IMPROVING FORECASTS AND BEHAVIOURAL INSIGHTS BY APPLYING MODEL AVERAGING ACROSS MULTIPLE CHOICE MODELS

Thomas O. Hancock (Corresponding Author)
Choice Modelling Centre & Institute for Transport Studies
University of Leeds
tstoh@leeds.ac.uk

Stephane Hess
Choice Modelling Centre & Institute for Transport Studies
University of Leeds
S.Hess@its.leeds.ac.uk

Andrew Daly
Choice Modelling Centre & Institute for Transport Studies
University of Leeds
andrew@alogit.com

James Fox
RAND Europe
jfox@rand.org

ABSTRACT

Despite the frequent use of model averaging in many disciplines from weather forecasting to health outcomes, it is not yet an idea often considered in travel behaviour or choice modelling. The idea behind model averaging is that a single model can be created by calculating contribution weights for a set of candidate models, depending on their relative performance, thus creating an 'average'. In this paper, we demonstrate that this idea can be used effectively for travel behaviour modelling. We identify two key opportunities for model averaging. The first is the situation where an analyst faces the difficult choice between a number of advanced models, all with some desirable properties. The second is the situation where advanced models cannot be used due to the size of the data and/or choice sets. Our tests demonstrate that in both cases, model averaging results in a consistent improvement in model fit for both estimation and in forecasting with subsets of validation samples. Additionally, we demonstrate that model averaging can be used to obtain more *reliable* elasticities and welfare measures by averaging across outputs obtained from the set of candidate models.

1. INTRODUCTION

Whilst there have thus far been very limited applications in choice modelling using model averaging, it is a popular method elsewhere with Bayesian model averaging being used regularly in medical statistics (Hoeting et al., 1999), ecology (Wintle et al., 2003) and biology (Posada and Buckley, 2004). Additionally, ensembles are often used to combine neural networks (Gazder and Ratrout, 2015; Moretti et al., 2015). Typically, model averaging or ensemble methods can be used to allow a modeller to establish a single model by calculating relative contribution weights for a set of candidate models. Within health, Bayesian model averaging has been successfully used to improve the prediction of who is at risk of a stroke (Volinsky et al., 1997), at risk of a coronary event (Wang et al., 2004) and to understand the relation between arsenic levels and cancer rates (Morales et al., 2006). Additionally, model averaging is often used for pooling forecasts from different models. This is particularly common for meteorological forecasting, with model averaging having been used to predict the surface temperature of the ocean (Raftery et al., 2005) and also wind speeds (Sloughter et al., 2010). It is also used in other fields for tasks such as predicting levels of economic inflation (Wright, 2009).

Choice modellers, by contrast, may often consider a set of candidate models, detail the advantages and disadvantages of each, but then subjectively choose only one model to use in the main application or reporting of key results. Consequently, it seems surprising that model averaging has not yet made the transition into mainstream choice modelling given that it can capture the benefits from a number of models and combine them into one model. We test this idea by averaging across a large number of candidate models as well as looking at the impacts on important model outputs such as cost and time elasticities. This is the first reason for applying model averaging, i.e. the case where multiple candidate models all have advantages and disadvantages and there is no clear cut case for choosing which is best. A second and rather different context arises in the case of very large-scale applications, either with large datasets or large choice sets, where we may not always be able to use as complex a model as we might otherwise choose to use due to the computational running time of complex models. This reason for model averaging will become even more timely

in the context of increasing reliance on big data.

The remainder of this paper is organised as follows. First, we present a methodology section demonstrating how we apply model averaging and how to get outputs such as elasticities from model averaging. Next, we present our empirical applications, where we use model averaging to improve model fit in both estimation and forecasting and to obtain welfare measures averaged across a set of candidate models. The final section summarises our findings and presents directions for future research.

2. METHODOLOGY

2.1. Model averaging in estimation

To apply model averaging, we first estimate a number of different individual models, where say $L(C_n \mid m, \Omega_m)$ gives the likelihood of the sequence of choices C_n observed for person n, conditional on using model m, where this model uses a vector of parameters Ω_m. We have that:

$$L(C_n \mid m, \Omega_m) = \int_{\beta_m} \prod_{t=1}^{T_n} P_m\left(j^*_{n,t} \mid \beta_m\right) f_m(\beta_m \mid \Omega_m)\, \mathrm{d}\beta_m. \tag{1}$$

In this general notation, we have that $P_m\left(j^*_{n,t} \mid \beta_m\right)$ gives the probability of the observed choice $j^*_{n,t}$ for decision maker n in choice situation t, conditional on using model m, where we allow for a general notation such that the parameters β_m are distributed according to $f_m(\beta_m \mid \Omega_m)$. Of course, it is possible that no random heterogeneity is used in which case the integral drops out, or that a latent class structure is used, replacing the integral with a weighted summation.

An analyst will estimate M different such individual models, of differing form, each yielding a set of parameters and a likelihood at the individual level $L(C_n \mid m, \Omega_m)$. The set of M models could combine models using different distributions for random parameters, models with different socio-demographic specifications, models of different types, etc. The model averaging process then computes the overall likelihood for person n as the weighted average across M models, with:

$$L_n(\pi_n, \Omega) = \sum_{m=1}^{M} \pi_{m,n} L(C_n \mid m, \Omega_m), \tag{2}$$

where $\sum_{m=1}^{M} \pi_{m,n} = 1$ and $0 \leq \pi_{m,n} \leq = 1$. This overall likelihood is conditional on the vector of weights $\pi_n = \left\langle \pi_{1,n}, \ldots, \pi_{M,n} \right\rangle$ and the combined parameter estimates from the different models $\Omega = \langle \Omega_1, \ldots, \Omega_M \rangle$.

Of course, this structure takes the form of a latent class model, but two core differences apply.

Firstly, in a latent class model, an analyst simultaneously estimates the class allocation probabilities and the within class probabilities. In model averaging, individual models are estimated for the entire sample, and then the weights for these models are estimated, conditional on the parameters obtained during the individual model estimations. Model averaging is thus a sequential rather than simultaneous process. This is clearly computationally much easier, but also in fact allows

a situation where the individual models come from different teams of analysts. In fact, the estimation of the weights in Equation 2 does not require the parameters of the individual models, or even the mathematical formulation of the probabilities for individual models, but simply relies on the person-specific likelihoods obtained with the individual models. Model averaging will almost inevitably lead to a lower model fit than the estimation of a simultaneous structure, but of course the general situation is one where this simultaneous structure is not feasible to be estimated.

Secondly, in a latent class model, it is generally the case that the same overall model structure is used in different classes, though this is by no means necessary (cf. Hess et al., 2012). In model averaging, a different model specification, in terms of model structure and/or e.g. utility specification, is required for the different models as the separate estimation of the same structure for different m would of course yield the same fit and parameter estimates.

Just as in standard latent class approaches, it is entirely possible to specify a class allocation model for the model weights, i.e. making π_n a function of characteristics of the individual n.

2.2. Model averaging in application

The use of model averaging produces a new likelihood at the person level, $L_n(\pi_n, \Omega)$, where the overall model averaging log-likelihood (across N people) is given by:

$$LL(\Omega, \pi) = \sum_{n=1}^{N} log\left(L_n(\pi_n, \Omega)\right). \tag{3}$$

This overall log-likelihood will be at least as good as the log-likelihood for the best model out of the set of M different models. However, model fit alone is not the key reason for model averaging[1]. The output from the model averaging estimation process is a vector of weights for different models, where these are potentially individual specific, i.e. $\pi_{m,n}$ for person n and model m. These weights can then be used in model application, with two key uses.

Firstly, let $P_n(j \mid S, m, \Omega_m)$ give the probability of individual n choosing a specific alternative j out of a choice set S, conditional on model m, where this probability may again require integration. We can then compute the probability of this alternative j under model averaging as:

$$P_n(j \mid S, \pi_n, \Omega) = \sum_{m=1}^{M} \pi_{m,n} P_n(j \mid S, m, \Omega_m), \tag{4}$$

and an analyst can then for example use these weighted predictions in sample enumeration or other forecasting.

Secondly, let $W_n(j \mid S, m, \Omega_m)$ be some model output for individual n and choice set S, conditional on model m. This could for example be a willingness-to-pay (WTP) measure from model m or an elasticity measure. It is then similarly possible to compute a model average version of this

[1] For a full discussion of different model averaging methodologies and benefits, readers should refer to Claeskens and Hjort (2008).

output, using:

$$W_n(j \mid S, \pi_n, \Omega) = \sum_{m=1}^{M} \pi_{m,n} W_n(j \mid S, m, \Omega_m), \quad (5)$$

Any measures such as WTP or elasticities thus need to first be calculated for the individual models before being averaged across models. The calculation will likely differ across models and may involve simulation for some of the models if they incorporate random heterogeneity. If this is the case, it is advisable to use the entire distributions in model averaging rather than just relying on the moments from individual models if some non-normal distributions are included.

The key advantage of this process is that the calculation of these predictions or derived measures is informed by the results of a number of different models, and is thus potentially more robust to mis-specification of the individual models. It is similarly possible to compute variances for the outputs of Equation 3 and 5, though we rely just on the mean outputs in the present paper.

3. EMPIRICAL APPLICATION

In this section, we first give details on the different datasets used in this paper. Then, for each dataset, we demonstrate how model averaging can improve model fit. We then consider elasticities and willingness-to-pay outputs from model averaging.

3.1. Data

We use three different datasets for trialling model averaging. The first is a typical SP dataset with the latter two more complex RP datasets. We detail these first before applying model averaging across different models used on each dataset.

3.1.1. UK data

The first dataset that we consider involves public transport commuters living in the UK each making ten choices between three alternatives in a stated preference survey. A total of 368 participants completed the survey resulting in 3,680 choices. Each choice task involves an invariant reference trip and two hypothetical alternatives. Each alternative is described by travel time (in minutes), fare (in £), rate of crowded trips, rate of delays (both out of 10 trips), the average length of delays (across delayed trips) and the cost and availability of a delay information service (in £). Full details of the dataset are given by Hess and Stathopoulos (2013).

3.1.2. Sydney dataset

The second dataset that we use for model averaging comes from a Household Travel Survey (HTS-06) that was carried out in Sydney between 2004 and 2006 (Bureau of Transport Statistics, 2012). For this dataset, seven possible modes are established (car driver, car passenger, taxi, walk, bicycle, bus or train) and a large number of destination zones are defined (2,277 travel zones). For the

purposes of this paper we consider only 5,173 home-work tours. Level of service and attraction measures were assembled such that attributes could be derived for travel times, costs, waiting times and distances. For a full description of the data and its components, readers should refer to Fox (2015).

3.1.3. California dataset

The final dataset comes from the 2012 California Household Travel Survey (California Department of Transportation, 2013). For this dataset, there are 6,718 choices, with car, bus, rail and air as mode alternatives and 58 destination zones (the different counties in California). Again, we have attraction attributes for the different destinations and times, costs, and distances associated with the different travel modes.

3.2. Model averaging in estimation

Our first aim is to use model averaging to improve model fit. We can also test whether these models are overfitting through the use of validation subsets. Whilst there are many examples of possible uses of model averaging for travel behaviour modelling, we consider three different datasets to cover three common issues: the specification of heterogeneity, the definition of non-linearity, and the nesting structure in models for large-scale datasets.

3.2.1. Model averaging for specification of heterogeneity

A good first example of when model averaging might be useful is in the consideration of distributions for parameters within a mixed logit (MMNL) model. There is extensive literature on the choice of distributions and it is often clear that different specifications yield relatively similar fit but often substantially different model outputs, making the choice of a final distribution difficult for analysts (Börjesson et al., 2012; Hess et al., 2017), while the use of non-parametric distributions is still beyond the reach of most modellers despite recent innovations on this approach (Fosgerau and Mabit, 2013).

For the UK dataset, we first test the use of continuous distributions. For fare, time, crowding and rate of delays, we use either negative log-normal or negative log-uniform distributions depending on the model (see Table 1). We use negative log-normal distributions for the remaining four attributes. This results in 16 different MMNL models, for which the model fits are given in Table 1, where we also show the percentage of individuals whose choices are best described by each model (labelled as 'best fit' in Table 1).

The best fitting model here is version 15, which has negative log-uniform distributions for fare, time and crowding. A model with negative log-normal distributions for all parameters actually has better fit for more individual participants (13.59% compared to 7.07%) and consequently this model must have a larger range of fits for the individuals to have worse overall fit, which could be a result of long tails (Hess et al., 2017). However, more crucially, there is not much difference between the model fits and this means that there is scope for model averaging.

TABLE 1 : Log-likelihoods for 16 MMNLs with different combinations of distributions for the UK dataset

MMNL	Time	Fare	Crowding	Rate of delays	Log-likelihood	Best fit	MA Share
1	lognormal	lognormal	lognormal	lognormal	-3,034.16	**13.59%**	7.18%
2	lognormal	lognormal	lognormal	loguniform	-3,030.67	5.16%	0.00%
3	lognormal	lognormal	loguniform	lognormal	-3,019.60	4.62%	0.00%
4	lognormal	lognormal	loguniform	loguniform	-3,015.35	4.35%	0.00%
5	lognormal	loguniform	lognormal	lognormal	-3,027.83	7.34%	0.00%
6	lognormal	loguniform	lognormal	loguniform	-3,015.46	8.42%	7.84%
7	lognormal	loguniform	loguniform	lognormal	-3,001.06	3.80%	0.00%
8	lognormal	loguniform	loguniform	loguniform	-2,996.96	4.35%	3.26%
9	loguniform	lognormal	lognormal	lognormal	-2,982.40	6.79%	3.73%
10	loguniform	lognormal	lognormal	loguniform	-2,983.74	8.15%	15.21%
11	loguniform	lognormal	loguniform	lognormal	-2,980.24	5.43%	14.65%
12	loguniform	lognormal	loguniform	loguniform	-2,990.15	6.25%	0.00%
13	loguniform	loguniform	lognormal	lognormal	-2,982.85	4.08%	0.00%
14	loguniform	loguniform	lognormal	loguniform	-2,978.60	5.43%	9.70%
15	loguniform	loguniform	loguniform	lognormal	**-2,963.14**	7.07%	**34.70%**
16	loguniform	loguniform	loguniform	loguniform	-2,985.48	5.16%	3.73%

We apply model averaging across the 16 mixed logit models, i.e. estimating the 16 model specific weights[2], where we do not make these individual specific in our application. This results in a log-likelihood of -2,945, which as expected is better than that of any of the individual models. No formal statistical test is used here as it is not a process of simultaneously estimating all the parameters for all the models on a single dataset. The estimated weights are given in the 'MA share' column in Table 1. We see that the model with the best individual log-likeliood obtains the largest share but in addition see non-trivial shares for a substantial subset of other models. Crucially, this includes model 1, which had the worst individual fit, but also the largest share of respondents where this model produced the best fit out of all 16 models. This confirms that model averaging can be a successful approach for incorporating results from models that work well for only a subset of individuals. We also test to see whether the results from model averaging are overfitting by using out-of-sample validation. In this case, we split the dataset into five sections. For each section, we first estimate the parameters for all 16 mixed logit models individually on the first 80% of the data before calculating the log-likelihood of the remaining 20% validation set with the estimated parameters found for the initial set. We then apply model averaging across the 16 MMNLs for each of the five estimation subsets, before applying the resulting model averaging structure in each set to the appropriate holdout sample. The results of this are also shown in Table 2, where, for space reasons, we only ever show the fits for the five most contributing MMNLs in model averaging.

Across both the estimated and forecasted subsets, we consistently see that the model averaging approach has better fit than that of the best fitting MMNL model for each estimation subset. Note that across the five different subsets, four different combinations of distributions result in the best model fit (Models 14, 12, 15, 9 and 15 respectively across the different subsets). This highlights the difficult task of choosing distributions and further reinforces the potential benefits of model averaging. In addition, the MMNL model that offers the best performance in estimation is not the one with the best performance in the holdout sample in three out of five cases, while model averaging always produces a log-likelihood on the holdout that is at least as good as the best MMNL fit. As in the full sample, we again see that models that do not fit well across the subset can still contribute to the model average, with the best fitting model only twice receiving the largest share across the five subsets, and 13 out of the 16 models are at least once one of the top five contributors to the model average. Additionally, no single model is the largest contributor to more than one holdout subset.

3.2.2. Model averaging for linearity assumptions

We next test model averaging on revealed preference (RP) datasets, which can be more complex. As choices in RP data often include both mode and destination choice, the models that can be applied often have to be simpler due to both the vast size of some RP datasets and also due to the large number of alternatives generated if a modeller is trying to predict the precise zone or area an individual has chosen to travel to, together with the mode. Consequently large-scale models are often simple in structure as our usual more complex models such as mixed logit quickly become computationally infeasible. Model averaging avoids computational problems by creating a more

[2]We use a logit model for class allocation, with 15 constants estimated.

TABLE 2 : Estimation and holdout sample results for model averaging for the UK dataset

	Best individual MMNL		Estimation					Holdout Sample			
			Model averaging	Most contributing MMNLs			MA LL	Model averaging	Individual MMNLs		MA LL
	version	LL	LL	Version	LL	Share	Improvement	LL	Version	LL	Improvement
Full	15	-2,963	-2,945	15	-2,963	34.7%	18	n/a			
				10	-2,984	15.2%	39				
				11	-2,980	14.7%	35				
				14	-2,963	9.7%	18				
				6	-3,015	7.8%	70				
Holdout 1	14	-2,347	-2,327	11	-2,355	20.7%	28	-625	11	-632	7
				10	-2,350	18.6%	23		10	-637	12
				13	-2,354	14.5%	27		13	-629	4
				2	-2,390	10.7%	63		2	-653	28
				14	-2,347	9.8%	20		14	-629	4
Holdout 2	12	-2,405	-2,383	12	-2,405	24.7%	22	-558	12	-562	4
				9	-2,422	19.2%	39		9	-565	7
				16	-2,408	18.5%	25		16	-564	6
				6	-2,424	14.4%	41		6	-573	15
				3	-2,438	11.1%	55		3	-572	14
Holdout 3	15	-2,354	-2,326	16	-2,356	17.6%	30	-622	16	-626	4
				8	-2,356	15.0%	30		8	-631	9
				15	-2,354	13.9%	28		15	-629	7
				13	-2,369	12.5%	43		13	-627	5
				1	-2,413	12.1%	87		1	-633	11
Holdout 4	9	-2,362	-2,333	8	-2,362	24.1%	29	-615	8	-622	7
				9	-2,362	20.4%	29		9	-635	20
				3	-2,377	18.9%	44		3	-628	13
				15	-2,371	10.3%	38		15	-615	0
				12	-2,370	8.5%	37		12	-629	14
Holdout 5	15	-2,378	-2,347	15	-2,378	22.7%	31	-587	15	-595	8
				6	-2,388	22.5%	41		6	-597	10
				12	-2,396	10.8%	49		12	-601	14
				9	-2,392	8.7%	45		9	-593	6
				11	-2,381	7.3%	34		11	-597	10

complex model from averaging across a number of simpler models.

A key interest in large scale modelling is the specification of the utility function notably in terms of linearity assumptions (Daly, 2010; Stathopoulos and Hess, 2012). We therefore trial combinations of parameters for models on our Sydney HTS-06 data. In these models we use just the mode choice, for which there are a total of 5,173 choices, each with 7 alternatives. As there are a number of level of service attributes across the alternatives, we use four main attribute types and trial each with or without a logarithmic transformation applied to the set of attributes. The four parameter types that we consider are costs sensitivities (three different income groups), in-vehicle travel time sensitivities (bus, car, train, bus connection for train), other times sensitivities (access time, time until next service, time until subsequent service) and distance sensitivities (car, walking and bus distances). We additionally have a number of socio-demographic measures included in the specification of the models, which are based on a model for both mode and destination (detailed in Table 4.11 of Fox 2015). As we do not consider destination choice here, we do not use attraction variables. We trial all 16 different combinations of linear and logarithmic transformations of attributes. This gives us the model results displayed in Table 3.

TABLE 3 : Results from combinations of linear and logarithmic transformations of attributes on the Sydney HTS-06 mode choice data

Model	Cost	IVT	OT	Distance	Best fit	MA16 Share	Log-likelihood
1	linear	linear	linear	linear	5.2%	0.0%	-2,784.74
2	linear	linear	linear	log	5.5%	0.0%	-2,803.43
3	linear	linear	log	linear	6.5%	66.4%	-2,771.52
4	linear	linear	log	log	10.0%	0.0%	-2,792.17
5	linear	log	linear	linear	4.3%	0.0%	-2,806.83
6	linear	log	linear	log	4.7%	6.6%	-2,814.47
7	linear	log	log	linear	3.3%	0.0%	-2,800.51
8	linear	log	log	log	8.4%	0.0%	-2,804.25
9	log	linear	linear	linear	4.1%	0.0%	-2,801.99
10	log	linear	linear	log	1.6%	0.0%	-2,799.90
11	log	linear	log	linear	5.5%	0.0%	-2,791.18
12	log	linear	log	log	2.9%	7.7%	-2,792.10
13	log	log	linear	linear	6.1%	0.0%	-2,839.87
14	log	log	linear	log	6.4%	19.4%	-2,823.12
15	log	log	log	linear	5.6%	0.0%	-2,838.38
16	log	log	log	log	8.6%	0.0%	-2,818.69
Model averaging across 16 models							-2,750.49

The best performing individual model (model 3) comprises of linear costs, in-vehicle travel times and distances but a logarithmic transformation for other travel times. When applying model averaging across the 16 simpler models, this model obtains 66% of the allocation. Crucially, the improvement from model averaging across the simpler models is 21 log-likelihood units. Notably, the second largest share goes to model 14, which is an opposite to model 3, in that it has a logarithmic transformation for cost, in-vehicle travel times and distances but not for other travel times. The

results suggest that model 3 provides the best fit due to it providing a steady performance for each observation. Model 4, as a contrast, is the 2nd best fitting model for the largest number of choices, but overall performs worse than model 3 by 20 units, demonstrating that it predicts some choices very well and others very badly. Consequently, the joint model established from model averaging is far less sensitive to outliers, which only have a strong impact if they are not well described by any of the contributing models.

Again, we trial model averaging across models run on the full dataset as well as models run on 80% estimation subsets and 20% validation subsets (See Table 4).

TABLE 4 : Model averaging log-likelihoods across the attribute treatment combinations for estimation and holdout samples for the Sydney HTS-06 mode choice data

	Full	Subset 1	Subset 2	Subset 3	Subset 4	Subset 5
Best individual	-2,771.52	-2,230.51	-2,162.17	-2,215.39	-2,229.62	-2,230.61
Model averaging	-2,750.49	-2,216.40	-2,145.19	-2,198.72	-2,212.30	-2,208.86
Best individual (from estimation)	n/a	-544.80	-614.94	-561.49	-548.73	-553.30
Model averaging	n/a	-538.86	-610.92	-556.53	-546.15	-549.31

Across all five holdout samples, model 3 again performs best in estimation. This is very different from the case of the mixed logit examples discussed earlier. However, in line with previous results, we again find that estimation and holdout model fits are consistently improved by averaging across all 16 models.

3.2.3. Model averaging for nesting structure

We use our California dataset to test model averaging across models estimating both mode and destination choice. We start by using a multinomial logit model (MNL), two nested logit models (mode over destination, NL (M>D), and destination over mode, NL (D>M)) and a cross nested logit model (CNL). A full description of these models is given by Outwater et al. (2015). We then apply model averaging over these four models. Moving to increasingly complex models results in improvements in model fit and this pattern continues as we move to a model averaging approach (MA), which results in a substantial improvement in model fit over the cross nested logit model (see Table 5).

Whilst the majority of the model averaging share is given to cross nested logit, model averaging improves model fit by a further 53 log-likelihood units by also giving a substantial share to the nested logit model with mode over destination. The implications of using a combined model with CNL and this nested logit model are further investigated through consideration of averaging across the model outputs in terms of the differences in elasticities (see Section 3.3.2) and estimates for the value of travel time (see Section 3.4.3).

3.3. Elasticities from model averaging

Elasticities are a key output from choice models estimated on RP data. We now look at the implications of model averaging in this context. Given that elasticities from different models can

TABLE 5 : The results from model averaging (MA) across four basic models applied to the California dataset

Model	**MNL**	**NL (D>M)**	**NL (M>D)**	**CNL**	**MA**
Log-likelihood	-19,276	-19,271	-19,220	-19,152	-19,099
Improvement over MNL		5	56	124	177
Improvement over NL(D>M)			51	119	172
Improvement over NL(M>D)				68	121
Improvement over CNL					53
Model averaging share	0.01%	0.02%	35.57%	64.40%	
Proportion of best fits for individuals	5.59%	19.70%	20.66%	54.05%	

be very contrasting, a further use of model averaging is that it can be used to derive an 'average' single elasticity. A number of elasticities from different models can be used, with an appropriate weight for the relative importance/performance of the model included. We test this using our two revealed preference datasets.

3.3.1. Sydney elasticities

Our first test of elasticities from model averaging uses the Sydney mode only choice data. This is a particularly relevant example, as elasticities from models with a logarithmic transform for cost are often too low, whilst linear cost models are often too high (Fox et al., 2009). The elasticities for a 10% increase in car costs are shown for the 17 different models tested in Table 6. It is noticeable that, whilst many of the elasticities across the different models are similar, the values estimated for train, bus and walking vary more substantially. Unsurprisingly, models with a logarithmic transformation of costs (models 9-16) tend to estimate lower values for alternatives that cost money and higher values for alternatives that do not have an associated cost. This is particularly the case for train and bus, for which elasticity values estimated by models 9-16 are up to half those of values estimated by models 1-8. It is worth noting that as only mode choice is estimated here, the car elasticities observed are lower than those typically observed (see Fox (2015) for elasticities from models predicting mode and destination choice for this data). Whilst model averaging gives a larger share to linear cost models, lower elasticities for train and bus (relative to model 3, which is likely to have been used if outputs from a single model were to be chosen) are found for the model average. As a result, it appears that model averaging may be able to avoid the issues of finding elasticities that are either too high or too low.

3.3.2. California elasticities

We also test different elasticities for the California dataset, where we estimate car cost and time elasticities for trips, trip length and distance. For number of tours, average tour length and total

TABLE 6 : Elasticities for a 10% increase in car cost for the Sydney mode choice models.

			Elasticities						
Model	MA Share 16	Log-likelihood	Car Driver	Car Passenger	Train	Bus	Bike	Walk	Taxi
1	0.0%	-2,784.74	-0.11	0.13	0.29	0.24	0.15	0.05	0.15
2	0.0%	-2,803.43	-0.10	0.12	0.26	0.22	0.18	0.05	0.13
3	**66.4%**	**-2,771.52**	**-0.11**	**0.13**	**0.28**	**0.23**	**0.14**	**0.05**	**0.14**
4	0.0%	-2,792.17	-0.10	0.11	0.26	0.21	0.17	0.05	0.12
5	0.0%	-2,806.83	-0.14	0.15	0.35	0.30	0.18	0.06	0.16
6	**6.6%**	**-2,814.47**	**-0.12**	**0.14**	**0.29**	**0.25**	**0.17**	**0.06**	0.14
7	0.0%	-2,800.51	-0.14	0.15	0.35	0.29	0.17	0.06	0.16
8	0.0%	-2,804.25	-0.11	0.13	0.29	0.24	0.16	0.05	0.13
9	0.0%	-2,801.99	-0.05	0.09	0.09	0.11	0.14	0.10	0.13
10	0.0%	-2,799.90	-0.08	0.13	0.14	0.16	0.20	0.13	0.18
11	0.0%	-2,791.18	-0.05	0.08	0.09	0.10	0.12	0.09	0.12
12	**7.7%**	**-2,792.10**	**-0.07**	**0.12**	**0.13**	**0.15**	**0.19**	**0.12**	**0.17**
13	0.0%	-2,839.87	-0.06	0.11	0.11	0.13	0.16	0.11	0.15
14	**19.4%**	**-2,823.12**	**-0.08**	**0.13**	**0.14**	**0.17**	**0.20**	**0.13**	**0.18**
15	0.0%	-2,838.38	-0.06	0.09	0.10	0.11	0.14	0.10	0.13
16	0.0%	-2,818.69	-0.07	0.12	0.13	0.15	0.18	0.12	0.16
Model Averaging		**-2,750.49**	**-0.10**	**0.13**	**0.24**	**0.21**	**0.16**	**0.07**	**0.15**

distance respectively, we define:

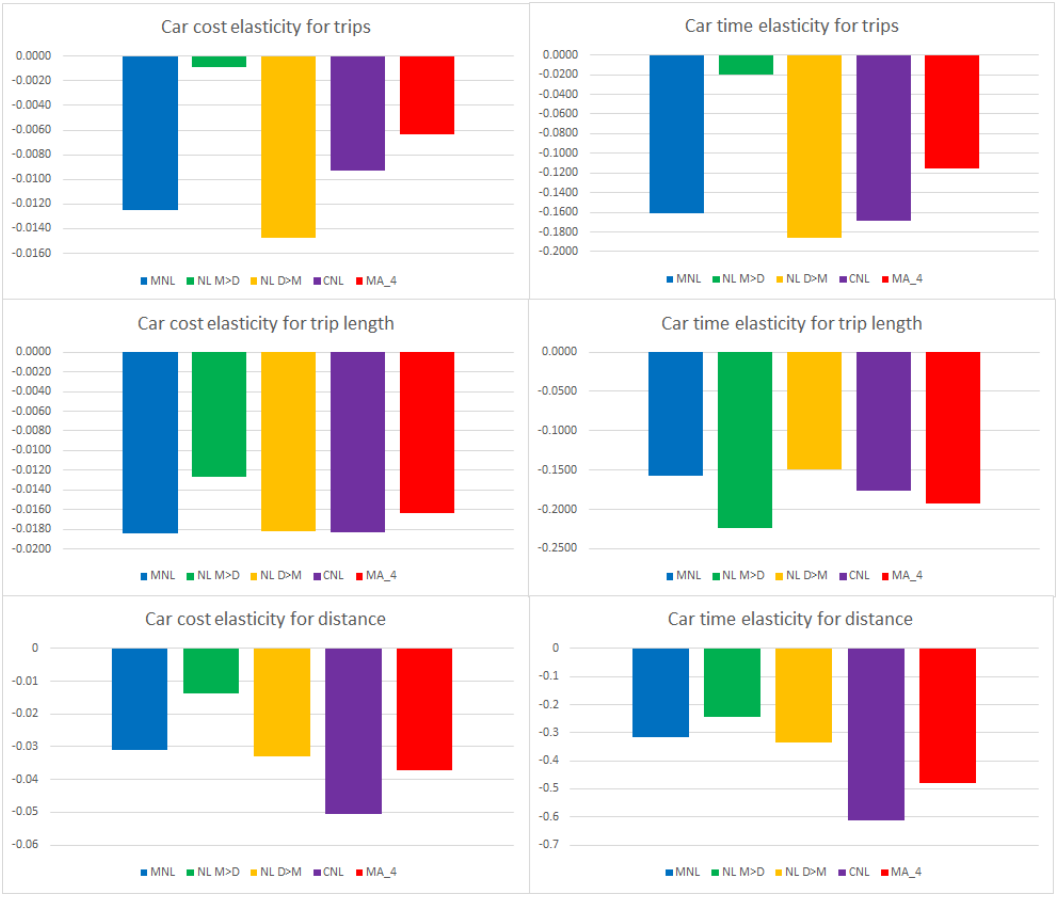
$$TourElasticity = log(\frac{ForecastedTours}{BaseTours})/log(1.1), \quad (6)$$

$$TourLengthElasticity = log(\frac{ForecastedTourLength}{BaseTourLength})/log(1.1), \quad (7)$$

$$DistanceElasticity = log(\frac{ForecastedTotalDistance}{BaseTotalDistance})/log(1.1). \quad (8)$$

Elasticities from the two best fitting candidate models (CNL and NL mode over destination, see Section 3.2.3) produce some very different[3] values (see the purple and green bars respectively in Figure 3). Consequently, if a single output for each elasticity is required, model averaging provides suitable values which take into account the relative performance of the different models.

FIGURE 1 : Elasticities from the four different candidate models and model averaging for the California dataset



[3]For a full review of the elasticities from these four candidate models, readers should refer to Outwater et al. (2015).

3.4. Willingness-to-pay outputs from model averaging

In this section, we explore welfare outputs from model averaging across the different models for the different datasets. We demonstrate how this can be done for both SP and RP datasets. For our SP dataset, we look at value of travel time as well as values for decreasing both the amount of crowding and the rate of delays. For our RP data, we take a more detailed look at the different estimates for the value of travel time.

3.4.1. Outputs from UK models

For the UK models, we first use the estimates from each of the 16 mixed logit models to obtain values[4] for the value of travel time (VTT, £/hour), value of crowding (VCR, amount paid in £ for 1/10 less crowded trips) and value of the rate of delays (VDE, amount paid in £ for 1/10 less delayed trips). We use the full distributions from the individual models in model averaging - the resulting means and standard deviations of these measures for each model and the model average is given in Table 7.

In comparison to the estimates obtained if we had simply used the best fitting mixed logit model (MMNL-15), results from model averaging suggest that the willingness to pay for changes in travel time and the rate of delays are not as high. The opposite is true for changes in the number of crowded trips, for which model averaging produces a higher estimate than MMNL-15. Notably, model averaging predicts a much wider standard deviation for the value of crowding.

3.4.2. Sydney VTT

Given that we use several different mode-specific travel time coefficients and three different income groups for our Sydney models, we can study a number of different travel time outputs from model averaging. We can compare the values for different groups of individuals as we have three cost coefficients in each model for three different income categories (1st: < $26k AUD, 2nd: $26-36.4k AUD, 3rd: > $36.4k AUD). We first obtain the value of travel time from all of the candidate models. As some of the models use logarithmic transformations for costs and times, we multiply these measures by a representative cost ($5.48) and divide by a representative time (49 minutes), as required. These outputs are detailed in Table 8.

It appears that, whilst the different models have fairly similar model fit, the value of travel times vary significantly, both across models and modes. The effect of income, however, is fairly consistent, with individuals of a higher income prepared to pay more to reduce time spent travelling. The difference between models is very significant, with the results from some models suggesting that individuals are willing to spend up to 10 times more than other models suggest. This means that if we were to pick a single model to use the outputs from, very different interpretations of the value of travel time could be made depending on which model is chosen. Without model averaging, it may be hard to know which models to rely more heavily on. The results from

[4]Note that as we use a logarithmic transformation for the cost attribute, we multiply values by 3, as this is the average cost of chosen alternatives (to the nearest pound).

TABLE 7 : Welfare measures obtained from the UK models

MMNL	TT	LF	CR	DE	Model LL	VTT mean	VTT sd	VCR mean	VCR sd	VDE mean	VDE sd	model share π_m
1	n	n	n	n	-3,034	3.390	6.812	0.291	0.725	0.265	0.926	7.2%
2	n	n	n	u	-3,031	3.332	5.866	0.321	0.921	0.244	0.723	0.0%
3	n	n	u	n	-3,020	3.099	5.538	0.381	1.275	0.231	0.675	0.0%
4	n	n	u	u	-3,015	3.382	6.532	0.396	1.367	0.274	0.820	0.0%
5	n	u	n	n	-3,028	3.195	4.755	0.366	1.252	0.258	0.868	0.0%
6	n	u	n	u	-3,015	3.113	4.399	0.441	2.866	0.228	0.438	7.8%
7	n	u	u	n	-3,001	3.672	6.221	0.375	0.813	0.298	1.052	0.0%
8	n	u	u	u	-2,997	2.945	4.226	0.319	0.695	0.185	0.468	3.3%
9	u	n	n	n	-2,982	3.809	7.620	0.297	0.803	0.205	0.506	3.7%
10	u	n	n	u	-2,984	3.945	8.084	0.338	0.964	0.231	0.593	15.2%
11	u	n	u	n	-2,980	3.871	8.407	0.405	1.376	0.228	0.625	14.6%
12	u	n	u	u	-2,990	3.962	9.251	0.330	0.969	0.265	0.772	0.0%
13	u	u	n	n	-2,983	3.624	5.730	0.305	0.882	0.256	1.076	0.0%
14	u	u	n	u	-2,979	3.612	5.889	0.310	0.876	0.193	0.348	9.7%
15	**u**	**u**	**u**	**n**	**-2,963**	**3.891**	**6.615**	**0.302**	**0.628**	**0.260**	**0.794**	**34.7%**
16	u	u	u	u	-2,985	3.691	6.406	0.325	0.659	0.199	0.433	3.7%
Model Averaging					**-2,945**	**3.737**	**6.951**	**0.335**	**1.019**	**0.235**	**0.658**	

TABLE 8 : Value of travel times (AUD/hr) obtained from the models for the Sydney choice-only data, across different modes and income categories.

			Car			Train			Bus			Access		
Model	LL	Share	1	2	3	1	2	3	1	2	3	1	2	3
1	-2,784.74	0.0%	9.10	12.33	15.33	2.42	3.28	4.08	7.20	9.75	12.12	3.50	4.74	5.89
2	-2,803.43	0.0%	7.77	10.98	13.93	0.12	0.17	0.22	6.15	8.68	11.02	3.40	4.80	6.10
3	**-2,771.52**	**66.4%**	**11.37**	**15.29**	**19.57**	**4.15**	**5.58**	**7.15**	**8.31**	**11.18**	**14.30**	**7.07**	**9.51**	**12.17**
4	-2,792.17	0.0%	9.68	13.51	17.65	1.70	2.37	3.10	7.05	9.84	12.86	7.08	9.88	12.91
5	-2,806.83	0.0%	1.02	1.31	1.56	4.99	6.39	7.61	0.25	0.32	0.38	1.86	2.39	2.84
6	-2,814.47	6.6%	4.37	5.88	7.16	4.46	6.01	7.31	2.12	2.85	3.47	2.77	3.73	4.54
7	-2,800.51	0.0%	2.59	3.32	3.96	4.70	6.03	7.20	0.64	0.82	0.98	5.35	6.85	8.19
8	-2,804.25	0.0%	6.88	9.20	11.34	3.84	5.14	6.33	2.99	4.01	4.94	6.86	9.18	11.31
9	-2,801.99	0.0%	20.88	21.92	21.90	3.34	3.51	3.51	13.68	14.36	14.35	8.15	8.55	8.55
10	-2,799.90	0.0%	11.46	11.91	12.04	0.31	0.33	0.33	7.91	8.22	8.30	4.80	4.98	5.04
11	-2,791.18	0.0%	27.14	28.21	28.40	7.46	7.75	7.80	17.21	17.89	18.00	15.81	16.43	16.54
12	-2,792.10	7.7%	14.35	14.80	15.06	2.52	2.60	2.64	9.58	9.88	10.05	9.53	9.83	10.00
13	-2,839.87	0.0%	3.80	3.94	3.92	11.81	12.25	12.19	-0.57	-0.59	-0.58	4.65	4.82	4.80
14	-2,823.12	19.4%	6.49	6.68	6.71	6.79	6.99	7.03	1.98	2.04	2.05	4.04	4.16	4.18
15	-2,838.38	0.0%	8.38	8.59	8.60	12.32	12.64	12.65	0.88	0.91	0.91	12.65	12.97	12.99
16	-2,818.69	0.0%	10.90	11.11	11.24	6.33	6.45	6.53	3.74	3.81	3.85	10.11	10.30	10.43
MA	**-2,750.49**		**10.19**	**12.96**	**15.91**	**4.56**	**5.66**	**6.79**	**6.77**	**8.76**	**10.89**	**6.39**	**8.12**	**9.95**

model averaging, however, appear reasonable.

3.4.3. California VTT

For our California data, we can calculate four mode-specific values of travel time from each of the different models. The results of these models are given in Table 9.

TABLE 9 : Value of travel time estimates by mode across the different models for the California data

		MNL	NL (D>M)	NL (M>D)	CNL	**MA**
Log-likelihood		-19,276	-19,271	-19,220	-19,152	**-19,099**
VTT	car	73.10	71.70	135.94	88.09	**105.11**
	bus	78.75	77.66	141.72	98.33	**113.76**
	rail	72.15	73.15	67.41	77.68	**74.03**
	air	24.01	25.85	3.84	42.40	**28.68**
Model share		0.01%	0.02%	35.57%	64.40%	

In this case, as there are only two models that contribute to the model average, model averaging provides a value that is close to halfway between the estimates for the value of travel time from the CNL model and the nested logit model with mode over destination. For air, this results in an estimate that is actually closer to the non-contributing MNL value.

4. CONCLUSIONS

Despite successful results in a number of fields including health, ecology and economics, model averaging has yet to make a transition into mainstream choice modelling. In this paper, we demonstrate that it is very simple to run and that it consistently improves model fit in both estimation and forecasting. Whilst we apply model averaging through the use of sequential latent class models, other methods are possible, with Bayesian methods used for model averaging typical in other disciplines (Wintle et al., 2003; Wang et al., 2004; Raftery et al., 2005). Consequently, future work could compare different model averaging methods. However, we find that model averaging using a simple sequential latent class structure provides many benefits.

We demonstrate that model averaging can be applied across a large number of candidate models. These models can be very similar, with model averaging proving effective when used across multiple mixed logit models with various different combinations of distributions for the parameters. The models can also be more different, such as in our nesting structures for large scale modelling. With complex models often infeasible to run when there are hundreds or even thousands of alternatives, model averaging provides a simple and efficient method for improving models, with consistent improvements in model fit found when applying it over a number of simple models.

Additionally, model averaging is less sensitive to outliers, as unlikely choices only have an impact on the model fit if they are outliers across all models contributing to the model average. This also means that model averaging is very good at making the most of models which are very accurate at describing some choices but less accurate for others. Consequently, the best fitting model may not contribute to a model average.

We show that model averaging always provides model fit at least as good as the best fitting candidate model. We have purposefully not conducted statistical tests for these improvements in fit. Indeed, model averaging should not be seen as a different model which can be compared to individual structures, such as a simultaneous latent class model with different models in each class. Indeed, for model averaging, the process only involves calculating a weighted average of the outputs from individual models and does not involve the reestimation of the parameters from the individual models, where these always come from individual models estimated on the full sample.

Whilst we only ever consider the use of constants for class allocation, more complex structures could easily be adopted. For example, the parameterisation of class allocation within model averaging could be performed very simply by using socio-economic attributes. A final key advantage of model averaging is that it is very easy to apply. A modeller does not even require knowledge of the individual models within the classes to apply model averaging. This means that, for example, practitioners could ask multiple researchers to apply models to the same dataset and then average across the models, for which they would only need the underlying log-likelihood contribution for each individual or observation in the dataset. This may go some way to mitigating risk, as well as having a chance of improving the model. Consequently, there are many advantages to be gained by applying model averaging for both applied and theoretical transport behaviour modellers.

ACKNOWLEDGEMENTS

Thomas Hancock and Stephane Hess would like to acknowledge the financial support by the European Research Council through the consolidator grant 615596-DECISIONS.

REFERENCES

Börjesson, M., Fosgerau, M., and Algers, S. (2012). Catching the tail: Empirical identification of the distribution of the value of travel time. *Transportation Research Part A: Policy and Practice*, 46(2):378–391.

Bureau of Transport Statistics (2012). Household Travel Survey 2010/11: Technical Documentation. *Bureau of Transport Statistics, Transport for New South Wales*.

California Department of Transportation (2013). 2010-2012 California Household Travel Survey Final Report.

Claeskens, G. and Hjort, N. L. (2008). Model selection and model averaging. Technical report, Cambridge University Press.

Daly, A. (2010). Cost damping in travel demand models.

Fosgerau, M. and Mabit, S. L. (2013). Easy and flexible mixture distributions. *Economics Letters*, 120(2):206–210.

Fox, J. (2015). *Temporal transferability of mode-destination choice models*. PhD thesis, University of Leeds.

Fox, J., Daly, A., and Patruni, B. (2009). Improving the treatment of cost in large scale models. In *European Transport Conference*. Citeseer.

Gazder, U. and Ratrout, N. T. (2015). A new logit-artificial neural network ensemble for mode choice modeling: a case study for border transport. *Journal of Advanced Transportation*, 49(8):855–866.

Hess, S., Daly, A., Dekker, T., Cabral, M. O., and Batley, R. (2017). A framework for capturing heterogeneity, heteroskedasticity, non-linearity, reference dependence and design artefacts in value of time research. *Transportation Research Part B: Methodological*, 96:126–149.

Hess, S. and Stathopoulos, A. (2013). A mixed random utility - random regret model linking the choice of decision rule to latent character traits. *Journal of Choice Modelling*, 9:27–38.

Hess, S., Stathopoulos, A., and Daly, A. (2012). Allowing for heterogeneous decision rules in discrete choice models: an approach and four case studies. *Transportation*, 39(3):565–591.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical science*, pages 382–401.

Morales, K. H., Ibrahim, J. G., Chen, C.-J., and Ryan, L. M. (2006). Bayesian model averaging with applications to benchmark dose estimation for arsenic in drinking water. *Journal of the American Statistical Association*, 101(473):9–17.

Moretti, F., Pizzuti, S., Panzieri, S., and Annunziato, M. (2015). Urban traffic flow forecasting through statistical and neural network bagging ensemble hybrid modeling. *Neurocomputing*, 167:3–7.

Outwater, M. L., Bradley, M., Ferdous, N., Bhat, C., Pendyala, R., Hess, S., Daly, A., and LaMondia, J. (2015). Tour-based national model system to forecast long-distance passenger travel in the united states. Technical report.

Posada, D. and Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic biology*, 53(5):793–808.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005). Using bayesian model averaging to calibrate forecast ensembles. *Monthly weather review*, 133(5):1155–1174.

Sloughter, J. M., Gneiting, T., and Raftery, A. E. (2010). Probabilistic wind speed forecasting using ensembles and bayesian model averaging. *Journal of the american statistical association*, 105(489):25–35.

Stathopoulos, A. and Hess, S. (2012). Revisiting reference point formation, gains–losses asymmetry and non-linear sensitivities with an emphasis on attribute specific treatment. *Transportation Research Part A: Policy and Practice*, 46(10):1673–1689.

Volinsky, C. T., Madigan, D., Raftery, A. E., and Kronmal, R. A. (1997). Bayesian model averaging in proportional hazard models: assessing the risk of a stroke. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(4):433–448.

Wang, D., Zhang, W., and Bakhai, A. (2004). Comparison of bayesian model averaging and stepwise methods for model selection in logistic regression. *Statistics in medicine*, 23(22):3451–3467.

Wintle, B. A., McCarthy, M. A., Volinsky, C. T., and Kavanagh, R. P. (2003). The use of bayesian model averaging to better represent uncertainty in ecological models. *Conservation Biology*, 17(6):1579–1590.

Wright, J. H. (2009). Forecasting us inflation by bayesian model averaging. *Journal of Forecasting*, 28(2):131–144.