Article:

Hancock, TO, Hess, S orcid.org/0000-0002-3650-2518, Daly, A et al. (1 more author) (2020) Using a sequential latent class approach for model averaging: Benefits in forecasting and behavioural insights. Transportation Research Part A: Policy and Practice, 139. pp. 429-454. ISSN 0965-8564

https://doi.org/10.1016/j.tra.2020.07.005

USING A SEQUENTIAL LATENT CLASS APPROACH FOR MODEL AVERAGING: BENEFITS IN FORECASTING AND BEHAVIOURAL INSIGHTS

Thomas O. Hancock (Corresponding Author)
Choice Modelling Centre & Institute for Transport Studies
University of Leeds
tratoh@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’. There are different ways of doing this, with a clear distinction between looking at the overall performance of each model or by doing this at the level of individual agents or observations. In this paper, we demonstrate that a relatively straightforward adaptation of latent class models can be used for the latter approach and show how this can be an effective method 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 using a sequential latent class framework 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. In terms of actual implementation of model averaging, we present a simple expectation-maximisation (EM) algorithm which can deal with very large numbers of candidate models within the same model averaging structure, unlike the typical case with classical estimation approaches for latent class.

1. INTRODUCTION

Travel behaviour modelling, and choice modelling in particular, places great emphasis on model specification, with studies often comparing and contrasting several mutually exclusive model structures on the same data. These comparisons are carried out on the basis of mathematical fit to the data, theoretical consistency with underlying hypotheses, and reasonableness of the substantive model outputs. This process often does not lead to a clear “winner”, and analysts need to make a highly consequential decision on which structure to put forward as the final model. Even in situations where one model is “superior” to the others overall, it is quite reasonable to expect that this model may be inferior to some of the rejected models for at least a non-trivial subpart of the data.

Other fields have tended to avoid this winner-take-all approach in model selection by using techniques commonly referred to as model averaging or ensemble methods. These can be used to allow a modeller to establish a single model by calculating relative contribution weights for a set of candidate models, with the underlying idea of allowing multiple competing structures to contribute to the final model/results. 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) or a coronary event (Wang et al., 2004), and to understand the relation between arsenic levels and cancer rates (Morales et al., 2006). Bayesian model averaging is 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). 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). A key advantage of model averaging is that the structures that contribute to the model average can be very different from each other, using diverse methods, and even be produced by different sets of researchers.

The use of model averaging in choice modelling is far more limited, despite some promising work. Rose et al. (2009) for example demonstrate how model averaging can be used in the generation of efficient stated choice experiments. Furthermore, there has also been work comparing the use of different model averaging procedures for the allocation of model weights for multinomial and ordered logit models, with Zhao et al. (2019) developing a method based on cross-validation. Other alternatives include asymptotically optimal model averaging Wan et al. (2014) (which has been extended for use in MACML-estimated multinomial probit models by Batram and Bauer 2017) and Bayesian model averaging over multinomial logit models (Sevcikova and Raftery, 2013). However, despite the variety of methods for model averaging and the fact that it can combine the benefits from a number of models into one model, it is not yet common practice within choice modelling.

The lack of previous applications of model averaging in transport behaviour research specifically may in part be due to a lack of understanding about model averaging methods. Firstly, model averaging may be perceived to be a complex undertaking as analysts may not understand that a joint estimation of the overall model is not required. Secondly, the decision on the approach used for determining the weights of individual models to the overall structure may be seen as arbitrary. This is a perfectly understandable concern in the situation where model averaging is carried out using sample level (aggregate) measures of fit, i.e. assigning model weights based on aggregate model measures such as AIC or BIC[1]. Finally, the main use of model averaging in other fields has been for prediction and there has thus far not been an emphasis on the fact that standard outputs such as elasticities and welfare measures can still be provided after the use of model averaging.

This paper addresses all three of the issues above by relying on a sequential latent class approach for model averaging. Individual model structures are estimated on the full sample and their individual-level contributions to the overall sample level likelihood are then used in a latent class structure that only estimates class allocation probabilities. The sequential nature of the approach addresses the concern about complexity. The reliance on individual-level probabilities within a latent class model means that no arbitrary decisions are required on how weights should be calculated, instead making use of maximum likelihood estimation. A key benefit of this approach is that models that work well for subgroups of decision-makers but offer poor overall performance can still contribute to the model average. Finally, the class allocation probabilities produced by the sequential latent class structure can then also be used to produce weighted averages of other model outputs such as willingness-to-pay measures or elasticities.

We consider two key cases that occur frequently within travel behaviour modelling for which there is clear scope for the introduction of this type of model averaging. The first is to apply model

[1] Akaike and Bayesian Information Criterion, respectively.

averaging across multiple candidate models that all have advantages and disadvantages, where there is no clear cut case for choosing which is best. One obvious example of this is in the choice of distribution(s) within mixed logit models (Guo and Wilson, 2007; Hess, 2010; Tjiong, 2015). A second and rather different context in which the benefits of implementing model averaging are clear is in the case of very large-scale applications, either with large datasets or large choice sets. In these cases, the use of complex models may not be possible for computational reasons, and combining several simpler models may have benefits. Obvious examples of this includes choice modelling applications where we aim to predict both travel mode and destination (Fox, 2015; Outwater et al., 2015) or choice modelling in the context of big data (Zannat and Choudhury, 2019; Tang et al., 2020).

The remainder of this paper is organised as follows. First, we present a methodology section demonstrating how we apply model averaging with sequential latent class models and demonstrate how to produce outputs such as elasticities from model averaging. This is followed by three separate sections with empirical work on three different datasets. The final section summarises our findings and presents directions for future research.

2. METHODOLOGY

In this section, we discuss how model averaging can be carried out using a simple sequential latent class approach. We look separately at estimation and application.

2.1. Model averaging in estimation

Let us assume that we have a dataset containing the choices made by N different individuals, where individual n makes T_n separate choices, with $T_n \geq 1$.

To apply model averaging, we first determine a set of M different candidate models that are suitable for the data at hand. These differences between the models can arise for a variety of reasons. In the simplest form, they could relate to the specification of the value functions (such as utility), for example using different socio-demographic interactions, different treatments of non-linearity, or different specifications of random heterogeneity. The differences could be more fundamental than that, with differences in the actual model structure, for example looking at different models from the family of Generalised Extreme Value (GEV) models. Finally, the models could be based on different behavioural paradigms, for example looking at random utility maximisation, random regret minimisation, etc.

In the most standard approach, a single analyst (or team of analysts) will then estimate the M different models on the data. This is not an actual requirement of model averaging, with the possibility that the different models are contributed by different teams, which is an inherent strength of the approach as this can lead to a more heterogeneous set of inputs into the model averaging process.

Ignoring the possibility of non-parametric models[2], the estimation of each one of the M models will involve finding the values for the parameters of that model which maximise the likelihood of the choices C in the data. We would have that, for model m:

$$L_m(C \mid \Omega_m) = \prod_{n=1}^{N} L_m(C_n \mid \Omega_m) \tag{1}$$

where C_n is the set of choices for individual n. The specific functional form for $L_m(C_n \mid \Omega_m)$ will vary across models[3].

At convergence, model m will give us a set of estimates $\widehat{\Omega_m}$ for the vector Ω, such that:

$$\widehat{\Omega_m} = \underset{\Omega \in \Theta}{\arg\max}\, L_m(C \mid \Omega_m), \tag{2}$$

where Θ is the set of real numbers. Estimation will yield M sets of vectors of optimal parameters, i.e. $\widehat{\Omega_m}$ for model m, as well as M measures of mathematical fit to the data, i.e. $L_m\left(C \mid \widehat{\Omega_m}\right)$ for model m, obtained by using $\widehat{\Omega_m}$ in Equation 1.

In its simplest form, model averaging would involve computing weights for each of the M models as a function of the relative differences across models in $L_m\left(C \mid \widehat{\Omega_m}\right)$, or some other measure of model fit. This however looks only at fit at the sample level, and ignores the possibility that different models will work differently well for individual people in a sample population.

In the sequential latent class approach, we instead rely on the likelihood at the level of individual decision-makers. In particular, we have that, at the estimated set of parameters Ω_m, the likelihood of the observed choices for person n, using model m, is given by $L_m\left(C_n \mid \widehat{\Omega_m}\right)$. We group together the estimates from the M different models, giving $\widehat{\Omega} = \left\langle \widehat{\Omega_1}, \ldots, \widehat{\Omega_M} \right\rangle$. The likelihood function for the model averaging structure is then given by:

$$L_{MA}\left(C \mid \pi, \widehat{\Omega}\right) = \prod_{n=1}^{N} \sum_{m=1}^{M} \pi_{m,n} L_m\left(C_n \mid \widehat{\Omega_m}\right), \tag{3}$$

where $\pi_{m,n}$ is an estimated weight for model m for person n, $\sum_{m=1}^{M} \pi_{m,n} = 1$ and $0 \leq \pi_{m,n} \leq 1$. With $\pi_n = \left\langle \pi_{1,n}, \ldots, \pi_{M,n} \right\rangle$ representing the weights for person n, we have that $\pi = \langle \pi_1, \ldots, \pi_N \rangle$.

[2] Such models can also be used in model averaging of the type described here if they can provide likelihoods at the individual level, but their presence in the model average will preclude the calculation of other possible outputs, such as willingness-to-pay measures, although the same also applies for some parametric models, if they are not grounded in the appropriate behavioural paradigm.

[3] For example, if model m is of the mixed logit type, we would have $L_m(C_n \mid \Omega_m) = \int_{\beta_m} \prod_{t=1}^{T_n} P_m\left(j^*_{n,t} \mid \beta_m\right) f_m(\beta_m \mid \Omega_m)\, \mathrm{d}\beta_m$. In this example, 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 which would be of the Multinomial Logit type, where the parameters β_m are distributed according to $f_m(\beta_m \mid \Omega_m)$.

In many applications of model averaging, π_n would be the set to be the same $\forall n$, though it is easily possible to link π_n to characteristics of person n in estimation.

Of course, the likelihood of the model averaging structure in Equation 3 is dependent on the vector $\widehat{\Omega}$. This combines the estimates from the M different models contributing to the model average. Crucially, in estimation of the model averaging structure, these parameters are kept fixed at the estimates from the individual models, hence the $\widehat{}$ notation, and only π is estimated.

This last point provides the key contrast between the sequential latent class approach used for model averaging and the typical simultaneous approach used in standard latent class applications. In the latter, an analyst simultaneously estimates the class allocation parameters ($\pi_{m,n}$) and the parameters driving 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 3 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 offer a model fit that is bounded below by the fit of the best fitting of the M 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 often difficult or impossible to estimate.

A further difference arises in that, in a simultaneous 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.

Model averaging such as discussed here can be carried out using any package capable of latent class estimation, where for all models, we use Apollo (Hess and Palma, 2019). Latent class models are well known to have complex likelihood function that can lead to problems with convergence to poor local optima. While this issue is alleviated to some extent with the sequential latent class approach used in model averaging, care is still required, and we advocate the use of an expectation-maximisation (EM) approach rather than using classical estimation. For a detailed discussion of EM algorithms, see Train (2009, ch. 14). In our case, we rely on a class allocation model without covariates, i.e. $\pi_{m,n} = \pi_m, \forall n$, making the use of an EM approach especially straightforward. In particular, the following iterative process is used:

1. Definition of starting model weights π_m, where we set these to $\pi_m = \frac{1}{M}, \forall m$.

2. Calculate likelihood of the model, using equation 3 and store this as L_1, i.e.

$$L_{MA}^{(1)}\left(C \mid \pi, \widehat{\Omega}\right) = \prod_{n=1}^{N}\sum_{m=1}^{M}\pi_m L_m\left(C_n \mid \widehat{\Omega_m}\right) \tag{4}$$

3. Calculate posterior model weights for each individual conditional on the model specific likelihoods for that individual, using:

$$h_{m,n} = \frac{\pi_m L_m\left(C_n \mid \widehat{\Omega_m}\right)}{\sum_{m=1}^{M}\pi_m L_m\left(C_n \mid \widehat{\Omega_m}\right)} \tag{5}$$

4. Update the model weights as follows:

$$\pi_m = \frac{\sum_n^N h_{m,n}}{\sum_n^N \sum_{m=1}^{M} h_{m,n}} \tag{6}$$

5. Calculate likelihood of the model with new model weights, using equation 3 and store this as L_2

6. If $L_2 - L_1$ is less than a predefined limit (we chose 10^{-5}), convergence has been reached. Otherwise, return to step 2 with the new values for π

We use a two-stage implementation of this algorithm. After completing the original algorithm, there is a possibility of some models being retained in the model averaging with very low weights, i.e. not contributing in any meaningful manner. We eliminate any models that obtain less than a 1% share in the first round, and repeat the above algorithm until convergence a second time with the reduced set of models.

2.2. Model averaging in application

To use model averaging in application, we rely on the estimates for the model averaging weights, i.e. $\widehat{\pi}$ obtained by maximising Equation 3, i.e.:

$$\widehat{\pi} = \arg\max_{\pi\in\Theta} \widehat{L}_{MA}\left(C \mid \pi, \widehat{\Omega}\right), \tag{7}$$

where this itself is conditional on the estimates $\widehat{\Omega}$ obtained by optimising the M individual models.

In application, we use $\widehat{\pi} = \left\langle \widehat{\pi_1}, \ldots, \widehat{\pi_N} \right\rangle$ and $\widehat{\Omega}$. If $\widehat{\pi}$ is generic, i.e. not linked to the characteristics of individual decision-makers, we have $\widehat{\pi_n} = \widehat{\pi}\,\forall n$, and the application to a sample different from that used in estimation does not necessitate any additional steps. If $\widehat{\pi}$ is a function of characteristics of the decision-makers, i.e. $\pi_n = f\left(\widehat{\gamma}, z_n\right)$, where $\widehat{\gamma}$ is estimated during the model averaging, then individual-level weights simply need to be computed for the application sample of decision makers.

The most obvious use of model averaging in application concerns forecasting of choices. Let $P_m\left(j_n \mid S_n, \widehat{\Omega_m}\right)$ give the probability of individual n choosing a specific alternative j out of a choice set S_n, conditional on model m, where S_n describes the characteristics of the alternatives faced by person n, where these could be different from the levels used in estimation. The calculation of $P_m\left(j_n \mid S_n, \widehat{\Omega_m}\right)$ is thus equivalent to making a forecast of a choice probability with a single model m. We can then compute the probability of this alternative j under model averaging as:

$$P_{MA}\left(j_n \mid S_n, \widehat{\pi_n}, \widehat{\Omega}\right) = \sum_{m=1}^{M} \widehat{\pi_{m,n}} P_m\left(j_n \mid S_n, \widehat{\Omega_m}\right), \tag{8}$$

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

When producing forecasts, we thus use the actual model averaging structure for the forecasts, combining predictions from individual models and averaging across those, using the weights obtained by Equation 7. Elasticities and other measures related to changes in demand thus need to be calculated on the basis of these weighted predictions, rather than by looking at changes in demand from individual models. To explain this further, imagine a situation where we want to study the impact of a change in a given attribute. In our notation, this would lead to a new definition of the choice set, say S'_n, rather than S_n. To study the impact of this change, say on the demand for alternative j at the sample level, we would look at the predicted change in demand, relative to the original demand, i.e.:

$$\begin{aligned}\Delta_j &= \frac{\sum_{n=1}^{N}\left(P_{MA}\left(j_n \mid S'_n, \widehat{\pi_n}, \widehat{\Omega}\right) - P_{MA}\left(j_n \mid S_n, \widehat{\pi_n}, \widehat{\Omega}\right)\right)}{\sum_{n=1}^{N}\left(P_{MA}\left(j_n \mid S'_n, \widehat{\pi_n}, \widehat{\Omega}\right)\right)} \\ &= \frac{\sum_{n=1}^{N}\left(\sum_{m=1}^{M}\widehat{\pi_{m,n}} P_m\left(j_n \mid S'_n, \widehat{\Omega_m}\right) - \sum_{m=1}^{M}\widehat{\pi_{m,n}} P_m\left(j_n \mid S_n, \widehat{\Omega_m}\right)\right)}{\sum_{n=1}^{N}\left(\sum_{m=1}^{M}\widehat{\pi_{m,n}} P_m\left(j_n \mid S'_n, \widehat{\Omega_m}\right)\right)}.\end{aligned} \tag{9}$$

The reader will note that this is different from using the weighted average of the relative changes, i.e.:

$$\overline{\Delta_j} = \sum_{m=1}^{M} \widehat{\pi_{m,n}} \frac{\sum_{n=1}^{N}\left(P_m\left(j_n \mid S'_n, \widehat{\Omega_m}\right) - P_m\left(j_n \mid S_n, \widehat{\Omega_m}\right)\right)}{\sum_{n=1}^{N} P_m\left(j_n \mid S_n, \widehat{\Omega_m}\right)}. \tag{10}$$

Equation 9 is looking at the change in demand predicted by the final model averaging structure, while Equation 10 looks at the weighted average of predicted changes across the individual model components. The former measure is the one in line with the notion of model averaging.

Aside from predictions, the other key post estimation output of a choice model is the computation of marginal rates of substitution (MRS), which, if the denominator is a cost sensitivity, give us willingness-to-pay (WTP) measures. Let $W_m\left(n \mid S_n, \widehat{\Omega_m}\right)$ be some model output for individual n and choice set S_n, conditional on model m and the estimated parameters for that model[4]. It is

[4] In many cases, the dependence on S_n will not apply and is only shown here for the sake of generality.

then similarly possible to compute a model average version of this output, using:

$$W_{MA}\left(n \mid S_n, \widehat{\pi}_n, \widehat{\Omega}\right) = \sum_{m=1}^{M} \widehat{\pi_{m,n}} W_m \left(j_n \mid S_n, \widehat{\Omega_m}\right), \tag{11}$$

Any measures such as WTP thus need to be calculated first 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 8 and 11, though we rely only on the mean outputs in the present paper.

3. APPLICATION TO SP ROUTE CHOICE DATA

Our first application makes use of a typical stated preference (SP) dataset, where our focus for model averaging is on combining the results from multiple Mixed Multinomial Logit (MMNL) models making different assumptions about the shape of distribtions for random heterogeneity in sensitivities across respondents. 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 seminal innovations on this approach (Fosgerau and Mabit, 2013).

3.1. Data

The dataset that we consider involves public transport commuters living in the UK each making ten choices between three alternatives in a SP survey. A total of 368 participants completed the survey resulting in 3,680 choices. Each choice task includes an invariant reference trip (with the attribute values collected before the decision-maker completes the SP questions) and two hypothetical alternatives with attribute values that are pivoted around those of the reference trip. In the scenarios, each alternative was described by six attributes: travel time (in minutes), fare (in £), rate of crowded trips (frequency of having to stand out of 10 trips), rate of delays (frequency of delays out of 10 trips), the average length of delays (across delayed trips) and the provision of a delay information service (either not available, available at a small cost of £0.30 per journey, or available for free). Full details of the dataset are given by Hess and Stathopoulos (2013).

3.2. Specification of individual components and model averaging

In our MMNL models, the utility (net of the extreme value error term) for alternative i in choice task t for individual n is specified as:

$$\begin{aligned} V_{int} = \delta_i + \beta_{n,TT} \cdot TT_{int} + \beta_{n,LF} \cdot log(F_{int}) + \beta_{n,CR} \cdot CR_{int} + \beta_{n,RD} \cdot RD_{int} \\ + \beta_{n,AD} \cdot AD_{int} + \beta_{n,ED} \cdot ED_{int} + \beta_{n,CI} \cdot CI_{int} + \beta_{n,FI} \cdot FI_{int}. \end{aligned} \tag{12}$$

In this specification, we include alternative specific constants (ASC) for alternatives 1 and 2, i.e. δ_1 and δ_2, i.e. fixing $\delta_3 = 0$. We use the continuous attributes of travel time (TT), the natural logarithm of fare (F) given earlier findings about non-linear response, crowding (CR), rate of delays (RD) and average delays (AD). In addition, we include a new variable of expected delay (ED), which is the interaction between RD and AD. Finally, the delay information attribute is dummy coded, where we set the base of no information service to zero, and estimate an effect for the charged information service (CI) and the free information service (FI).

We estimate $M = 16$ different models on this data. In each model, the ASCs are kept fixed across respondents, i.e. not random. The eight marginal utility (β) coefficients are allowed to vary randomly across respondents. Our focus in testing the distributional assumptions (and thus the use of model averaging) is on the first four attributes, where we look at all 16 possible combinations of negative lognormal and negative loguniform distributions, i.e. distributions where the logarithm of the negative of the parameter follows a normal or uniform distribution. For example, in specification 1, all four are negative lognormal (LN-), in specification 2, the rate of delays is negative loguniform (LU-), and the other three are negative lognormal, etc. For the remaining four attributes (ED, AD, CI, FI), we always use lognormal distributions. For ED and AD, the use of a negative lognormal distributions is an obvious choice, given the undesirable nature of these attributes. Preliminary tests also showed that respondents preferred the absence of a delay information service to a charged one, so that a negative lognormal distribution was used for CI, and a positive lognormal (LN+) distribution for FI.

The model fits for the 16 different MMNL models are given in Table 1, where the corresponding model parameter estimates are given in the Appendix, in Table A1[5]. The best fitting model across the different specifications is version 15, which has negative loguniform distributions for fare, time and crowding. Table 1 also shows the percentage of individuals whose choices are best described by each model (labelled as 'Best model for x% of respondents' in Table 1). The model using negative lognormal distributions for all parameters actually has the worst sample level fit but obtains the best fit for more individual participants than any other model. This, together with the small overall differences between the sample level model fits supports the hypothesis of different models working differently well for different individuals and means that there is clear scope for model averaging.

We next apply model averaging across the 16 mixed logit models, i.e. estimating the 16 model specific weights using the EM algorithm discussed earlier. The use of model averaging results in a

[5] With both lognormal and loguniform distributions, the analyst estimates the parameters for the distribution of the logarithm of the marginal utility coefficient (possibly of the negative of that coefficient). In Table A1, we then show par_1 and par_2, where e.g. for the first block of parameters, these are for the distribution of $log(-\beta_{n,TT})$.

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

Model	Type of distribution				Overall	Best model for	MA
	Time	Fare	Crowding	Rate of delays	Log-likelihood	x% of respondents	Share
1	LN-	LN-	LN-	LN-	-3,034.16	**13.59%**	7.59%
2	LN-	LN-	LN-	LU-	-3,030.67	5.16%	0.00%
3	LN-	LN-	LU-	LN-	-3,019.60	4.62%	0.00%
4	LN-	LN-	LU-	LU-	-3,015.35	4.35%	0.00%
5	LN-	LU-	LN-	LN-	-3,027.83	7.34%	0.00%
6	LN-	LU-	LN-	LU-	-3,015.46	8.42%	8.16%
7	LN-	LU-	LU-	LN-	-3,001.06	3.80%	0.00%
8	LN-	LU-	LU-	LU-	-2,996.96	4.35%	3.26%
9	LU-	LN-	LN-	LN-	-2,982.40	6.79%	1.90%
10	LU-	LN-	LN-	LU-	-2,983.74	8.15%	16.17%
11	LU-	LN-	LU-	LN-	-2,980.24	5.43%	14.87%
12	LU-	LN-	LU-	LU-	-2,990.15	6.25%	0.00%
13	LU-	LU-	LN-	LN-	-2,982.85	4.08%	0.00%
14	LU-	LU-	LN-	LU-	-2,978.60	5.43%	9.54%
15	LU-	LU-	LU-	LN-	**-2,963.14**	7.07%	**36.19%**
16	LU-	LU-	LU-	LU-	-2,985.48	5.16%	2.33%
Model averaging					**-2,945.47**		

log-likelihood of -2,945.47, 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 model averaging process retained 9 out of the 16 models, and their weights are shown in the 'MA share' column in Table 1. We see that the model with the best individual log-likelihood obtains the largest share but we in addition see non-trivial shares for a substantial subset of other models. Crucially, this includes model 1, which had the worst sample level 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.

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

	Estimation Sample					Holdout Sample		
	Model averaging LL	Most contributing MMNLs Version	LL	Share	MA LL Improvement	Model averaging LL	MMNL LL	MA LL Improvement
Full	-2,945.47	15	-2,963.14	36.19%	17.67	n/a		
		10	-2,983.74	16.17%	38.27			
		11	-2,980.24	14.87%	34.77			
		14	-2,978.60	9.54%	33.13			
		6	-3,015.46	8.16%	69.99			
Holdout 1	-2,326.82	11	-2,355.35	20.42%	28.53	-624.81	-631.51	6.70
		10	-2,350.06	18.17%	23.24		-637.22	12.41
		13	-2,353.99	14.29%	27.17		-628.67	3.86
		2	-2,389.52	10.56%	62.70		-652.93	28.12
		14	-2,347.18	9.75%	20.36		-629.08	4.27
Holdout 2	-2,382.56	12	-2,405.45	24.68%	22.89	-558.21	-561.54	3.33
		9	-2,421.76	19.19%	39.20		-565.16	6.95
		16	-2,407.67	18.44%	25.11		-564.02	5.81
		6	-2,423.79	14.35%	41.23		-573.01	14.81
		3	-2,437.56	11.10%	55.01		-571.94	13.73
Holdout 3	-2,326.41	16	-2,355.67	17.44%	29.26	-621.63	-626.38	4.76
		8	-2,355.90	15.00%	29.49		-630.66	9.03
		15	-2,353.65	13.76%	27.25		-628.59	6.96
		13	-2,369.48	12.43%	43.07		-627.18	5.56
		1	-2,412.55	12.07%	86.14		-632.62	10.99
Holdout 4	-2,333.89	8	-2,362.21	24.18%	28.32	-614.88	-621.61	6.73
		9	-2,361.90	20.38%	28.01		-635.30	20.42
		3	-2,376.59	18.87%	42.70		-628.50	13.62
		15	-2,371.26	10.27%	37.37		-614.83	-0.05
		12	-2,370.15	8.52%	36.26		-628.56	13.68
Holdout 5	-2,346.93	15	-2,378.26	22.70%	31.33	-586.73	-595.39	8.66
		6	-2,388.22	22.50%	41.29		-596.86	10.13
		12	-2,396.10	10.86%	49.17		-601.15	14.43
		9	-2,391.96	8.49%	45.03		-592.52	5.79
		11	-2,381.30	7.20%	34.37		-596.67	9.94

We also test to see whether the results from model averaging are overfitting by using out-of-sample validation. To do this, we first split the dataset into five subsets of 80% of the data, where, for each subset, we first repeat the exact same process as described above for the full sample, i.e. estimating the parameters for all 16 mixed logit models and then estimating the weights for these models using a latent class structure. In the next step, we calculate the log-likelihood on the remaining 20% of the data, i.e. our hold-out sample, using the 16 separate models as well as the model averaging structure, each time with the parameters obtained from estimation on the 80%

sample. The results of this process are shown in Table 2, where, for brevity, we only ever show the fits for the five most contributing MMNLs in model averaging on the estimation data.

Across all five holdout runs, we see that model averaging obtains a better fit in estimation, where this is of course in line with expectation. 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, with a different "optimal" specification arising even across these datasets which share the majority of the sample. As in the full sample, we again see that models which do not fit well at the sample level can still contribute to the model average, with the best fitting model only twice receiving the largest share across the five subsets, and with 13 out of the 16 models appearing at least once in the top five contributors to the model average. Additionally, no single model is the largest contributor to model averaging in more than one holdout subset. Crucially, 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 the performance on the holdout sample is always superior for the model averaging model compared to the best fitting MMNL model on the estimation data. This highlights that model averaging is potentially more robust to overfitting than using a single model structure.

3.3. Computation of outputs from model averaging

In this section, we look at value of travel time as well as values for changing the amount of crowding and the rate of delays. We first use the estimates from each of the 16 MMNL models to obtain model-specific values[6] 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).

In our models, the individual coefficients follow random distributions, and as a result, so do the monetary valuations. We use the full distributions from the individual models in model averaging, i.e. we do not simply take the weighted average of moments of the distributions but produce an overall set of draws with an unequal distribution of draws from the individual distributions, representing the weights of each model. To explain this further, the distribution of the WTP in model averaging is represented by R draws, where in our case, we set $R = 10^6$. As model 1 has a weight of 7.2% in model averaging, $0.072 \cdot R$ draws will be produced from model 1 to contribute to the set of R draws. The means and standard deviations of the WTP measures for each model and the model average are given in Table 3. In comparison with the estimates obtained if we had simply used the best fitting mixed logit model (MMNL-15, highlighted in Table 3), results from model averaging suggest that the willingness to pay for changes in travel time and the rate of delays are lower by 3.9% and 8.5%, respectively. The opposite is true for changes in the number of crowded trips, for which model averaging produces a valuation that is 10.8% higher than that for MMNL-15. Model averaging also produces a much wider standard deviation for the value of crowding, by a factor of 62.6%.

[6] 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 3 : Welfare measures obtained from the UK models

Model	MA Share	Log-likelihood	VTT		VCR		VDE	
			mean	sd	mean	sd	mean	sd
1	7.6%	-3,034.16	3.3901	6.8120	0.2913	0.7249	0.2653	0.9255
2	0.0%	-3,030.67	3.3323	5.8661	0.3209	0.9212	0.2440	0.7226
3	0.0%	-3,019.60	3.0992	5.5377	0.3807	1.2754	0.2310	0.6749
4	0.0%	-3,015.35	3.3816	6.5317	0.3964	1.3673	0.2738	0.8196
5	0.0%	-3,027.83	3.1945	4.7553	0.3662	1.2518	0.2579	0.8680
6	8.2%	-3,015.46	3.1130	4.3986	0.4405	2.8656	0.2281	0.4376
7	0.0%	-3,001.06	3.6717	6.2209	0.3752	0.8128	0.2975	1.0523
8	3.3%	-2,996.96	2.9446	4.2261	0.3195	0.6953	0.1851	0.4678
9	1.9%	-2,982.40	3.8090	7.6200	0.2973	0.8028	0.2048	0.5058
10	16.2%	-2,983.74	3.9449	8.0836	0.3383	0.9637	0.2309	0.5933
11	14.9%	-2,980.24	3.8708	8.4069	0.4052	1.3764	0.2279	0.6246
12	0.0%	-2,990.15	3.9624	9.2506	0.3299	0.9689	0.2652	0.7720
13	0.0%	-2,982.85	3.6243	5.7299	0.3053	0.8823	0.2565	1.0757
14	9.5%	-2,978.60	3.6119	5.8889	0.3103	0.8760	0.1932	0.3483
15	**36.2%**	**-2,963.14**	**3.8910**	**6.6145**	**0.3025**	**0.6279**	**0.2596**	**0.7941**
16	2.3%	-2,985.48	3.6906	6.4057	0.3250	0.6588	0.1988	0.4331
Model Averaging		**-2,945.47**	**3.7405**	**7.0017**	**0.3352**	**1.0208**	**0.2374**	**0.6727**

4. APPLICATION TO RP MODE CHOICE DATA

We next test model averaging on revealed preference (RP) datasets, which can be more complex, both in terms of number of individuals and the size of the choice set. As a result, the models that can be applied often have to be simpler in structure as more complex models such as mixed logit quickly become computationally infeasible. Model averaging avoids these computational problems by creating a more complex model by 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), and this is the focus of our second application.

4.1. Data

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 considered (car driver (CD), car passenger (CP), taxi (TX), walk (WK), bicycle (BK), train (TR) or bus (BS)) and a large number of destination zones are defined (2,277 travel zones). For the purposes of this paper, we consider 5,173 home-work tours, where we focus on mode choice only. Level of service and attraction measures were assembled for each alternative such that attribute values could be derived for in-vehicle travel time, cost, access time, waiting times for public transport modes (time until next service and time until subsequent service) and distance. Details of the parameters used for the models for this dataset are given in Table A2, and readers are invited to refer to Fox (2015) for a full description of the data and its components, and also a discussion of the use of attraction measures.

4.2. Specification of individual components and model averaging

We group the attributes such that we have four parameter types: cost sensitivities (where there are three different income groups, β_{cost}, β_{cost_2} and β_{cost_3}), in-vehicle travel time (IVT) sensitivities (bus ($\beta_{\text{bus-time}}$), car ($\beta_{\text{car-time}}$), train ($\beta_{\text{train-time}}$), bus connection for train ($\beta_{\text{rail-bus-connect-time}}$)), other time (OT) sensitivities (access time ($\beta_{\text{access-time}}$), time until next service ($\beta_{\text{first-wait}}$), time until subsequent service ($\beta_{\text{second-wait}}$)) and distance sensitivities (car ($\beta_{\text{cp-dist}}$), walking ($\beta_{\text{wk-dist}}$) and bike ($\beta_{\text{bk-dist}}$) 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 (which is detailed in Table 4.11 of Fox 2015). The model used involves a complex utility specification, where full details of this are shown in Table A2. We then try linear and logarithmic specifications for the parameters by attribute type, using the four groups outlined above. This leads to 16 different combinations of linear and logarithmic transformations of attributes, and gives us the model results displayed in Table 4, with the model estimates given in Tables A3 and A4.

The best performing individual model (model 3) uses 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 66.6% of the allocation, where the improvement from model averaging over this model is 21 log-likelihood units. However, 4 other models are

TABLE 4 : Results from combinations of linear and logarithmic transformations of attributes on the Sydney HTS-06 mode choice data

Model	Attribute treatment				Overall	Best model for	MA
	Cost	IVT	OT	Distance	Log-likelihood	x% of respondents	Share
1	linear	linear	linear	linear	-2,784.74	5.18%	0.00%
2	linear	linear	linear	log	-2,803.43	5.48%	0.00%
3	linear	linear	log	linear	**-2,771.52**	6.50%	**66.60%**
4	linear	linear	log	log	-2,792.17	**9.95%**	0.00%
5	linear	log	linear	linear	-2,806.83	4.26%	0.00%
6	linear	log	linear	log	-2,814.47	4.68%	5.83%
7	linear	log	log	linear	-2,800.51	3.25%	0.00%
8	linear	log	log	log	-2,804.25	8.41%	1.34%
9	log	linear	linear	linear	-2,801.99	4.14%	0.00%
10	log	linear	linear	log	-2,799.90	1.57%	0.00%
11	log	linear	log	linear	-2,791.18	5.46%	0.00%
12	log	linear	log	log	-2,792.10	2.89%	6.69%
13	log	log	linear	linear	-2,839.87	6.06%	0.00%
14	log	log	linear	log	-2,823.12	6.44%	19.53%
15	log	log	log	linear	-2,838.38	5.60%	0.00%
16	log	log	log	log	-2,818.69	8.61%	0.00%
Model averaging					**-2,750.48**		

also included in the final set from model averaging. 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. This model is in fact the third worst fitting model at the sample level and the high weight in model averaging again shows how a model that works well for some people but badly overall can obtain a high weight in model averaging. 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. Notably, model 4, which is the best model for the largest percentage of respondents, and also performs relatively well at the sample level, does not get a share. This is likely a result of it being similar in structure but inferior in overall model fit to model 3.

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 5). 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 the 5 models retained by model averaging.

TABLE 5 : Model averaging log-likelihoods across the attribute treatment combinations for estimation and holdout samples for the Sydney HTS-06 mode choice data

	Estimation					Holdout Sample		
	Model averaging	Most contributing MMNLs			MA LL	Model averaging	MMNL	MA LL
	LL	Version	LL	Share	Improvement	LL	LL	Improvement
Full	-2,750.48	3	-2,771.52	66.60%	21.04	n/a		
		14	-2,823.12	19.53%	72.64			
		12	-2,792.10	6.69%	41.62			
		6	-2,814.47	5.58%	63.99			
		8	-2,804.25	1.34%	53.77			
Holdout 1	-2,216.40	3	-2,230.51	70.67%	14.11	-538.90	-544.80	5.90
		14	-2,278.31	19.68%	61.91		-549.82	10.91
		12	-2,249.97	6.23%	33.57		-545.95	7.05
		16	-2,272.48	2.18%	56.08		-550.71	11.80
		1	-2,243.81	1.25%	27.41		-545.27	6.36
Holdout 2	-2,145.19	3	-2,162.17	62.44%	16.98	-610.57	-614.94	4.37
		12	-2,174.79	15.54%	29.60		-623.56	12.99
		6	-2,198.73	10.74%	53.54		-623.58	13.01
		14	-2,212.04	6.04%	66.84		-617.86	7.29
		4	-2,172.02	3.62%	26.83		-626.51	15.94
Holdout 3	-2,198.76	3	-2,215.39	56.26%	16.64	-556.64	-561.49	4.85
		14	-2,257.75	12.84%	58.99		-569.38	12.74
		6	-2,252.54	12.51%	53.78		-566.65	10.01
		11	-2,228.74	7.92%	29.98		-567.16	10.52
		1	-2,221.77	4.12%	23.01		-569.47	12.83
Holdout 4	-2,212.34	3	-2,229.62	52.51%	17.29	-546.13	-548.73	2.60
		14	-2,270.23	13.40%	57.89		-563.16	17.04
		6	-2,250.48	12.19%	38.14		-556.84	10.71
		8	-2,246.49	6.82%	34.15		-549.18	3.05
		4	-2,239.66	3.98%	27.33		-551.99	5.86
Holdout 5	-2,208.77	3	-2,230.61	51.57%	21.84	-549.23	-553.30	4.08
		14	-2,255.60	27.24%	46.83		-561.25	12.02
		11	-2,236.46	9.48%	27.69		-558.46	9.23
		16	-2,250.30	7.83%	41.53		-555.52	6.29
		6	-2,261.23	3.88%	52.46		-559.30	10.07

4.3. Computation of outputs from model averaging

For the Sydney study, we can compare the value of time measures for different groups of individuals as we have three cost coefficients in each model for three different annual 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 6.

We see that, whilst the different models have fairly similar model fit, the values of travel time 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 differences between models are very significant, with the results from some models suggesting that individuals have valuations that are up to 10 times larger than the findings for other models. This makes the selection of one of the models a highly consequential decision. The results from model averaging, however, appear reasonable, with the added confidence that they combined results from a number of models that each offer comparable performance in explaining the choices in the data.

Elasticities are a key output from choice models, particularly for those estimated on RP data. We now look at the implications of model averaging in this context. Given that elasticities from different models can be very contrasting, a further use of model averaging is that it can be used to derive a single elasticity from an ‘average’ model. For the Sydney data, we calculate the elasticity for tours for all modes in response to a increase in the cost of car. 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). In particular, let $T_{j,base\,car\,cost}$ be the predicted number of tours by mode j at the base costs for car in the data, and let $T_{j,1.01\cdot car\,cost}$ be the predicted number of tours by mode j after a 1% increase in car costs. Both these quantities would be obtained by using Equation 8. The elasticity would then be calculated as :

$$E_{T_j,car\,cost} = \log(\frac{T_{j,1.01\cdot car\,cost}}{T_{j,base\,car\,cost}})/\log(1.01). \tag{13}$$

The tour elasticities in response to car costs are shown for the 16 different models tested in Table 7. 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. In line with the results of Fox et al. (2009), we see less of an impact on the share for car driver for models with a logarithmic transformation of costs (models 9-16). Consequently, more trips are transferred to train and bus under models 1-8, for which elasticity values are up to double those estimated by models 9-16. It is worth noting that as only mode choice is estimated here, the car elasticities observed are lower than those from models predicting mode and destination choice for this data (see Fox 2015). Whilst model averaging gives a larger share to linear cost models, lower elasticities for train and bus are found for the model average when compared with model 3, which is likely to have been used in the absence of model averaging (e.g. the cross-elasticity for train reduces from 0.2693 to 0.2371). As a result, it appears that the use of model averaging may allow a modeller to avoid the issue of finding elasticities that are either too high (through the use of linear attributes) or too low (through the use of logarithmic attributes).

TABLE 6 : Value of travel times (AUD/hr) obtained from the models for the Sydney choice-only data, across different modes and income categories (IC1, IC2 and IC3).

Model	MA Share	Log-likelihood	Car			Train			Bus			Access		
			IC1	IC2	IC3	IC1	IC2	IC3	IC1	IC2	IC3	IC1	IC2	IC3
1	0.00%	-2,784.74	9.1016	12.3341	15.3345	2.4234	3.2841	4.0830	7.1952	9.7506	12.1227	3.4965	4.7384	5.8910
2	0.00%	-2,803.43	7.7686	10.9756	13.9278	0.1209	0.1709	0.2168	6.1456	8.6827	11.0181	3.4003	4.8040	6.0961
3	**66.60%**	**-2,771.52**	**11.3727**	**15.2868**	**19.5670**	**4.1548**	**5.5848**	**7.1484**	**8.3142**	**11.1758**	**14.3049**	**7.0728**	**9.5071**	**12.1690**
4	0.00%	-2,792.17	9.6828	13.5127	17.6517	1.6983	2.3700	3.0960	7.0546	9.8450	12.8605	7.0814	9.8823	12.9093
5	0.00%	-2,806.83	1.0217	1.3098	1.5585	4.9862	6.3921	7.6058	0.2463	0.3157	0.3756	1.8624	2.3876	2.8409
6	5.83%	-2,814.47	4.3663	5.8800	7.1572	4.4614	6.0082	7.3132	2.1193	2.8541	3.4741	2.7683	3.7280	4.5378
7	0.00%	-2,800.51	2.5881	3.3150	3.9633	4.7048	6.0263	7.2047	0.6388	0.8182	0.9782	5.3473	6.8492	8.1887
8	1.34%	-2,804.25	6.8791	9.2047	11.3421	3.8407	5.1392	6.3325	2.9935	4.0055	4.9357	6.8599	9.1790	11.3105
9	0.00%	-2,801.99	20.8780	21.9172	21.8958	3.3445	3.5109	3.5075	13.6809	14.3618	14.3478	8.1491	8.5547	8.5464
10	0.00%	-2,799.90	11.4625	11.9060	12.0352	0.3142	0.3264	0.3299	7.9094	8.2154	8.3046	4.7990	4.9847	5.0388
11	0.00%	-2,791.18	27.1439	28.2111	28.3959	7.4605	7.7538	7.8046	17.2093	17.8859	18.0030	15.8066	16.4281	16.5357
12	6.69%	-2,792.10	14.3500	14.7996	15.0593	2.5177	2.5966	2.6421	9.5773	9.8773	10.0507	9.5335	9.8322	10.0048
13	0.00%	-2,839.87	3.7963	3.9377	3.9177	11.8119	12.2519	12.1898	-0.5665	-0.5876	-0.5847	4.6492	4.8224	4.7980
14	19.53%	-2,823.12	6.4854	6.6814	6.7138	6.7877	6.9928	7.0268	1.9804	2.0403	2.0502	4.0392	4.1612	4.1814
15	0.00%	-2,838.38	8.3800	8.5927	8.6020	12.3238	12.6367	12.6504	0.8833	0.9058	0.9068	12.6536	12.9749	12.9889
16	0.00%	-2,818.69	10.8960	11.1051	11.2367	6.3306	6.4521	6.5285	3.7380	3.8097	3.8549	10.1099	10.3039	10.4260
Model Averaging		**-2,750.48**	**10.1475**	**12.9421**	**15.9195**	**4.5727**	**5.6780**	**6.8212**	**6.7285**	**8.7224**	**10.8685**	**6.3905**	**8.1426**	**10.0066**

TABLE 7 : Elasticities for an increase in car cost for the Sydney mode choice models.

Model	MA Share	Log-likelihood	Tour car cost elasticity						
			Car Driver	Car Passenger	Train	Bus	Bike	Walk	Taxi
1	0.00%	-2,784.74	-0.1084	0.1259	0.2786	0.2339	0.1414	0.0522	0.1414
2	0.00%	-2,803.43	-0.0985	0.1147	0.2524	0.2093	0.1682	0.0505	0.1237
3	**66.60%**	**-2,771.52**	**-0.1033**	**0.1199**	**0.2693**	**0.2182**	**0.1326**	**0.0471**	**0.1342**
4	0.00%	-2,792.17	-0.0958	0.1090	0.2496	0.1993	0.1645	0.0474	0.1181
5	0.00%	-2,806.83	-0.1322	0.1455	0.3392	0.2941	0.1689	0.0624	0.1596
6	5.83%	-2,814.47	-0.1122	0.1291	0.2850	0.2465	0.1661	0.0562	0.1368
7	0.00%	-2,800.51	-0.1290	0.1405	0.3384	0.2776	0.1600	0.0575	0.1515
8	1.34%	-2,804.25	-0.1087	0.1207	0.2833	0.2321	0.1553	0.0520	0.1299
9	0.00%	-2,801.99	-0.0536	0.0929	0.0944	0.1118	0.1364	0.0992	0.1293
10	0.00%	-2,799.90	-0.0766	0.1297	0.1397	0.1627	0.1980	0.1284	0.1791
11	0.00%	-2,791.18	-0.0484	0.0846	0.0861	0.0985	0.1235	0.0898	0.1178
12	6.69%	-2,792.10	-0.0726	0.1236	0.1339	0.1510	0.1883	0.1212	0.1707
13	0.00%	-2,839.87	-0.0628	0.1050	0.1133	0.1316	0.1561	0.1132	0.1458
14	19.53%	-2,823.12	-0.0789	0.1307	0.1449	0.1691	0.1988	0.1314	0.1794
15	0.00%	-2,838.38	-0.0548	0.0923	0.1003	0.1111	0.1354	0.0995	0.1280
16	0.00%	-2,818.69	-0.0708	0.1184	0.1319	0.1476	0.1780	0.1180	0.1623
Model Averaging		**-2,750.48**	**-0.0971**	**0.1228**	**0.2371**	**0.2060**	**0.1515**	**0.0691**	**0.1456**

5. APPLICATION TO RP MODE AND DESTINATION CHOICE DATA

5.1. Data

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 (number of hospitals, employment and other services) for the different destinations and travel times, travel costs, and distances associated with the different travel modes.

5.2. Specification of individual components and model averaging

We use our California dataset to test model averaging across models looking jointly at mode and destination choice. We start by using a multinomial logit model (MNL). There are 4 different modes and 58 different destination zones. The utility for mode i and destination j is given by:

$$U_{ij} = \delta_i + \beta_F \cdot F_i + \beta_{TT_i} \cdot TT_i + log\left(exp(sz_h) \cdot h_j + exp(sz_e) \cdot e_j + exp(sz_o) \cdot o_j\right), \qquad (14)$$

where δ_i is an alternative specific constant,[7] F_i and TT_i are the travel fare and time for alternative i, respectively, β_F is the coefficient estimated for travel fare and β_{TT_i} is the mode-specific coefficient for travel time. The ‘size’ of destination j is then calculated using coefficients sz_h, sz_e and sz_o for the relative importance of the number of hospitals (h_j), employment (e_j) and other services (o_j). A full description of these models is given by Outwater et al. (2015). We next develop four nested logit models. One key decision for modellers considering destination choice is the definition of the destination boundaries, which can be arbitrary. In this case study, California could be split into, for example, 58 counties or 10 regions (See Figure 1).

The four nested logit models make use of the following nesting structures:

1. $NL_{destination}$. The use of 58 nests, one for each county, with 10 different nesting parameters, one for each region. Thus counties that are in the same region have the same nesting parameter, but the alternatives for different counties are not nested together.

2. NL_{region}. The use of 10 nests, one for each region, with a different nesting parameter for each region.

3. NL_{NSEW}. The use of 4 nests, with different nesting parameters, where the regions are grouped according to location: north (1 and 2), east (4 and 6), west (3, 5 and 8) and south (7, 9 and 10).

4. NL_{N-S}. This model also uses 4 nests, but this time with regions grouped depending on how far north they are: north (1 and 2), centre-north (3 and 4), centre-south (5 and 6) and south (7, 8, 9 and 10).

[7] For δ_{air}, a constant is estimated with additional shifts added depending on the travel distance.

FIGURE 1 : Californian regions and counties (retrieved from CA Government 2020)

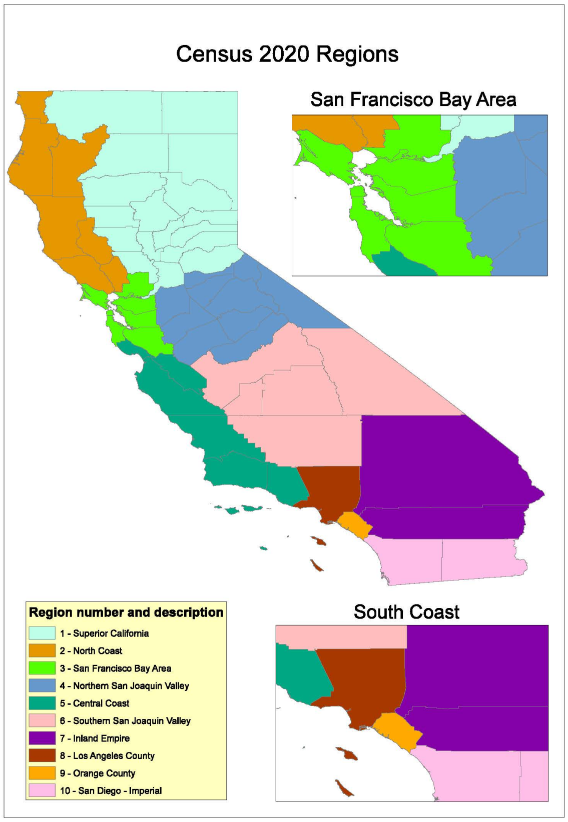

Table 8 shows that all four nested logit models give improvements in model fit over the MNL model, where Table A5 gives the estimated parameters for the five different models. Turning to model averaging, we see that three of the four NL models are retained, but the MNL model is not included in the final averaging. The worst fitting NL model, NL_{NSEW}, which provides the best fit for 10.98% of individuals, does not contribute to the model average. This is a result of the fact that the only nesting parameter that is significantly different from one is for the north regions. As this grouping is also in the NL_{N-S} model, the NL_{NSEW} becomes an intermediary model between MNL and NL_{N-S}. The shares for the three contributing models are approximately in line with the log-likelihoods of the models: NL_{region} is the best performing model and also receives the largest share.

5.3. Computation of outputs from model averaging

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. In this case, the NL$_{destination}$ model provides values that are higher than MNL, whereas the other nested logit models provide lower values than MNL. The result of using model averaging is that we obtain values that are closer to MNL, despite this model having been excluded from the averaging. Thus, by using model averaging, we again avoid more extreme values, with, for example, the best performing

TABLE 8 : The results from model averaging (MA) across five basic models applied to the California dataset

Model	pars	nests	free	fixed	Overall Log-likelihood	Gain over MNL	Best model for x% of respondents	MA Share
MNL	14	0	0	0	-23,955.26	0.00	1.15%	0.00%
NL$_{destination}$	21	10	7	3	-23,925.90	29.36	6.75%	28.22%
NL$_{region}$	17	10	3	7	**-23,921.72**	33.54	38.86%	**41.40%**
NL$_{NSEW}$	15	4	1	3	-23,939.90	15.36	10.98%	0.00%
NL$_{N-S}$	16	4	2	2	-23,926.94	28.32	**42.26%**	30.38%
Model averaging					**-23,904.58**	50.68		

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

Model	MA Share	Log-likelihood	VTT			
			car	bus	rail	air
MNL	0.00%	-23,955.26	14.2597	21.5192	18.5619	6.9961
NL$_{destination}$	28.22%	-23,925.90	17.0652	22.0728	22.5708	7.4018
NL$_{region}$	**41.40%**	**-23,921.72**	**13.5825**	**20.4702**	**18.2597**	**4.5117**
NL$_{NSEW}$	0.00%	-23,939.90	13.4850	20.7312	17.7946	6.2388
NL$_{N-S}$	30.38%	-23,926.94	13.4428	20.6955	17.3794	6.0621
MA		**-23,904.58**	**14.5229**	**20.9910**	**19.2089**	**5.7983**

model, NL_region, providing a particularly low estimate for air.

We also test different elasticities for the California dataset, where we estimate car cost and time elasticities for tours by mode, distance travelled by mode, and total distance travelled across modes. These elasticities are given in Tables 10 and 11.

TABLE 10 : Tour elasticities from the five different candidate models and model averaging for the California dataset

Model	MA Share	Log-likelihood	Car cost tour elasticity				Car time tour elasticity			
			car	bus	rail	air	car	bus	rail	air
MNL	0.00%	-23,955.26	-0.0184	0.4862	0.4710	0.4589	-0.0374	0.9796	0.9523	0.9369
NL$_{destination}$	28.22%	-23,925.90	-0.0208	0.5075	0.4833	0.5989	-0.0501	1.2180	1.1697	1.4221
NL$_{region}$	**41.40%**	**-23,921.72**	**-0.0197**	**0.5093**	**0.4993**	**0.4955**	**-0.0380**	**0.9749**	**0.9602**	**0.9582**
NL$_{NSEW}$	0.00%	-23,939.90	-0.0189	0.4991	0.4807	0.4714	-0.0362	0.9497	0.9189	0.9086
NL$_{N-S}$	30.38%	-23,926.94	-0.0189	0.4999	0.4782	0.4723	-0.0361	0.9483	0.9122	0.9070
Model averaging		**-23,904.58**	**-0.0198**	**0.5060**	**0.4885**	**0.5164**	**-0.0408**	**1.0367**	**1.0056**	**1.0686**

TABLE 11 : Distance elasticities from the five different candidate models and model averaging for the California dataset

Model	Car cost distance elasticity					Car time distance elasticity				
	car	bus	rail	air	total	car	bus	rail	air	total
MNL	-0.0194	0.0521	0.0506	0.0477	-0.0149	-0.0394	0.1047	0.1019	0.0978	-0.0302
NL$_{destination}$	-0.0180	0.0582	0.0534	0.0626	-0.0129	-0.0436	0.1378	0.1282	0.1492	-0.0313
NL$_{region}$	**-0.0202**	**0.0550**	**0.0539**	**0.0518**	**-0.0153**	**-0.0391**	**0.1049**	**0.1033**	**0.1004**	**-0.0297**
NL$_{NSEW}$	-0.0200	0.0534	0.0515	0.0491	-0.0154	-0.0384	0.1013	0.0982	0.0949	-0.0295
NL$_{N-S}$	-0.0200	0.0533	0.0511	0.0492	-0.0154	-0.0383	0.1009	0.0972	0.0948	-0.0294
Model averaging	**-0.0195**	**0.0554**	**0.0529**	**0.0539**	**-0.0147**	**-0.0401**	**0.1133**	**0.1083**	**0.1120**	**-0.0301**

Whilst the elasticities from most of the nested logits are similar to those of the MNL model, the NL$_{destination}$ model produces larger deviations. Calculating the elasticities from the model bringing together three different nesting structures provides suitable values which take into account the relative performance of the different models (as the NL$_{destination}$ model performs best), without giving undue influence to extreme results from any one model.

6. 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 out-of-sample 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. To deal with the issue of convergence to poor local optima, we present a simple expectation-maximisation (EM) algorithm which can deal with very large numbers of candidate models within the same model averaging structure, unlike is typically the case with classical estimation approaches for latent class.

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. 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 simpler models based on different utility specifications or nesting structures.

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 the most to a model average, or may in fact not contribute at all.

We show that model averaging always provides model fit at least as good as the best fitting candidate model. We have purposely 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. The use of the EM algorithm still remains possible, as discussed for discrete mixtures by Train (2009, chapter 14). 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, an analyst could ask multiple researchers or teams of researchers to develop models for the same dataset, and then estimate the model averaging across these models, for which they would only need the underlying log-likelihood contribution for each individual or observation in the dataset. This would allow an analyst to combine insights from different researchers or teams, with different skills/background, producing a more robust overall model.

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

Batram, M. and Bauer, D. (2017). Model selection and model averaging in MACML-estimated MNP models. *arXiv preprint arXiv:1704.00183*.

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*.

CA Government (2020). Census 2020 Regions. https://census.ca.gov/regions/.

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

Daly, A. (2010). Cost damping in travel demand models: Report of a study for the department for transport. Technical report, RAND Corporation.

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.

Guo, Z. and Wilson, N. H. (2007). Modeling effects of transit system transfers on travel behavior: case of commuter rail and subway in Downtown Boston, Massachusetts. *Transportation Research Record*, 2006(1):11–20.

Hess, S. (2010). Conditional parameter estimates from mixed logit models: distributional assumptions and a free software tool. *Journal of Choice Modelling*, 3(2):134–152.

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 Palma, D. (2019). Apollo: a flexible, powerful and customisable freeware package for choice model estimation and application, www.apollochoicemodelling.com.

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. In *Transportation Research Board 94th Annual Meeting*.

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.

Rose, J. M., Scarpa, R., and Bliemer, M. C. (2009). Incorporating model uncertainty into the generation of efficient stated choice experiments: A model averaging approach. *Institute of transport and logistics studies. Working paper*.

Sevcikova, H. and Raftery, A. (2013). mlogitbma: Bayesian model averaging for multinomial logit model. *R package version 0.1-6, URL http://CRAN. R-project. org/package= mlogitBMA*.

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.

Tang, T., Liu, R., and Choudhury, C. (2020). Incorporating weather conditions and travel history in estimating the alighting bus stops from smart card data. *Sustainable Cities and Society*, 53:101927.

Tjiong, L. (2015). Re-estimating UK appraisal values for non-work travel time savings using random coefficient logit model. *Transportation Research Procedia*, 8:50–61.

Train, K. (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press, Cambridge, MA, second edition edition.

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.

Wan, A. T., Zhang, X., and Wang, S. (2014). Frequentist model averaging for multinomial and ordered logit models. *International Journal of Forecasting*, 30(1):118–128.

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.

Zannat, K. E. and Choudhury, C. F. (2019). Emerging big data sources for public transport planning: A systematic review on current state of art and future research directions. *Journal of the Indian Institute of Science*, pages 1–19.

Zhao, S., Zhou, J., and Yang, G. (2019). Averaging estimators for discrete choice by m-fold cross-validation. *Economics Letters*, 174:65–69.

APPENDIX: CANDIDATE MODEL OUTPUTS

UK data

Table A1 gives the parameter estimates and robust t-ratios for the 16 different mixed logit models for the UK dataset.

We estimate lognormal or loguniform distributions for each β-coefficient to capture inter-individual heterogeneity only. Thus, for example, the coefficient for travel fare under a lognormal distribution is estimated with:

$$\beta_{n,LF} = -exp(\mu_{log(-\beta_{LF})} + \sigma_{log(-\beta_{LF})}) \cdot \xi_{n,LF}, \tag{A1}$$

where $\mu_{log(-\beta_{LF})}$ is the estimated mean for the log of $-\beta_{LF}$, $\sigma_{log(-\beta_{LF})}$ is the corresponding standard deviation and $\xi_{n,LF}$ is an inter-individual level standard normally distributed error term.

Equivalently, the coefficient for travel fare under a loguniform distribution is estimated with:

$$\beta_{n,LF} = -exp(a_{log(-\beta_{LF})} + b_{log(-\beta_{LF})}) \cdot \xi_{n,LF}, \tag{A2}$$

where $a_{log(-\beta_{LF})}$ and $b_{log(-\beta_{LF})}$ are the estimated offset and range, respectively, for the uniform distribution of the log of $-\beta_{LF}$ and $\xi_{n,LF}$ is now an inter-individual level standard uniform distributed error term. To see how to simply estimate these models, please refer to Hess and Palma (2019). In Table A1, par_1 refers to the $\mu_{log(-\beta)}$ for negative lognormal distributions and $a_{log(-\beta)}$ for negative loguniform distributions, with par_2 referring to $\sigma_{log(-\beta)}$ and $b_{log(-\beta)}$, respectively. For positive coefficients, the minus sign is dropped.

Sydney data

For the Sydney dataset, there are a large number of 'level of service' attributes as well as socio-demographic variables that are used in the specification for the utility of the different modes. The full specification for each mode are given in Table A2.

A number of indicator variables (I) are used to indicate that the parameter is only used in certain cases. The attributes, which are labelled X, (with the indices for individual and choice task dropped) can all either be entered into the utility as they are (in which case $I_{NC} = 1$ for cost attributes, $I_{NT} = 1$ for in-vehicle time attributes, $I_{NR} = 1$ for other time attributes and I_{ND} for distance attributes) or with a logarithmic transformation applied (corresponding to indicators I_{LC}, I_{LT}, I_{LR} and I_{LD}). The use of either natural or logarithmic transformations of attributes results in our 16 different model specifications. A number of socio-demographic parameters (ξ) are also included along with additional alternative specific constants that are only applied in some cases (such as I_{CBD}, which corresponds to whether the destination is in the CBD). This results in, for example, the utility for walking being calculated as:

$$\begin{aligned} U_{WK} = \delta_{WK} & \\ & + \beta_{\text{wk-dist}} \cdot I_{ND} \cdot X_{\text{slow-dist}} \\ & + \beta_{\text{log-walk-dist}} \cdot I_{LD} \cdot log(X_{\text{slow-dist}}) \\ & + \delta_{\text{WK-CBD}} \cdot I_{CBD}, \end{aligned} \tag{A3}$$

where δ_{WK} is an alternative specific constant for the utility of walking and $\delta_{\text{WK-CBD}}$ is an additional constant added in the case where $I_{CBD} = 1$, which is when the destination is in the CBD. $\beta_{\text{wk-dist}}$ is the marginal utility for walking distance, $X_{\text{slow-dist}}$, which is used only in models which utilise natural distances ($I_{ND} = 1$ and $I_{LD} = 0$). Alternatively, models using a logarithmic transform of walking distance would have $I_{ND} = 0$ and $I_{LD} = 1$, resulting in the use of $\beta_{\text{log-walk-dist}}$ instead. A full description of the attributes and socio-demographics is given in Section 4.3 of Fox (2015).

California data

For the California dataset, the full parameter estimates for the MNL and four nested logit models is given in Table A5.

TABLE A1 : Parameter estimates for the mixed logit models for the UK data

Model number		Model 1		Model 2		Model 3		Model 4		Model 5		Model 6		Model 7		Model 8	
Log-likelihood		-3,034.16		-3,030.67		-3,019.60		-3,015.35		-3,027.83		-3,015.46		-3,001.06		-2,996.96	
		est.	rob.t-rat.	est.	rob.t-rat.	est.	rob.t-rat.	est.	rob.t-rat.	est.	rob.t-rat.	est.	rob.t-rat.	est.	rob.t-rat.	est.	rob.t-rat.
TT	dist.	LN-		LN-		LN-		LN-		LN-		LN-		LN-		LN-	
	par_1	-2.83	-23.51	-2.82	-19.67	-2.81	-22.94	-2.78	-20.83	-4.75	-11.50	-4.52	-12.63	-4.64	-11.15	-4.72	-3.53
	par_2	-0.71	-8.09	-0.59	-4.43	-0.60	-6.77	-0.58	-5.28	3.60	6.22	3.34	6.51	3.55	6.77	3.73	2.20
LF	dist.	LN-		LN-		LN-		LN-		LU-		LU-		LU-		LU-	
	par_1	1.94	21.18	1.87	14.96	1.96	21.44	1.97	21.33	1.88	17.86	1.93	18.18	2.01	22.17	2.09	23.09
	par_2	1.04	14.20	1.02	9.12	1.03	14.35	1.09	12.48	0.99	11.40	1.04	12.32	1.06	13.69	1.10	13.26
CR	dist.	LN-		LN-		LU-		LU-		LN-		LN-		LU-		LU-	
	par_1	-1.38	-8.63	-1.50	-7.43	-6.97	-22.55	-6.33	-24.33	-1.51	-6.92	-1.38	-7.57	-6.56	-22.21	-4.61	-7.76
	par_2	-0.94	-8.63	1.11	9.93	8.48	24.44	7.72	31.70	1.08	10.25	1.09	10.97	8.09	24.29	5.61	7.43
RD	dist.	LN-		LU-		LN-		LU-		LN-		LU-		LN-		LU-	
	par_1	-1.80	-4.56	-6.11	-8.97	-1.75	-5.29	-5.26	-5.48	-1.80	-3.32	-4.66	-3.24	-1.65	-2.84	-4.87	-5.40
	par_2	1.25	4.44	6.88	8.22	1.11	4.65	6.04	5.65	1.00	2.02	5.14	2.77	1.03	1.53	5.66	5.55
AD	dist.	LN-		LN-		LN-		LN-		LN-		LN-		LN-		LN-	
	par_1	-3.40	-9.20	-3.79	-2.00	-3.58	-6.62	-4.89	-2.94	-3.69	-5.47	-3.91	-6.98	-3.66	-3.70	-3.88	-6.35
	par_2	0.91	3.09	1.06	0.40	1.13	2.43	1.98	2.76	0.99	1.92	1.28	2.50	1.22	1.22	1.57	5.03
ED	dist.	LN-		LN-		LN-		LN-		LN-		LN-		LN-		LN-	
	par_1	-4.53	-4.95	-2.80	-5.35	-3.50	-5.39	-2.32	-7.08	-2.94	-4.07	-2.74	-6.86	-3.39	-4.93	-2.47	-7.78
	par_2	-2.12	-5.71	-1.43	-6.86	-1.64	-5.81	0.20	0.38	-1.48	-6.18	-1.46	-10.20	-1.63	-5.11	-0.33	-2.16
CI	dist.	LN-		LN-		LN-		LN-		LN-		LN-		LN-		LN-	
	par_1	-2.57	-2.35	-2.46	-2.38	-3.27	-1.05	-7.35	-0.39	-2.21	-3.10	-2.75	-2.67	-2.84	-1.66	-1.79	-3.46
	par_2	-1.48	-2.71	-1.38	-4.10	-2.08	-1.39	-4.42	-0.46	-1.31	-5.93	-1.49	-4.54	-1.93	-2.09	-0.77	-3.56
FI	dist.	LN+		LN+		LN+		LN+		LN+		LN+		LN+		LN+	
	par_1	-0.87	-4.10	-0.78	-3.75	-0.81	-3.52	-0.85	-2.34	-1.12	-2.06	-0.76	-3.29	-0.77	-4.16	-0.76	-3.75
	par_2	0.29	1.01	0.19	0.42	0.26	0.72	0.49	1.31	-0.79	-1.54	0.35	0.83	0.10	0.20	0.28	0.75
	δ_1	0.66	8.63	0.69	8.42	0.68	8.41	0.73	8.39	0.69	8.61	0.72	8.98	0.72	8.80	0.76	9.26
	δ_2	0.25	3.91	0.27	4.04	0.28	4.21	0.29	4.29	0.27	3.98	0.28	4.14	0.29	4.21	0.30	4.23
Model number		Model 9		Model 10		Model 11		Model 12		Model 13		Model 14		Model 15		Model 16	
Log-likelihood		-2,982.40		-2,983.74		-2,980.24		-2,990.15		-2,982.85		-2,978.60		-2,963.14		-2,985.48	
		est.	rob.t-rat.	est.	rob.t-rat.	est.	rob.t-rat.	est.	rob.t-rat.	est.	rob.t-rat.	est.	rob.t-rat.	est.	rob.t-rat.	est.	rob.t-rat.
TT	dist.	LU-		LU-		LU-		LU-		LU-		LU-		LU-		LU-	
	par_1	-2.81	-22.70	-2.77	-22.36	-2.77	-23.42	-2.79	-23.67	-4.52	-11.73	-4.68	-10.27	-4.73	-9.43	-4.97	-9.79
	par_2	-0.70	-5.60	-0.68	-5.80	-0.82	-7.47	-0.67	-8.83	3.40	6.04	3.68	6.07	3.79	6.12	4.03	5.71
LF	dist.	LN-		LN-		LN-		LN-		LU-		LU-		LU-		LU-	
	par_1	0.12	1.06	0.21	1.06	0.11	0.44	0.22	0.28	0.20	1.05	0.27	1.53	0.20	1.14	0.22	1.16
	par_2	3.76	31.95	3.58	14.52	3.77	11.55	3.70	2.93	3.67	14.49	3.59	15.19	3.75	15.99	3.66	16.24
CR	dist.	LN-		LN-		LU-		LU-		LN-		LN-		LU-		LU-	
	par_1	-1.54	-7.36	-1.80	-6.06	1.13	4.20	-4.78	-4.78	-1.54	-7.85	-1.46	-8.55	-4.42	-6.15	-4.28	-6.00
	par_2	1.35	10.48	1.67	7.73	-5.73	-5.63	5.87	4.99	1.25	10.27	1.24	16.18	5.35	5.96	5.25	6.13
RD	dist.	LN-		LU-		LN-		LU-		LN-		LU-		LN-		LU-	
	par_1	-1.91	-3.61	-4.35	-4.79	-1.82	-5.99	-6.82	-33.42	-2.09	-5.32	-4.06	-4.25	-1.74	-6.27	-5.35	-3.34
	par_2	1.36	5.38	4.89	4.68	1.40	9.68	7.62	30.34	1.52	7.29	4.38	4.03	1.29	7.47	5.96	3.29
AD	dist.	LN-		LN-		LN-		LN-		LN-		LN-		LN-		LN-	
	par_1	-3.64	-4.43	-4.73	-9.40	-4.78	-6.59	-7.01	-4.92	-4.03	-4.25	-3.92	-7.26	-4.05	-5.30	-4.54	-3.63
	par_2	1.24	1.88	-2.74	-10.97	2.20	5.77	2.90	5.32	1.54	3.85	1.54	5.93	1.62	3.52	1.96	3.35
ED	dist.	LN-		LN-		LN-		LN-		LN-		LN-		LN-		LN-	
	par_1	-3.50	-2.78	-2.74	-6.90	-2.40	-6.30	-2.10	-9.13	-2.59	-4.11	-3.62	-5.89	-3.58	-5.88	-2.45	-4.42
	par_2	-1.77	-3.72	1.25	7.84	0.59	3.13	0.99	8.75	1.18	4.17	2.54	7.57	-1.91	-7.24	1.25	5.37
CI	dist.	LN-		LN-		LN-		LN-		LN-		LN-		LN-		LN-	
	par_1	-476.82	-3.36	-4.90	-0.73	-16.90	-2.06	-30.50	-0.64	-3.96	-1.16	-7.43	-1.74	-5.75	-1.97	-240.20	-0.88
	par_2	-235.40	-3.40	-3.10	-0.85	-9.22	-2.14	-17.90	-0.65	-2.66	-1.46	-4.35	-2.17	-4.09	-2.37	-92.21	-0.88
FI	dist.	LN+		LN+		LN+		LN+		LN+		LN+		LN+		LN+	
	par_1	-0.80	-3.11	-0.80	-2.50	-0.72	-3.10	-0.89	-3.23	-0.95	-2.11	-0.89	-3.26	-1.10	-3.41	-0.83	-3.99
	par_2	-0.43	-1.19	0.37	0.57	-0.12	-0.09	-0.62	-1.82	-0.62	-1.10	-0.65	-2.49	-0.79	-3.50	-0.69	-4.00
	δ_1	0.74	9.81	0.76	9.47	0.77	9.94	0.77	9.87	0.74	8.94	0.78	9.90	0.79	9.59	0.82	10.44
	δ_2	0.28	4.19	0.31	4.50	0.31	4.40	0.32	4.45	0.29	4.03	0.29	3.97	0.31	4.27	0.31	4.18

TABLE A2 : Parameters included in the specifications for the utilities for the different modes in the Sydney data

Parameter	Car Driver (CD)	Car Passenger (CP)	Train (TR)	Bus (BS)	Bicycle (BK)	Walk (WK)	Taxi (TX)
δ_{CD}	1						
δ_{CP}		1					
δ_{TR}			1				
δ_{BS}				1			
δ_{BK}					1		
δ_{WK}						1	
δ_{TX}							1
β_{cost_1}	$I_{NC} \cdot I_{IG_1} \cdot CD_{fact} \cdot X_{\text{car-cost}}$	$I_{NC} \cdot I_{IG_1} \cdot CP_{fact} \cdot X_{\text{car-cost}}$	$I_{NC} \cdot I_{IG_1} \cdot X_{\text{train-fare}}$	$I_{NC} \cdot I_{IG_1} \cdot X_{\text{bus-fare}}$			$I_{NC} \cdot I_{IG_1} \cdot X_{\text{taxi-fare}}$
β_{cost_2}	$I_{NC} \cdot I_{IG_2} \cdot CD_{fact} \cdot X_{\text{car-cost}}$	$I_{NC} \cdot I_{IG_2} \cdot CP_{fact} \cdot X_{\text{car-cost}}$	$I_{NC} \cdot I_{IG_2} \cdot X_{\text{train-fare}}$	$I_{NC} \cdot I_{IG_2} \cdot X_{\text{bus-fare}}$			$I_{NC} \cdot I_{IG_2} \cdot X_{\text{taxi-fare}}$
β_{cost_3}	$I_{NC} \cdot I_{IG_3} \cdot CD_{fact} \cdot X_{\text{car-cost}}$	$I_{NC} \cdot I_{IG_3} \cdot CP_{fact} \cdot X_{\text{car-cost}}$	$I_{NC} \cdot I_{IG_3} \cdot X_{\text{train-fare}}$	$I_{NC} \cdot I_{IG_3} \cdot X_{\text{bus-fare}}$			$I_{NC} \cdot I_{IG_3} \cdot X_{\text{taxi-fare}}$
$\beta_{\text{car-time}}$	$I_{NT} \cdot X_{\text{car-time}}$	$I_{NT} \cdot X_{\text{car-time}}$					$I_{NT} \cdot X_{\text{car-time}}$
$\beta_{\text{rail-time}}$			$I_{NT} \cdot X_{\text{train-time}}$				
$\beta_{\text{bus-time}}$				$I_{NT} \cdot X_{\text{bus-time}}$			
$\beta_{\text{rail-bus-connect-time}}$			$I_{NT} \cdot X_{\text{railbus-time}}$				
$\beta_{\text{access-time}}$			$I_{NR} \cdot X_{\text{rail-walk}}$	$I_{NR} \cdot X_{\text{bus-walk}}$			
$\beta_{\text{first-wait}}$			$I_{NR} \cdot X_{\text{first-wait}}$	$I_{NR} \cdot X_{\text{first-wait}}$			
$\beta_{\text{second-wait}}$			$I_{NR} \cdot X_{\text{second-wait}}$	$I_{NR} \cdot X_{\text{second-wait}}$			
$\beta_{\text{wk-dist}}$						$I_{ND} \cdot X_{\text{slow-dist}}$	
$\beta_{\text{bk-dist}}$					$I_{ND} \cdot X_{\text{slow-dist}}$		
$\beta_{\text{cp-dist}}$		$I_{ND} \cdot X_{\text{car-dist}}$					$I_{ND} \cdot X_{\text{car-dist}}$
$\beta_{\text{log-cost}_1}$	$I_{LC} \cdot I_{IG_1} \cdot CD_{fact} \cdot log(X_{\text{car-cost}})$	$I_{LC} \cdot I_{IG_1} \cdot CP_{fact} \cdot log(X_{\text{car-cost}})$	$I_{LC} \cdot I_{IG_1} \cdot log(X_{\text{train-fare}})$	$I_{LC} \cdot I_{IG_1} \cdot log(X_{\text{bus-fare}})$			$I_{LC} \cdot I_{IG_1} \cdot log(X_{\text{taxi-fare}})$
$\beta_{\text{log-cost}_2}$	$I_{LC} \cdot I_{IG_2} \cdot CD_{fact} \cdot log(X_{\text{car-cost}})$	$I_{LC} \cdot I_{IG_2} \cdot CP_{fact} \cdot log(X_{\text{car-cost}})$	$I_{LC} \cdot I_{IG_2} \cdot log(X_{\text{train-fare}})$	$I_{LC} \cdot I_{IG_2} \cdot log(X_{\text{bus-fare}})$			$I_{LC} \cdot I_{IG_2} \cdot log(X_{\text{taxi-fare}})$
$\beta_{\text{log-cost}_3}$	$I_{LC} \cdot I_{IG_3} \cdot CD_{fact} \cdot log(X_{\text{car-cost}})$	$I_{LC} \cdot I_{IG_3} \cdot CP_{fact} \cdot log(X_{\text{car-cost}})$	$I_{LC} \cdot I_{IG_3} \cdot log(X_{\text{train-fare}})$	$I_{LC} \cdot I_{IG_3} \cdot log(X_{\text{bus-fare}})$			$I_{LC} \cdot I_{IG_3} \cdot log(X_{\text{taxi-fare}})$
$\beta_{\text{log-car-time}}$	$I_{LT} \cdot log(X_{\text{car-time}})$	$I_{LT} \cdot log(X_{\text{car-time}})$					$I_{LT} \cdot log(X_{\text{car-time}})$
$\beta_{\text{log-rail-time}}$			$I_{LT} \cdot log(X_{\text{train-time}})$				
$\beta_{\text{log-bus-time}}$				$I_{LT} \cdot log(X_{\text{bus-time}})$			
$\beta_{\text{log-rail-bus-connect-time}}$			$I_{LT} \cdot log(X_{\text{railbus-time}})$				
$\beta_{\text{log-access-time}}$			$I_{LR} \cdot log(X_{\text{rail-walk}})$	$I_{LR} \cdot log(X_{\text{bus-walk}})$			
$\beta_{\text{log-first-wait}}$			$I_{LR} \cdot log(X_{\text{first-wait}})$	$I_{LR} \cdot log(X_{\text{first-wait}})$			
$\beta_{\text{log-second-wait}}$			$I_{LR} \cdot log(X_{\text{second-wait}})$	$I_{LR} \cdot log(X_{\text{second-wait}})$			
$\beta_{\text{log-walk-dist}}$						$I_{LD} \cdot log(X_{\text{slow-dist}})$	
$\beta_{\text{log-bike-dist}}$					$I_{LD} \cdot log(X_{\text{slow-dist}})$		
$\beta_{\text{log-carP-dist}}$		$I_{LD} \cdot log(X_{\text{car-dist}})$					$I_{LD} \cdot log(X_{\text{car-dist}})$
$\zeta_{\text{other-car-user}}$	$I_{\text{other-car-user}}$						
$\zeta_{\text{company-cars}}$	$I_{\text{company-cars}}$						
$\zeta_{\text{male-driver}}$	I_{male}						
$\zeta_{under25}$	$I_{under25}$						
$\zeta_{\text{other-driver}}$		$I_{\text{other-driver}}$					
$\zeta_{\text{higher-income-rail}}$			I_{IG_3}				
$\zeta_{\text{ft-worker}}$			$I_{\text{ft-worker}}$				
$\zeta_{\text{male-bike}}$					I_{male}		
δ_{iz}	I_{IZ}						
δ_{bef830}	I_{bef830}						
$\delta_{\text{CD-CBD}}$	I_{CBD}						
$\delta_{\text{CP-CBD}}$		I_{CBD}					
$\delta_{\text{TR-CBD}}$			I_{CBD}				
$\delta_{\text{BS-CBD}}$				I_{CBD}			
$\delta_{\text{BK-CBD}}$					I_{CBD}		
$\delta_{\text{WK-CBD}}$						I_{CBD}	
$\delta_{\text{TX-CBD}}$							I_{CBD}

TABLE A3 : Parameter estimates for models 1-8 for the Sydney data

Model number	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6		Model 7		Model 8	
Log-likelihood	-2,784.74		-2,803.43		-2,771.52		-2,792.17		-2,806.83		-2,814.47		-2,800.51		-2,804.25	
Parameter	est.	rob.t-rat.	est.	rob.t-rat.	est.	rob.t-rat.	est.	rob.t-rat.	est.	rob.t-rat.	est.	rob.t-rat.	est.	rob.t-rat.	est.	rob.t-rat.
δ_{CD}	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA
δ_{CP}	-5.0727	-18.62	-4.1064	-12.97	-5.0950	-18.40	-4.0844	-12.78	-5.1631	-18.85	-4.2520	-13.39	-5.1978	-18.74	-4.2868	-13.34
δ_{TR}	-1.7096	-5.34	-1.8236	-5.77	1.7714	3.46	1.6205	3.19	-3.8044	-7.68	-4.6632	-9.26	-0.5514	-0.96	-1.1924	-2.06
δ_{BS}	-2.3391	-9.36	-2.4616	-9.97	0.8942	1.98	0.7351	1.64	-2.8671	-8.65	-3.2886	-9.64	0.1470	0.33	-0.0091	-0.02
δ_{BK}	-6.5355	-10.55	-4.5545	-7.42	-6.4993	-10.59	-4.3984	-7.24	-6.8587	-9.64	-5.6821	-8.24	-7.2621	-10.03	-5.9427	-8.63
δ_{WK}	-0.6619	-2.81	0.6106	2.44	-0.6068	-2.57	0.7155	2.86	-0.8958	-2.89	-0.1819	-0.52	-1.1147	-3.71	-0.4257	-1.26
δ_{TX}	-4.6062	-12.72	-3.8605	-9.50	-4.6684	-12.91	-3.8641	-9.60	-4.4546	-13.11	-3.8750	-9.93	-4.5284	-13.44	-3.9461	-10.25
β_{cost_1}	-0.0015	-5.56	-0.0014	-5.58	-0.0014	-5.53	-0.0014	-5.64	-0.0017	-6.82	-0.0015	-6.34	-0.0016	-6.86	-0.0014	-6.39
β_{cost_2}	-0.0011	-2.72	-0.0010	-2.69	-0.0011	-2.70	-0.0010	-2.73	-0.0013	-3.45	-0.0011	-3.12	-0.0013	-3.46	-0.0011	-3.12
β_{cost_3}	-0.0009	-3.74	-0.0008	-3.56	-0.0008	-3.71	-0.0007	-3.59	-0.0011	-4.96	-0.0009	-4.34	-0.0011	-5.05	-0.0009	-4.41
$\beta_{\text{car-time}}$	-0.0226	-4.84	-0.0182	-4.57	-0.0270	-6.81	-0.0221	-6.40	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA
$\beta_{\text{rail-time}}$	-0.0060	-1.38	-0.0003	-0.08	-0.0099	-2.44	-0.0039	-1.14	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA
$\beta_{\text{bus-time}}$	-0.0178	-6.38	-0.0144	-5.90	-0.0197	-7.53	-0.0161	-6.99	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA
$\beta_{\text{rail-bus-connect-time}}$	-0.0179	-4.60	-0.0148	-4.19	-0.0227	-6.24	-0.0194	-5.83	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA
$\beta_{\text{access-time}}$	-0.0087	-1.23	-0.0080	-1.18	0.0000	NA	0.0000	NA	-0.0052	-0.89	-0.0069	-1.05	0.0000	NA	0.0000	NA
$\beta_{\text{first-wait}}$	-0.0234	-3.57	-0.0222	-3.44	0.0000	NA	0.0000	NA	-0.0203	-3.26	-0.0204	-3.22	0.0000	NA	0.0000	NA
$\beta_{\text{second-wait}}$	-0.0387	-6.01	-0.0416	-6.55	0.0000	NA	0.0000	NA	-0.0531	-8.26	-0.0547	-8.40	0.0000	NA	0.0000	NA
$\beta_{\text{wk-dist}}$	-0.5494	-12.48	0.0000	NA	-0.5656	-12.83	0.0000	NA	-0.5369	-8.96	0.0000	NA	-0.5799	-9.37	0.0000	NA
$\beta_{\text{bk-dist}}$	-0.1363	-5.24	0.0000	NA	-0.1451	-5.47	0.0000	NA	-0.1059	-3.75	0.0000	NA	-0.1210	-3.87	0.0000	NA
$\beta_{\text{cp-dist}}$	-0.0232	-4.24	0.0000	NA	-0.0227	-4.12	0.0000	NA	-0.0243	-4.73	0.0000	NA	-0.0233	-4.57	0.0000	NA
$\beta_{\text{log-cost}_1}$	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA
$\beta_{\text{log-cost}_2}$	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA
$\beta_{\text{log-cost}_3}$	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA
$\beta_{\text{log-car-time}}$	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	-0.1382	-0.77	-0.5253	-2.52	-0.3367	-1.95	-0.7929	-4.21
$\beta_{\text{log-rail-time}}$	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	-0.6746	-5.36	-0.5367	-4.08	-0.6121	-5.00	-0.4427	-3.55
$\beta_{\text{log-bus-time}}$	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	-0.0333	-0.24	-0.2550	-1.73	-0.0831	-0.59	-0.3451	-2.41
$\beta_{\text{log-rail-bus-connect-time}}$	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	-0.0298	-0.61	-0.0711	-1.38	-0.0701	-1.43	-0.1183	-2.37
$\beta_{\text{log-access-time}}$	0.0000	NA	0.0000	NA	-0.8140	-6.47	-0.7831	-6.29	0.0000	NA	0.0000	NA	-0.6957	-5.51	-0.7907	-6.15
$\beta_{\text{log-first-wait}}$	0.0000	NA	0.0000	NA	-0.3321	-4.49	-0.3228	-4.40	0.0000	NA	0.0000	NA	-0.3479	-4.82	-0.3645	-4.97
$\beta_{\text{log-second-wait}}$	0.0000	NA	0.0000	NA	-0.4470	-5.03	-0.4812	-5.54	0.0000	NA	0.0000	NA	-0.7199	-7.99	-0.7537	-8.33
$\beta_{\text{log-walk-dist}}$	0.0000	NA	-2.8088	-20.88	0.0000	NA	-2.9000	-21.81	0.0000	NA	-2.9890	-15.00	0.0000	NA	-3.2345	-17.09
$\beta_{\text{log-bike-dist}}$	0.0000	NA	-1.6460	-10.38	0.0000	NA	-1.7641	-11.68	0.0000	NA	-1.6301	-7.66	0.0000	NA	-1.8652	-9.08
$\beta_{\text{log-carP-dist}}$	0.0000	NA	-0.5522	-5.94	0.0000	NA	-0.5646	-6.09	0.0000	NA	-0.5184	-5.59	0.0000	NA	-0.5105	-5.57
$\zeta_{\text{other-car-user}}$	-1.5998	-16.57	-1.6225	-16.25	-1.5981	-16.47	-1.6190	-16.11	-1.5982	-16.52	-1.6230	-16.33	-1.5957	-16.42	-1.6216	-16.18
$\zeta_{\text{company-cars}}$	0.7789	6.26	0.8226	6.26	0.7831	6.30	0.8280	6.29	0.7660	6.12	0.8045	6.15	0.7590	6.10	0.7984	6.12
$\zeta_{\text{male-driver}}$	0.2223	2.36	0.2411	2.47	0.2199	2.33	0.2391	2.44	0.2506	2.66	0.2650	2.74	0.2569	2.73	0.2708	2.80
$\zeta_{under25}$	-0.4097	-2.85	-0.4205	-2.84	-0.4061	-2.85	-0.4198	-2.86	-0.4380	-3.03	-0.4396	-2.98	-0.4347	-3.04	-0.4403	-3.01
$\zeta_{\text{other-driver}}$	1.8271	7.16	1.8402	7.27	1.8420	7.07	1.8479	7.22	1.8497	7.18	1.8620	7.31	1.8687	7.14	1.8830	7.29
$\zeta_{\text{higher-income-rail}}$	-0.0664	-0.48	-0.0417	-0.31	-0.0563	-0.41	-0.0335	-0.25	-0.0836	-0.62	-0.0723	-0.53	-0.0728	-0.54	-0.0626	-0.47
$\zeta_{\text{ft-worker}}$	-0.2170	-1.52	-0.2073	-1.47	-0.2117	-1.48	-0.2005	-1.42	-0.2155	-1.53	-0.2187	-1.56	-0.2199	-1.57	-0.2264	-1.61
$\zeta_{\text{male-bike}}$	2.2415	4.03	2.1794	4.03	2.2064	4.06	2.1749	4.02	2.2247	4.08	2.2092	4.08	2.2258	4.07	2.2072	4.07
δ_{iz}	-1.4286	-4.26	0.1355	0.26	-1.3876	-4.12	0.2385	0.44	-1.5558	-4.58	-0.1615	-0.32	-1.6222	-4.84	-0.1489	-0.29
δ_{bef830}	-0.4531	-4.11	-0.5105	-4.42	-0.4684	-4.22	-0.5311	-4.55	-0.4916	-4.42	-0.4900	-4.25	-0.5014	-4.48	-0.5008	-4.30
$\delta_{\text{CD-CBD}}$	-2.4075	-4.90	-2.5303	-5.13	-2.5094	-5.11	-2.6297	-5.34	-2.3490	-4.72	-2.4597	-4.94	-2.4903	-5.01	-2.5918	-5.22
$\delta_{\text{CP-CBD}}$	-2.1677	-3.97	-2.1278	-3.93	-2.2116	-4.04	-2.1828	-4.03	-2.2353	-4.09	-2.1480	-3.97	-2.3205	-4.25	-2.2225	-4.11
$\delta_{\text{TR-CBD}}$	-0.3645	-0.69	-0.3458	-0.66	-0.5537	-1.06	-0.5480	-1.06	-0.3396	-0.64	-0.3206	-0.61	-0.5927	-1.14	-0.5879	-1.13
$\delta_{\text{BS-CBD}}$	-0.4886	-0.96	-0.4567	-0.90	-0.6904	-1.36	-0.6796	-1.35	-0.5652	-1.11	-0.5658	-1.12	-0.8678	-1.71	-0.8986	-1.78
$\delta_{\text{BK-CBD}}$	-0.3052	-0.56	-0.4338	-0.82	-0.2030	-0.37	-0.3900	-0.73	-0.4203	-0.79	-0.6059	-1.16	-0.4157	-0.78	-0.6295	-1.20
$\delta_{\text{WK-CBD}}$	-1.5208	-2.07	-1.3833	-1.89	-1.5207	-2.04	-1.4084	-1.92	-1.6040	-2.19	-1.5115	-2.09	-1.6606	-2.26	-1.5717	-2.17
$\delta_{\text{TX-CBD}}$	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA

TABLE A4 : Parameter estimates for models 9-16 for the Sydney data

Model number	Model 9		Model 10		Model 11		Model 12		Model 13		Model 14		Model 15		Model 16	
Log-likelihood	-2,801.99		-2,799.90		-2,791.18		-2,792.10		-2,839.87		-2,823.12		-2,838.38		-2,818.69	
Parameter	est	rob.t-rat.	est	rob.t-rat.	est	rob.t-rat.	est	rob.t-rat.	est	rob.t-rat.	est	rob.t-rat.	est	rob.t-rat.	est	rob.t-rat.
δ_{CD}	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA
δ_{CP}	-6.3811	-16.61	-5.8252	-14.65	-6.2590	-16.44	-5.7172	-14.53	-6.6764	-17.45	-6.0505	-15.60	-6.5039	-17.29	-5.9128	-15.51
δ_{TR}	-1.5759	-4.85	-1.5598	-4.92	2.2201	4.43	1.9845	3.98	-4.4540	-9.21	-4.8599	-9.79	-0.8991	-1.54	-1.2894	-2.19
δ_{BS}	-2.1179	-8.29	-2.1288	-8.43	1.4183	3.27	1.1724	2.72	-3.1994	-8.83	-3.4025	-8.72	0.2082	0.45	0.0645	0.14
δ_{BK}	-8.1555	-11.48	-6.6286	-9.38	-7.9735	-11.18	-6.3556	-9.17	-9.0417	-11.71	-7.9736	-10.86	-9.2894	-11.68	-8.0438	-10.95
δ_{WK}	-2.0670	-5.61	-1.3104	-3.36	-1.8665	-5.11	-1.1029	-2.86	-2.7248	-7.20	-2.2540	-5.89	-2.7720	-7.38	-2.3263	-6.21
δ_{TX}	-5.3919	-13.36	-3.5681	-6.95	-5.5148	-13.57	-3.6413	-7.14	-5.2925	-13.31	-3.7494	-7.49	-5.4969	-13.81	-3.9902	-8.10
β_{cost_1}	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA
β_{cost_2}	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA
β_{cost_3}	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA
$\beta_{\text{car-time}}$	-0.0269	-6.38	-0.0218	-5.94	-0.0316	-8.70	-0.0259	-8.11	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA
$\beta_{\text{rail-time}}$	-0.0043	-1.04	-0.0006	-0.16	-0.0087	-2.26	-0.0045	-1.34	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA
$\beta_{\text{bus-time}}$	-0.0176	-6.55	-0.0150	-6.15	-0.0200	-7.99	-0.0173	-7.56	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA
$\beta_{\text{rail-bus-connect-time}}$	-0.0190	-5.07	-0.0161	-4.70	-0.0244	-6.91	-0.0213	-6.49	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA
$\beta_{\text{access-time}}$	-0.0105	-1.46	-0.0091	-1.35	0.0000	NA	0.0000	NA	-0.0068	-1.10	-0.0077	-1.20	0.0000	NA	0.0000	NA
$\beta_{\text{first-wait}}$	-0.0218	-3.36	-0.0182	-2.90	0.0000	NA	0.0000	NA	-0.0160	-2.53	-0.0152	-2.42	0.0000	NA	0.0000	NA
$\beta_{\text{second-wait}}$	-0.0471	-8.00	-0.0455	-7.72	0.0000	NA	0.0000	NA	-0.0628	-10.34	-0.0610	-9.88	0.0000	NA	0.0000	NA
$\beta_{\text{wk-dist}}$	-0.5981	-11.91	0.0000	NA	-0.6104	-12.24	0.0000	NA	-0.6033	-8.94	0.0000	NA	-0.6502	-9.37	0.0000	NA
$\beta_{\text{bk-dist}}$	-0.1467	-5.07	0.0000	NA	-0.1556	-5.28	0.0000	NA	-0.1170	-3.56	0.0000	NA	-0.1341	-3.68	0.0000	NA
$\beta_{\text{cp-dist}}$	-0.0209	-4.06	0.0000	NA	-0.0202	-3.84	0.0000	NA	-0.0210	-4.35	0.0000	NA	-0.0191	-3.99	0.0000	NA
$\beta_{\text{log-cost}_1}$	-0.4237	-5.39	-0.6259	-6.50	-0.3826	-4.93	-0.5945	-6.38	-0.4823	-6.25	-0.6272	-6.88	-0.4186	-5.59	-0.5619	-6.54
$\beta_{\text{log-cost}_2}$	-0.4036	-4.95	-0.6026	-6.16	-0.3681	-4.53	-0.5764	-6.05	-0.4649	-5.81	-0.6089	-6.55	-0.4082	-5.21	-0.5514	-6.23
$\beta_{\text{log-cost}_3}$	-0.4040	-5.28	-0.5961	-6.38	-0.3657	-4.82	-0.5665	-6.23	-0.4673	-6.23	-0.6059	-6.85	-0.4078	-5.59	-0.5449	-6.51
$\beta_{\text{log-car-time}}$	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	-0.2701	-1.48	-0.6001	-3.01	-0.5175	-2.98	-0.9033	-4.98
$\beta_{\text{log-rail-time}}$	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	-0.8403	-6.27	-0.6281	-4.62	-0.7610	-5.91	-0.5248	-4.12
$\beta_{\text{log-bus-time}}$	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0403	0.26	-0.1833	-1.14	-0.0545	-0.36	-0.3099	-2.03
$\beta_{\text{log-rail-bus-connect-time}}$	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	-0.0296	-0.60	-0.0650	-1.28	-0.0755	-1.55	-0.1171	-2.35
$\beta_{\text{log-access-time}}$	0.0000	NA	0.0000	NA	-0.8921	-7.20	-0.8361	-6.75	0.0000	NA	0.0000	NA	-0.7814	-6.30	-0.8381	-6.62
$\beta_{\text{log-first-wait}}$	0.0000	NA	0.0000	NA	-0.3267	-4.38	-0.2736	-3.67	0.0000	NA	0.0000	NA	-0.3300	-4.44	-0.3216	-4.26
$\beta_{\text{log-second-wait}}$	0.0000	NA	0.0000	NA	-0.5437	-6.65	-0.5179	-6.28	0.0000	NA	0.0000	NA	-0.8662	-10.30	-0.8470	-9.81
$\beta_{\text{log-walk-dist}}$	0.0000	NA	-3.1922	-20.53	0.0000	NA	-3.2674	-21.37	0.0000	NA	-3.3813	-16.11	0.0000	NA	-3.6174	-18.00
$\beta_{\text{log-bike-dist}}$	0.0000	NA	-1.9353	-11.21	0.0000	NA	-2.0451	-12.38	0.0000	NA	-1.8927	-8.15	0.0000	NA	-2.1283	-9.45
$\beta_{\text{log-carP-dist}}$	0.0000	NA	-0.6999	-7.49	0.0000	NA	-0.7037	-7.50	0.0000	NA	-0.6351	-6.90	0.0000	NA	-0.6048	-6.62
$\zeta_{\text{other-car-user}}$	-1.5890	-16.52	-1.5928	-15.96	-1.5887	-16.47	-1.5935	-15.89	-1.5789	-16.52	-1.5852	-16.08	-1.5783	-16.45	-1.5876	-16.00
$\zeta_{\text{company-cars}}$	0.7606	6.21	0.7890	6.12	0.7656	6.24	0.7944	6.14	0.7461	6.08	0.7716	6.03	0.7404	6.05	0.7668	6.00
$\zeta_{\text{male-driver}}$	0.2392	2.55	0.2533	2.61	0.2371	2.53	0.2510	2.58	0.2659	2.86	0.2777	2.90	0.2708	2.92	0.2827	2.95
$\zeta_{under25}$	-0.4756	-3.44	-0.4981	-3.52	-0.4712	-3.43	-0.4932	-3.51	-0.5294	-3.83	-0.5354	-3.80	-0.5222	-3.82	-0.5292	-3.79
$\zeta_{\text{other-driver}}$	1.8353	7.14	1.8866	7.38	1.8442	7.09	1.8959	7.34	1.8549	7.17	1.9128	7.41	1.8768	7.16	1.9367	7.41
$\zeta_{\text{higher-income-rail}}$	-0.0875	-0.63	-0.0990	-0.71	-0.0798	-0.58	-0.0923	-0.67	-0.0726	-0.53	-0.0988	-0.72	-0.0573	-0.43	-0.0856	-0.63
$\zeta_{\text{ft-worker}}$	-0.1770	-1.26	-0.1907	-1.37	-0.1673	-1.19	-0.1815	-1.30	-0.1631	-1.18	-0.1895	-1.37	-0.1619	-1.18	-0.1923	-1.39
$\zeta_{\text{male-bike}}$	2.2002	4.05	2.1748	4.02	2.1956	4.04	2.1671	4.01	2.2180	4.07	2.2091	4.07	2.2148	4.04	2.2013	4.05
δ_{iz}	-1.8387	-5.39	-0.2070	-0.40	-1.7479	-5.11	-0.0796	-0.15	-2.1186	-6.28	-0.5715	-1.19	-2.1292	-6.37	-0.5179	-1.07
δ_{bef830}	-0.4310	-3.93	-0.5091	-4.39	-0.4428	-4.01	-0.5239	-4.48	-0.4833	-4.38	-0.5035	-4.37	-0.4889	-4.40	-0.5078	-4.37
$\delta_{\text{CD-CBD}}$	-2.6573	-5.31	-2.5222	-5.00	-2.7589	-5.52	-2.6274	-5.22	-2.6244	-5.24	-2.5088	-4.97	-2.7740	-5.56	-2.6693	-5.31
$\delta_{\text{CP-CBD}}$	-2.1655	-4.02	-2.0966	-3.89	-2.2087	-4.10	-2.1411	-3.97	-2.1721	-4.03	-2.1084	-3.90	-2.2313	-4.14	-2.1736	-4.02
$\delta_{\text{TR-CBD}}$	-0.1856	-0.37	-0.1716	-0.34	-0.3713	-0.75	-0.3585	-0.72	-0.0469	-0.09	-0.0861	-0.17	-0.2899	-0.58	-0.3504	-0.70
$\delta_{\text{BS-CBD}}$	-0.4228	-0.85	-0.4390	-0.89	-0.6192	-1.25	-0.6346	-1.29	-0.3990	-0.80	-0.4964	-1.00	-0.6871	-1.39	-0.8126	-1.64
$\delta_{\text{BK-CBD}}$	-0.3595	-0.62	-0.5261	-0.97	-0.2287	-0.39	-0.4543	-0.83	-0.4798	-0.82	-0.7177	-1.31	-0.4327	-0.73	-0.7119	-1.29
$\delta_{\text{WK-CBD}}$	-1.5243	-2.08	-1.4053	-1.93	-1.4983	-2.04	-1.4024	-1.92	-1.5396	-2.10	-1.5340	-2.10	-1.5511	-2.11	-1.5661	-2.15
$\delta_{\text{TX-CBD}}$	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA

TABLE A5 : Parameter estimates for all models for the California data

Model	MNL		NL		NL		NL		NL	
Version	basic		destination		region		NSEW		N-S	
Log-likelihood	-23,955.26		-23,925.90		-23,921.72		-23,939.90		-23,926.94	
Free parameters	14		21		17		15		16	
Parameters	est.	rob. t-rat(0)	est.	rob. t-rat(0)	est.	rob. t-rat(0)	est.	rob. t-rat(0)	est.	rob. t-rat(0)
δ_{car}	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA
δ_{bus}	-4.0868	-18.41	-3.4724	-15.18	-3.9682	-18.91	-4.0880	-18.40	-4.0871	-18.22
δ_{rail}	-3.5272	-8.92	-2.7558	-7.34	-3.2675	-9.18	-3.5220	-8.93	-3.5419	-8.90
δ_{air}	-4.8046	-8.23	-4.1379	-7.73	-4.5718	-7.95	-4.7415	-8.12	-4.7222	-8.09
$\beta_{time_{car}}$	-0.0036	-14.69	-0.0037	-15.91	-0.0034	-14.59	-0.0035	-14.03	-0.0035	-14.00
$\beta_{time_{bus}}$	-0.0054	-8.01	-0.0048	-9.51	-0.0052	-8.24	-0.0053	-7.91	-0.0053	-7.82
$\beta_{time_{rail}}$	-0.0047	-3.12	-0.0049	-3.61	-0.0046	-3.47	-0.0046	-3.07	-0.0045	-2.98
$\beta_{time_{air}}$	-0.0018	-1.20	-0.1620	-1.21	-0.0011	-0.80	-0.0016	-1.11	-0.0016	-1.08
β_{cost}	-0.0150	-7.58	-0.0131	-6.76	-0.0152	-7.86	-0.0154	-7.70	-0.0155	-7.69
dist$_{air_{500}}$	0.7668	0.65	0.6556	0.68	0.5817	0.51	0.8231	0.70	0.8165	0.70
dist$_{air_{600}}$	0.9117	0.85	0.4853	0.55	0.8873	0.85	0.8468	0.79	0.8324	0.78
dist$_{air_{700}}$	0.4543	1.11	0.2702	0.77	0.3985	1.00	0.4270	1.05	0.4922	1.22
dist$_{air_{800}}$	0.0884	0.36	-0.0333	-0.15	0.1342	0.56	0.1513	0.62	0.0986	0.41
sz$_h$	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA
sz$_o$	0.9258	8.53	0.9038	8.35	0.6525	6.23	1.0100	9.24	0.9727	9.15
sz$_e$	-1.3666	-13.86	-1.3971	-14.19	-1.5275	-17.28	-1.3459	-13.46	-1.3687	-14.29
Nesting pars.	est.	rob. t-rat(1)	est.	rob. t-rat(1)	est.	rob. t-rat(1)	est.	rob. t-rat(1)	est.	rob. t-rat(1)
λ_{reg_1}			0.6689	-6.82	0.9499	-3.76				
λ_{reg_2}			0.6954	-3.70	1.0000	NA				
λ_{reg_3}			0.8610	-2.23	1.0000	NA				
λ_{reg_4}			0.6886	-5.79	0.8730	-6.75				
λ_{reg_5}			0.8147	-2.99	1.0000	NA				
λ_{reg_6}			1.0000	NA	1.0000	NA				
λ_{reg_7}			1.0000	NA	1.0000	NA				
λ_{reg_8}			0.8758	-1.65	1.0000	NA				
λ_{reg_9}			1.0000	NA	1.0000	NA				
$\lambda_{reg_{10}}$			0.9412	-0.95	0.8885	-2.79				
λ_{north}							0.8227	-5.77	0.7912	-7.22
λ_{east}							1.0000	NA		
λ_{west}							1.0000	NA		
λ_{south}							1.0000	NA	1.0000	NA
$\lambda_{centre-north}$									0.8018	-4.46
$\lambda_{centre-south}$									1.0000	NA