emerging Technologies 98 (2019): 152-166.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232.

Hagenauer, J., Helbich, M., 2017. A comparative study of machine learning classifiers for modeling travel mode choice. Expert Syst. Appl. 78, 273–282.

Han, Y., Pereira, F.C., Ben-Akiva, M., Zegras, C., 2022. A neural-embedded discrete choice model: Learning taste representation with strengthened interpretability. Transp. Res. B Methodol. 163, 166–186.

Hess, S., Erath, A., Axhausen, K.W., 2008. Estimated value of savings in travel time in Switzerland: Analysis of pooled data. Transp. Res. Rec. 2082 (1), 43–55.

Hernandez, Jose Ignacio, Niek Mouter, and Sander van Cranenburgh. "An economically-consistent discrete choice model with flexible utility specification based on artificial neural networks." arXiv preprint arXiv:2404.13198 (2024).

Hillel, T., Bierlaire, M., Elshafie, M.Z., Jin, Y., 2021. A systematic review of machine learning classification methodologies for modelling passenger mode choice. Journal of Choice Modelling 38, 100221.

Kim, E.-J., Bansal, P., 2024. A new flexible and partially monotonic discrete choice model. Transp. Res. B Methodol. 183, 102947.

König, A., Abay, G., Axhausen, K., 2003. Time is money: the valuation of travel time savings in switzerland. In Proceedings of the 3rd Swiss Transportation Research Conference.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.

Mahajan, V., Katrakazas, C., Antoniou, C., 2020. Prediction of lane-changing maneuvers with automatic labeling and deep learning. Transp. Res. Rec. 2674 (7), 336–347.

Márquez-Neila, Pablo, Mathieu Salzmann, and Pascal Fua. "Imposing hard constraints on deep networks: Promises and limitations." arXiv preprint arXiv:1706.02025 (2017).

McFadden, D., 1973. Conditional logit analysis of qualitative choice behavior. Frontiers in Econometrics.

McFadden, D., 1978. Modeling the choice of residential location. Transportation Research Record [preprint] 673.

Mokhtarian, Patricia L., Ilan Salomon, and Lothlorien S. Redmond. "Understanding the demand for travel: It's not purely 'derived'." Innovation: the European journal of social science research 14.4 (2001): 355-380.

Omrani, H., 2015. Predicting travel mode of individuals by machine learning. Transp. Res. Procedia 10, 840–849.

Revelt, D., Train, K., 1998. Mixed logit with repeated choices: households' choices of appliance efficiency level. Rev. Econ. Stat. 80 (4), 647–657.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., Zhong, C., 2022. Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistic Surveys 16, 1–85.

Runje, Davor, and Sharath M. Shankaranarayana. "Constrained monotonic neural networks." International Conference on Machine Learning. PMLR, 2023.

Salomon, I., Mokhtarian, P.L., 2001. How Derived is the Demand for Travel. Transportation Research A 35, 695–719.

Sifringer, B., Lurkin, V., Alahi, A., 2020. Enhancing discrete choice models with representation learning. Transp. Res. B Methodol. 140, 236–261.

Son, H., Cho, S.W., Hwang, H.J., 2023. Enhanced physics-informed neural networks with augmented Lagrangian relaxation method (AL-PINNs). Neurocomputing 548, 126424.

Torres, C., Hanley, N., Riera, A., 2011. How wrong can you be? Implications of incorrect utility function specification for welfare measurement in choice experiments. J. Environ. Econ. Manag. 62 (1), 111–121.

Train, K.E., 2009. Discrete choice methods with simulation. Cambridge University Press.

Wang, Shenhao, Baichuan Mo, Stephane Hess, and Jinhua Zhao. "Comparing hundreds of machine learning classifiers and discrete choice models in predicting travel behavior: an empirical benchmark." arXiv preprint arXiv:2102.01130 (2021).

Wang, S., Mo, B., Zhao, J., 2020a. Deep neural networks for choice analysis: Architecture design with alternative-specific utility functions. Transp. Res. Part C Emerging Technol. 112, 234–251.

Wang, S., Wang, Q., Zhao, J., 2020b. Deep neural networks for choice analysis: Extracting complete economic information for interpretation. Transp. Res. Part C Emerging Technol. 118, 102701.

Wright, L.G., Onodera, T., Stein, M.M., Wang, T., Schachter, D.T., Zoey, Hu., McMahon, P.L., 2022. Deep physical neural networks trained with backpropagation. Nature 601 (7894), 549–555.

Zhao, X., 2020. Xiang Yan, Alan Yu, and Pascal Van Hentenryck. "Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models.". Travel Behav. Soc. 20, 22–35.