Combine and conquer: model averaging for out-of-distribution forecasting

Stephane Hess*[1,2] and Sander van Cranenburgh[2]

[1]University of Leeds, * s.hess@leeds.ac.uk
[2]Delft University of Technology

Short summary

Travel behaviour modellers are increasingly interested in using models from outside the traditional choice modelling area, first incorporating ideas from behavioural economics, such as in regret modelling, before looking at mathematical psychology and machine learning. A key question arises as to how well these different models perform in prediction, especially when predicting trips of different characteristics from those used in estimation. This paper first compares the elasticities and model fit of different models, bringing together models as diverse as logit, random regret, decision field theory and neural networks. We highlight differences in elasticities and also note that the prediction performance deteriorates at different rates for different models when moving further away from the estimation data. We then develop a model averaging approach that allows us to make the most of the entire collection of models and estimate weights for different models as a function of distance away from the estimation sample. **Keywords**: choice modelling; forecasting; machine learning; mathematical psychology; mode choice; model averaging

1 Introduction

The travel behaviour modelling literature has focussed extensively on two sorts of models, namely models for inference (henceforth inference models) and models for forecasting (henceforth forecasting models). Inference models aim to understand current travel behaviour (e.g. to recover the value of travel time), while forecasting models aim to forecast future travel behaviour in new settings (e.g. due to transport policies, such as toll roads and fuel levies). In other words, forecasting models are developed to generalise out-of-distribution.

Inference models and forecasting models are evaluated differently by analysts. When building an inference model, an analyst is keen that the model generates behaviourally plausible insights into causal factors and their relative impacts. Additionally, for inference models, it is well recognised that they should be able to replicate the behaviour in the empirical data as well as possible. In general, analysts perceive a high model fit as a proxy for a good model. The rationale is that the higher the likelihood of the empirical data given the model, the more reliable the results must be. Given this model evaluation approach, it should come as no surprise that researchers in the travel behaviour field are increasingly attracted by the comparatively good prediction performance of machine learning approaches (cf. Hagenauer & Helbich, 2017).

For forecasting models, the focus during evaluation is typically on elasticities. A widely held view is that forecasting models must produce behaviourally plausible elasticities, i.e. changes in demand in response to changes in journey characteristics. Furthermore, forecasting models are evaluated on their behavioural soundness. The dominant - although not necessarily evidenced-based - view is that forecasting models with a solid behavioural underpinning, such as Random Utility Maximisation (RUM) based discrete choice models, are better equipped to forecast behaviour under new settings than are models with a weak or no behaviour underpinning, such as e.g. machine learning models (cf. van Cranenburgh et al., 2022).

However, what is currently less well understood is how to value and incorporate model performance, i.e. in the model fit on the empirical data, when developing forecasting models. What is clear is that good model performance is not sufficient to establish that a model will make good out-of-distribution predictions. The fact that model parameters are estimated by maximising the model performance on the empirical data, i.e. how well they replicate current choices, is somewhat at odds with the aim of forecasting: to generalise out-of-distribution. The choice modelling literature has at times recognised that more advanced models that offer a better fit on the empirical data do not necessarily lead to better forecasts (see e.g. Fox et al., 2014). But the question is still

open on how to develop forecasting models considering the model's performance and behavioural underpinning. After all, it is intuitive that the performance of a model on the empirical data still pertains to relevant information on its ability to generalise out-of-distribution.
The aim of this paper is twofold. First, we aim to quantify the deterioration of prediction performance as a function of the "distance" between the training data and forecasting scenarios for models with varying levels of behavioural underpinning. Second, we aim to develop a model averaging-based approach that reduces the bias in forecasts by assigning different weights to different models depending on this "distance". Using this model, we aim to not only develop a flexible tool for combining models, but to craft rules-of-thumb for the conditions under which what sort of models perform best in terms of out-of-distribution forecasting.
The key hypotheses of the present paper are as follows:

1. Good absolute in-sample prediction performance does not necessarily translate into accuracy of elasticities.
2. Prediction performance for all models deteriorates as a function of the "distance" between the training data and scenarios for which a prediction is made.
3. Models with a solid behavioural underpinning will increasingly perform better than models without behavioural underpinning at increasing "distance".
4. Flexible models and models with a weak or no behaviour underpinning will suffer from a considerable spread in out-of-distribution generalisation performance.

2 Methodology

Datasets

We use two different revealed preference datasets in this study, both focussing on mode choice.
The first dataset comes from a large-scale survey conducted as part of the DECISIONS project carried out by the Choice Modelling Centre at the University of Leeds (Calastri et al., 2020). The data used for this work corresponds to the observed mode choice behaviour where after extensive data cleaning and data enrichment (Tsoleridis et al., 2022), 12,524 trips made by 540 individuals remained. For each trip, individuals travelled by one of six modes: car, bus, rail, taxi, cycling or walking. Attributes of the alternatives used in the models include in vehicle travel time, out of vehicle travel time, and travel cost.
The second dataset is the London mode choice data compiled by Hillel et al. (2018). This dataset contains four alternatives: walking, cycling, public transport (grouping together bus and rail) and driving. We use a sample of 81,086 trips. Attributes of the alternatives used in the models include in vehicle travel time, out of vehicle travel time, interchanges, and travel cost.

Model types

The following individual models were used in our analysis, combining models from traditional choice modelling, mathematical psychology and machine learning:

Logit models: Standard multinomial Logit models using three different specifications, namely a) linear in attributes; b) log-linear in attributes; and c) linear plus log-linear in attributes.

Nested logit models: Nested Logit using three different specifications, namely a) linear in attributes; b) log-linear in attributes; and c) linear plus log-linear in attributes. In terms of nesting structure, the DECISIONS models grouped together public transport options in a nest, while the London models grouped together motorised modes *vs* active modes.

Random regret minimisation (RRM): RRM models were used, treating attributes as linear, and with different constants depending on choice set size given varying mode availability in the DECISIONS data.

Decision field theory (DFT): DFT is a dynamic, stochastic model, introduced by Busemeyer & Townsend (1993). The key idea of the DFT model is that the preferences for different alternatives update over time whilst the decision-maker considers the different alternatives and their attributes. We use the implementation of T. O. Hancock et al. (2019).

MultiLayer Perceptron (MLP): This model comprises an input layer with input nodes, one or more hidden layers with hidden nodes, and and output layer with output nodes. In this model, signals propagate forward through the links connecting the nodes. The links have a numeric weight w, which is learned from the data. At each node, the weights are multiplied with the input value from the previous nodes and summed. Then the signal is propagated to the next layer using an activation function. We use tanh activation functions. In the output layer, a Softmax function (i.e. a logit) is applied to produce choice probabilities for each alternative.

XGBoost (XGB): The XGB model comprises a series of sequentially applied decision trees. A decision tree is a sequence of simple IF-THEN rules, optimised to classify data accurately. In the XGB model, each decision tree in the sequence 'corrects' the mispredictions of the models before it. This process is referred to as 'boosting'. The word 'gradient' relates to the notion that each subsequent decision tree is fitted on residuals of the trees before it.

Identifying the role of "distance"

To study the impact of distance from the estimation data, we first divided the samples into 10 subsets by distance, e.g. using the 10 percent of shortest trips for the first segment.
Distance segments 1 and 10 were excluded from the model fitting work to retain them for later out-of-distribution validation. Five separate models were then estimated for each model type on rolling subsets of the data combining 4 distance segments, e.g. segments 2-5 for the first model, 3-6 for the second, etc. Let us define $M_{m,g}$ to be the model of type m estimated on distance grouping g, where, e.g. with $g = 1$, we would use distance segments 2-5.
Finally, each of the estimated models was used to make a prediction of each trip in the data, independently of the distance segment for that trip, so e.g. also using models estimated on segments 2-5 only to predict mode choice for trips in distance segment 9.
Model estimation relied on 80% of the sample, with the remaining 20% kept for later out-of-sample and out-of-distribution prediction.

Model averaging approach

The final step of the work uses a model averaging approach, as outlined for example by S. H. A. D. Hancock Thomas O. & Fox (2020). Model averaging relies on a sequential latent class approach, where M different models have been estimated on the data, with model m giving a likelihood $P_{n,m}$ for the choice in observation n (working either at the person or observation level).
The model averaging log-likelihood is then given by

$$LL = \sum_{n=1}^{N} log \sum_{m=1}^{M} \pi_{n,m} P_{n,m}, \tag{1}$$

where $\pi_{n,m}$ is the estimated weight given to model m for observation/person n, where this is given by:

$$\pi_{n,m} = \frac{e^{\gamma_m' z_n}}{\sum_{l=1}^{M} e^{\gamma_l' z_n}}, \tag{2}$$

with an appropriate normalisation, where z_n are characteristics of person/observation n.
Model averaging typically estimates each model on the entire data and then computes the weights by considering how well each model fits for each of the data points.
We use a different approach. Specifically, we rely not only on the observations on which the models were estimated, given that each model uses only 4 distance segments, but include the prediction performance of every model on each of the trips. This allows us to study how the weight assigned by model averaging is a function of both the model type and how far, if at all, out-of-distribution, the model is. In other words, we expect that the weight given to a model decreases as a function of how far away the distance of the trip for each a mode choice is predicted is from the average distance of trips on which the model was estimatd.
With this in mind, let D_g be the average distance of trips in distance grouping g, and let d_n be the distance of the specific trip for which we make a prediction, where we work at the level of individual trips.
We then have that our log-likelihood is given by:

$$LL = \sum_{n=1}^{N} \sum_{g=1}^{G} log \sum_{m=1}^{M} \pi_{n,m,g} P_{n,m,g}, \tag{3}$$

where $G = 5$.
Each observation in the model averaging process thus uses the predictions for all the different model types for a given distance grouping. This approach produced better results than looking jointly at predictions from all GM models in a single row.
The model averaging weights are now specified as

$$\pi_{n,m,g} = \frac{e^{\delta_m + (\gamma_{d,m}(d_n < D_g) + \gamma_{i,m}(d_n > D_g)) log||D_g - d_n|}}{\sum_{l=1}^{M} e^{\delta_l + (\gamma_{d,l}(d_n < D_g) + \gamma_{i,l}(d_n > D_g)) log||D_g - d_n|}}, \tag{4}$$

where δ_m is a constant for model type m, and $\gamma_{d,m}$ and $\gamma_{i,m}$ are parameters capturing the influence on class allocation when the distance of a trip is below, respectively above the average distance of trips on which the model was estimated, where a non-linear transform was used.

3 Results and discussion

The work has produced a wealth of results of which we only focus on a subset in this brief paper.

Model fit comparisons

Figure 1 compares the model fit (using $\rho 2$) in prediction for MLP and linear logit, as a function of the estimation and prediction segment. A positive differences indicates better fit for MLP than logit. We note that the expected patterns emerge, with MLP overall predicting better in-distribution than logit, but losing out when moving out-of-distribution, though with some exceptions.

	Forecast_segment_1	Forecast_segment_2	Forecast_segment_3	Forecast_segment_4	Forecast_segment_5	Forecast_segment_6	Forecast_segment_7	Forecast_segment_8	Forecast_segment_9	Forecast_segment_10
Model_segments_2_5	-0.02105495	0.015811827	0.027874172	0.03394731	0.001921904	-0.01158405	-0.008914669	-0.09583537	0.32564889	0.07702223
Model_segments_3_6	-0.0160828	0.014424376	0.0236738	0.02070766	0.01798398	0.01204426	-0.001856788	-0.03780298	0.038396733	-0.00791557
Model_segments_4_7	-0.03849625	-0.004529412	0.007082065	0.03310145	0.027727421	0.04041223	0.030607856	-0.03929117	-0.133061398	-0.13602774
Model_segments_5_8	-0.17554066	-0.067432785	-0.035770125	-0.03242807	0.02202978	0.03880158	0.027496214	0.0434717	-0.034876721	-0.06171994
Model_segments_6_9	-1.04762794	-0.408353156	-0.180015921	-0.11351349	-0.100038998	-0.02941908	-0.02603404	-0.02049614	-0.001933407	-0.03624159

Figure 1: Model prediction comparison between MLP and linear logit: DECISIONS data

Elasticities

Figure 2 compares the car cost elasticities for a subset of models estimated on the DECISIONS data. We see differences as a function of which segment model was estimated on. For most models, we see decrease in elasticities for models estimated on longer distance trips, but not for MLP and especially XGB. This is an initial indication of differences in prediction results for different models.

Model averaging results

Figure 3 and 4 show the weights for different models generated by model averaging as a function of the difference in the distance of the trip under question and the average trip distance used in model estimation.Figure 3 shows the results for the DECISIONS data set; Figure 4 shows the results for the London data set.
The results show a diverse patterns. Firstly, for both data sets we see thatthe MLP and XGB models outperform the other models best close to the estimation distance,before tailing off with increasing distance. Second, the behavioural models take over for larger distances. Epecially notable is how well DFT performs for trips much longer than the estimation data on the DECISIONS data. But, in the London data this patterns is not visible. Third, for trips that are (much) shorter than the estimation data, the log-linear logit model performs well, on both data sets.

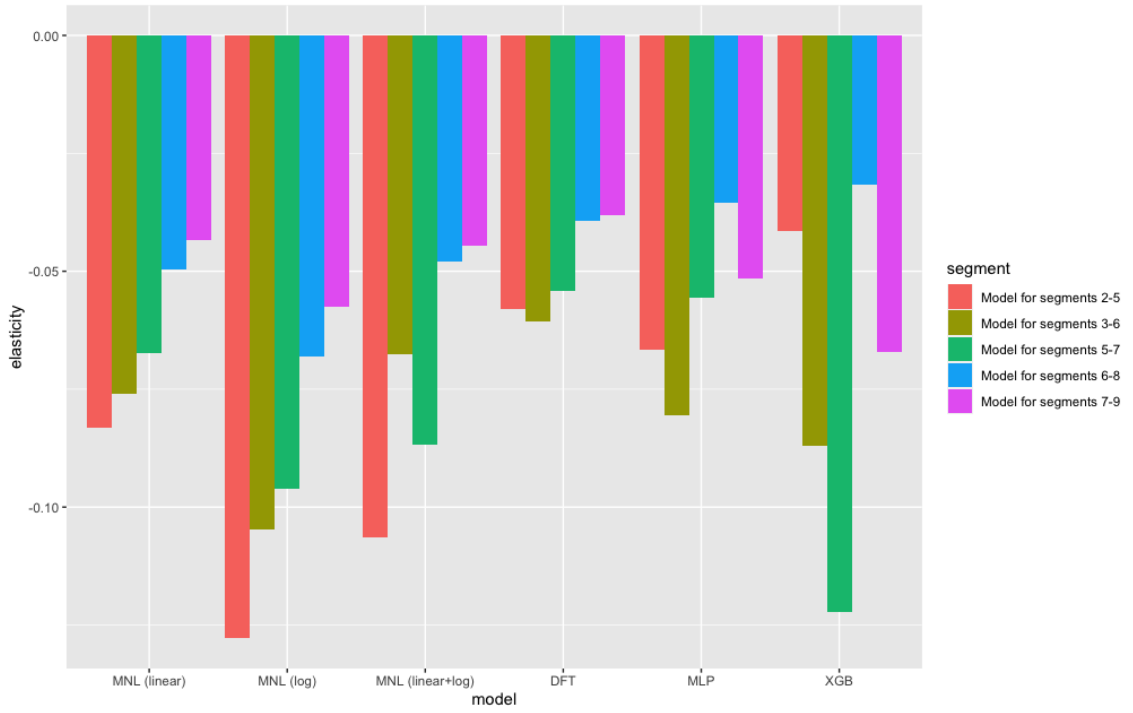

Figure 2: Car cost elasticity comparisons for select models: DECISIONS data

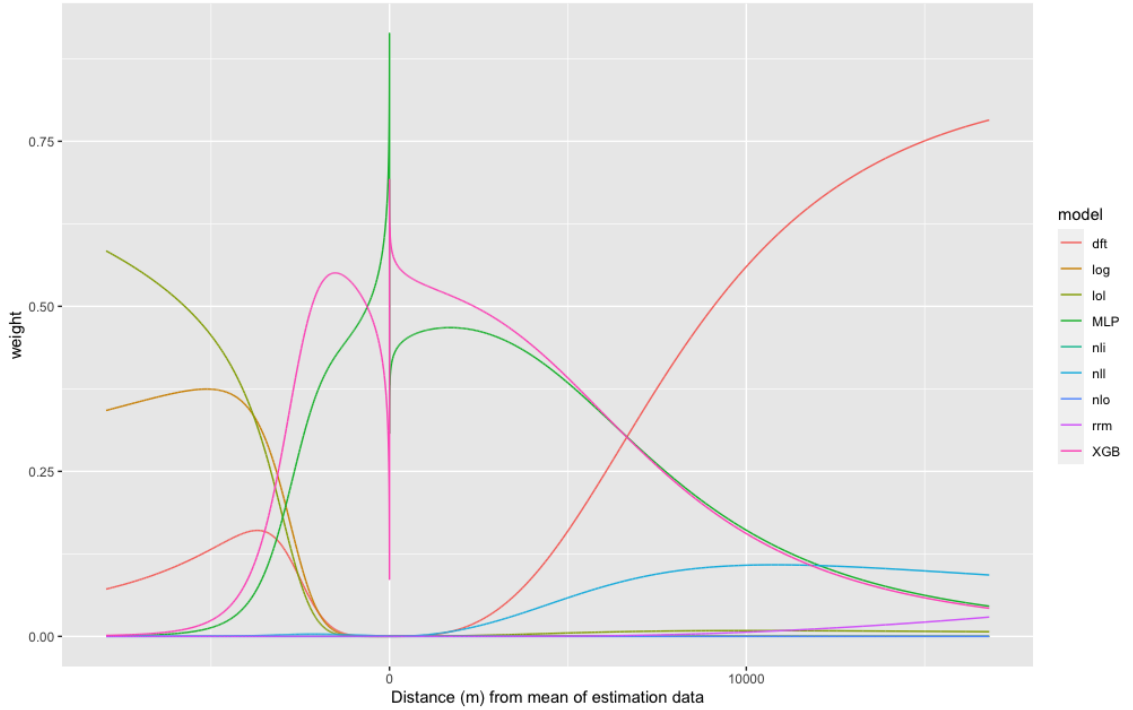

Figure 3: Model averaging weigths as a function of distance from estimation data: DECISIONS data (log=log-linear logit, lol=linear plus log-linear logit, nli=linear nested logit, nll=linear plus log-linear nested logit, nlo=log-linear nested logit, all other acronyms as in main text)

4 Conclusions

This work has taken an important step forward in combining insights from different modelling approaches for travel demand forecasting. Specifically, we have shown that different models predict choices differently well depending on how far away from the estimation sample the prediction takes

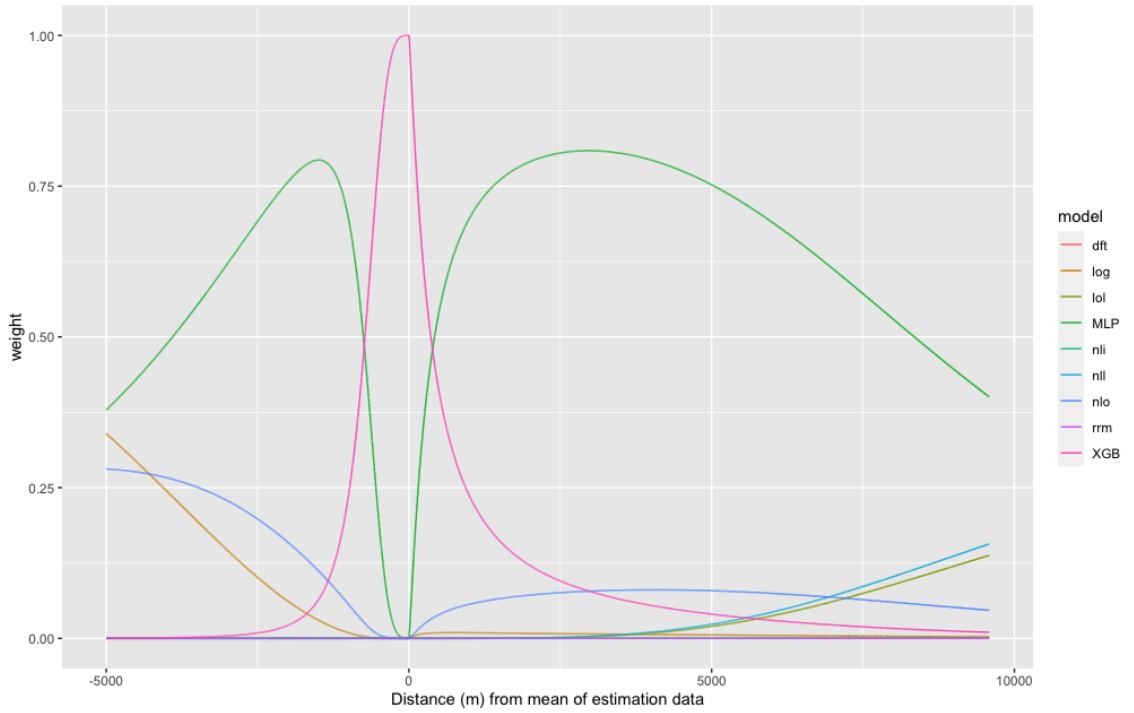

Figure 4: Model averaging weigths as a function of distance from estimation data: DECISIONS data (log=log-linear logit, lol=linear plus log-linear logit, nli=linear nested logit, nll=linear plus log-linear nested logit, nlo=log-linear nested logit, all other acronyms as in main text)

place. This result is not surprising in itself, but is quantified by our work.
Building forth on these insights, we contribute bydeveloping a model averaging-based approach that estimates weights for different models as a function of the distance away from the estimation data.
In the full paper, we present more detailed results, including on model fit, prediction out of sample, prediction completely out of distribution (i.e. distance segments 1 and 10), and full estimation results for the models. We also hypothesise about possible other attributes that could be used to measure 'distance' from the estimation data, going beyond trip distance alone.

Acknowledgements

Collate acknowledgements in this separate section at the end of the text, before the part of references. List here those individuals who provided help during the research (e.g., funding the project, providing language help, writing assistance or proof reading the article, etc.).

References

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, *100*(3), 432.

Calastri, C., dit Sourd, R. C., & Hess, S. (2020). We want it all: experiences from a survey seeking to capture social network structures, lifetime events and short-term travel and activity planning. *Transportation*, *47*(1), 175–201.

Fox, J., Daly, A., Hess, S., & Miller, E. (2014). Temporal transferability of models of mode-destination choice for the greater toronto and hamilton area. *Journal of Transport and Land Use*, *7*(2), 41–62.

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

Hancock, S. H. A. D., Thomas O., & Fox, J. (2020). Using a sequential latent class approach for model averaging: Benefits in forecasting and behavioural insights. *Transportation Research Part A: Policy and Practice*, *139*, 429-454.

Hancock, T. O., Hess, S., & Choudhury, C. F. (2019). An accumulation of preference: two alternative dynamic models for understanding transport choices. *Submitted*.

Hillel, T., Elshafie, M. Z. E. B., & Jin, Y. (2018). Recreating passenger mode choice-sets for transport simulation: A case study of london, uk. *Proceedings of the Institution of Civil Engineers - Smart Infrastructure and Construction*, *171*(1), 29-42.

Tsoleridis, P., Choudhury, C. F., & Hess, S. (2022). Deriving transport appraisal values from emerging revealed preference data. *Transportation Research Part A: Policy and Practice*, *165*, 225–245.

van Cranenburgh, S., Wang, S., Vij, A., Pereira, F., & Walker, J. (2022). Choice modelling in the age of machine learning - discussion paper. *Journal of Choice Modelling*, *42*, 100340.