Article:

Dumont, J, Giergiczny, M and Hess, S (2015) Individual level models vs. sample level models: contrasts and mutual benefits. Transportmetrica A: Transport Science. ISSN 2324-9935

https://doi.org/10.1080/23249935.2015.1018681

Individual level models vs. sample level models: contrasts and mutual benefits

Jeffrey Dumont* Marek Giergiczny† Stephane Hess‡

(authors in alphabetical order)

Abstract

With a view to better capturing heterogeneity across decision makers and improving prediction of choices, there is increasing interest in estimating separate models for each person. Almost exclusively, this work has however taken place outside the field of transport research. The aim of the present paper is twofold. We first wish to give an account of the potential benefits of a greater focus on individual level estimates in transport applications. Secondly, we wish to offer further insights into the relative benefits of sample level and individual level models (ILM) by drawing on a dataset containing an unusually large number (144) of decisions on holiday travel per individual. In addition to comparing existing approaches, we also put forward the use of a novel technique which draws on the relative benefits of both sample level and individual level models by estimating ILMs in a Bayesian fashion with priors drawn from a sample level model. Our results show only limited differences between ILMs and conditionals from sample level models when working with the full set of choices. When working with more realistic sample sizes at the person level, our results suggest that ILMs can offer better performance on the estimation data but that this is a result of overfitting which can lead to inferior prediction performance. Our proposed Bayesian ILM model offers good intermediary performance. The use of best-worst data rather than simple stated choice, as is done commonly in published ILM work, does not lead to major changes to these findings.

Keywords: individual level models; hierarchical bayes; heterogeneity; prediction

1 Introduction

Researchers have long believed that individuals differ in preferences, and substantial effort has gone into deterministic and random treatment of these differences in choice models. Indeed, and focussing on a transport context, a growing majority of academic applications in choice modelling now allow for some level of random heterogeneity across a sample of individual

*Corresponding Author, RSG, 55 Railroad Row, White River Junction, VT 05001, USA, Ph: 1-802-295-4999, Fax: 1-802-295-4999, E: jeff.dumont@rsginc.com

†University of Warsaw, Faculty of Economic Sciences, 44/50 Dluga Street, 00-241 Warsaw, Poland, Ph: 48-22-8313201, Fax: 48-22-8312846, E: mgiergiczny@wne.uw.edu.pl

‡Institute for Transport Studies, University of Leeds, University Road, Leeds LS2 9JT, UK, Ph: 44-0113-34-36611, Fax: 44-0113-343-5334, E: S.Hess@its.leeds.ac.uk

travellers. This can take the form of continuous Mixed Logit models (e.g. Brownstone et al., 2000) or latent class structures (e.g Ishaq et al., 2014), or a mixture of the two (e.g. Greene and Hensher, 2013). A growing number of studies also allow for the influence of unobserved (i.e. random) factors such as attitudes, perceptions and convictions (e.g. Daziano and Bolduc, 2013).

While the key emphasis in capturing heterogeneity has been on sample level models (i.e. one model for the entire sample) accommodating heterogeneity across individual respondents (see Train, 2009, for detailed overviews), there is a long established stream of work aimed at using separate models for each respondent. One of the first efforts to model individuals can be found in Beggs et al. (1981). The authors of this study estimated individual level models (ILM) using exploded logit models on ranked data. Based on statistical tests they found that allowing for individual level parameters leads to a significant improvement over the sample level models.

The ILMs in the Beggs et al. (1981) study had 10 parameters and were estimated using 15 observations per respondent. As a result, almost fifty percent of the models did not converge. A very similar pattern regarding convergence rate was found in another application of ILM at that time (cf. Chapman, 1984). ILMs require a large number of observations per person, and ranking data was seen as a convenient approach at the time of the Beggs et al. (1981) and Chapman (1984) work. Around the same time however, Hausman and Ruud (1987) and Ben-Akiva et al. (1991) found statistical differences in preferences across stages in ranking and determined that even after controlling for scale differences across stages, the hypothesis of a joint model was rejected. These findings more than likely played a key role in reducing researchers' interest in ILMs in subsequent years, and between the work of Chapman (1984) and Louviere et al. (2008), i.e. a period of almost twenty-five years, only limited work was conducted on modeling individual respondents' choices.

Interest in ILMs started afresh in 2008 when Louviere et al. (2008) proposed a new way of collecting a large number of observations per person by using efficient designs and a best-worst (BW) elicitation format. Louviere et al. argued that the problems with ranking identified by Hausman and Ruud (1987) and Ben-Akiva et al. (1991) could partly be due to the design used. These earlier studies relied on ranking choice data from one or a small number of choice tasks[1]. In contrast with these earlier applications, the study by Louviere et al. (2008) relied on multiple choice tasks composed of no more than five alternatives, which were prepared using optimal-in-difference designs (cf. Street et al., 2005; Burgess et al., 2006). Respondents in this study were asked a sequence of BW questions in alternating order - the argument being that BW is better than traditional ranking as it places less cognitive burden on respondents. Studies relying on BW tasks are also believed to be superior to the standard ranking approach as they take advantage of a person's propensity to respond more consistently to extreme options (cf. Flynn et al., 2007; Marley, 2009), by moving the focus away from middle ranked alternatives. The field then witnessed strong popularity of BW data in estimating ILMs after 2008 (e.g. Lancsar and Louviere, 2008; Ebling et al., 2010; Marshall et al., 2011; Giergiczny and Czajkowski, 2011). Nevertheless, it is important to also acknowledge growing doubts

[1] As an example, the study by Beggs et al. (1981) required respondents to rank 16 cars in just a single choice situation.

about the suitability of BW data, given evidence of differential drivers of preference for best and worst choices (cf. Giergiczny et al., 2013; Rose, 2014).

Apart from the design and the elicitation format, Louviere et al. (2008) proposed a weighted least squares (WLS) regression approach to fit models to individuals' choices. It is known that this estimation approach produces consistent, but inefficient, estimates (Louviere and Woodworth, 1983), and it was used as the first-step estimator in the iteratively reweighted least squares approach to obtaining maximum likelihood estimates (e.g. Green, 1984). An advantage of WLS compared to traditional ML estimators is that each individual level discrete choice model will converge successfully, even if not all attributes were used in making the choices. As a result, the majority of applications aimed at estimating individual level models in recent years have relied on WLS or ordinary least squares (OLS) (e.g. Louviere et al., 2008; Ebling et al., 2010; Marshall et al., 2011; Frischknecht et al., 2011b). OLS and WLS are recommended as *"simple, easy-to-use methods that are highly likely to produce correct answers"* (Frischknecht et al., 2011b). In the same line, Marshall et al. (2011) describe and discuss a comparison of the WLS approach with standard hierarchical Bayes (HB) estimation of a Mixed Logit model (MMNL), and report that both predict equally well. Frischknecht et al. (2011a) proposed a further method to estimate ILMs, in the form of a variant of maximum likelihood estimation known as the modified maximum likelihood (MML) method.

What is notable is that the work on individual level models has taken place almost completely outside the transport field. While insights into individual preferences may be of particular interest in a marketing context, where much research has been done, the same is true for transport, be it in terms of understanding the impact of policy decisions on subsets of the population, or in determining appropriate market segments (and pricing approaches) for new services and products (e.g. new cars). One obvious example comes in understanding the impact on distinct population segments of controversial policies such as road pricing (see e.g. Levinson, 2010). Furthermore, having access to estimates at an individual level can be a great asset in making more informed decisions on the specification for a sample level model, or in conducting cluster analysis, as discussed for example by Train (2009, chapter 11). This paper thus partly seeks to bring this literature to the wider attention of transport researchers.

Secondly, the paper has a methodological focus. While the above work has focused on the estimation of models for a single individual, there is also a parallel body of work on calculating respondent specific posterior distributions from the estimates of sample level models (see e.g. Train, 2009, chapter 11), with free software available (Hess, 2010). Individual parameter posterior distributions are also a natural output from HB models and their use is quite common practice in building simulators and for general forecasting, particularly in marketing (see e.g. Rossi et al. 2005 or Train 2009, chapter 12). The present paper first seeks to offer further insights into the relative benefits of sample level and individual level models in terms of explaining existing choices, understanding patterns of heterogeneity, and predicting future choices. In addition to comparing existing approaches, we also put forward the use of a novel technique which draws on the relative benefits of both sample level and individual level models by estimating ILMs in a Bayesian fashion with priors drawn from a sample level model. We conduct these comparisons with the help of a particularly suitable dataset which contains 144 choices per respondent.

The remainder of this paper is organized as follows. Section 2 presents the different estimation approaches applicable in this context, including our proposed hybrid approach. This is followed in Section 3 by a discussion of our data collection effort, and in Section 4 by the presentation of our empirical work. Finally, Section 5 summarises our findings and presents the conclusions of the research.

2 Estimation techniques

This section gives a brief overview of the main estimation techniques that are applicable and which are used in the present paper, including our proposed hybrid approach. It should also be noted that not every technique is applicable to each of the estimation runs, as discussed in Section 4.

2.1 Maximum likelihood estimation of ILMs

The most simplistic approach to producing individual level estimates is to perform model estimation at the level of individual respondents, i.e. using maximum likelihood estimation of ILMs, leading to an individual vector of parameters β_n for each respondent. As is well known, this is however likely to lead to major issues in estimation. Albert and Anderson (1984) identify cases in which maximum likelihood (ML) estimation will fail to converge. They classify data sets into three categories: complete data separation, quasi-complete data separation, and data overlap. Complete separation occurs when a combination of explanatory variables classifies responses without error according to a strict inequality. Quasi-complete separation occurs when a combination of explanatory variables classifies responses without error up to a non-strict inequality. In all other cases, data have overlap, which means no separation issues. ML estimation will fail to converge in cases of complete and quasi-complete separation[2] and it will converge only if data have overlap, i.e. there is no single rule that correctly predicts the actual choice in each observation. A simple example of data separation would be one where all choices for an individual can be explained on the basis of a single attribute, such as where a person always chooses the cheapest option, or always chooses the fastest option.

2.2 Conditionals from sample level MMNL models

An alternative to the estimation of ILMs is to estimate a sample level Mixed Multinomial Logit (MMNL) model and to produce conditional parameter estimates. For full details, see Train (2009, chapter 11).

In a MMNL model, we allow the vector β to follow a random distribution with parameters Ω. Making an assumption of intra-respondent homogeneity, following the work of Revelt and

[2] If maximum likelihood estimates do not exist for some coefficients because of quasi-complete separation, they may still exist for other variables in a choice model. In fact, if one leaves the offending variables in the model, then the coefficients, standard errors and test statistics for the remaining variables are still valid maximum likelihood estimates (Altman et al. (2003)).

Train (1998), we then write the likelihood of the observed sequence of choices for decision maker n as:

$$L_n(\Omega) = \int_\beta L_n(\beta) f(\beta \mid \Omega) \, \mathrm{d}\beta, \tag{1}$$

where $L_n(\beta) = \prod_{t=1}^{T_n} P_{ni^*t}(\beta)$, with $P_{ni^*t}(\beta)$ giving the probability of the observed choice for respondent n in task t, conditional on β.

Following estimation of the sample level distribution of β, further insights can then be obtained in a Bayesian manner, by calculating information relating to a given individual's sensitivities on the basis of the sample level model estimates and that individual's observed choices. In a continuous mixed logit context, these calculations are straightforward, as discussed for example by Train (2009, chapter 11).

Using Bayes' rule, we get:

$$L(\beta_n \mid C_n) = \frac{L_n(\beta) f(\beta \mid \Omega)}{L_n(\Omega)} \tag{2}$$

This gives us the probability of given values for β_n, conditional on the observed choices (C_n) for individual n, where it is important to remember that β_n is not observed but is distributed. It is then straightforward to calculate a conditional mean for β_n as:

$$\bar{\beta_n} = \int_{\beta_n} \beta_n L(\beta_n \mid C_n) \, \mathrm{d}\beta_n, \tag{3}$$

with similar calculations to obtain the corresponding variance or other measures.

2.3 Hierarchical Bayes Models

For the HB models estimated in this study, we assumed the model structure as described in Train (2009, chapter 12) and as implemented in the software package RSGHB (Dumont et al., 2012). To describe HB estimation, we first need to step back and discuss some basics of Bayesian inference. From Bayes theorem, we know that the posterior distribution K of a set of model parameters θ given a set of choices Y is

$$K(\theta|Y) \propto L(Y|\theta)k(\theta), \tag{4}$$

where $k(\theta)$ is often described as the prior. The goal of Bayesian estimation is to sample from this posterior distribution K to describe the model parameters.

In the context of a mixed logit model, the posterior distribution of b and W is defined by

$$K(\Omega|y) \propto \prod_n L(y_n|\Omega)k(\Omega), \tag{5}$$

with

$$k(\Omega) = k(b)k(W), \tag{6}$$

where b is the mean, W the covariance, and $k(b)$ is assumed to be normal with mean b_0 and extremely large variance s_o while $k(W)$ is assumed to be inverted Wishart, $IW(v, S)$. As with classical estimation, the likelihood function does not have a closed form and so must be approximated using simulation. This is typically done using a Metropolis Hastings algorithm. HB estimation requires iterative sampling from three posterior distributions $K(b), K(W), K(\beta)$ which is referred to as Gibbs Sampling.

With HB models, there is a need to specify values for b_0, v and S. In particular, the assumptions for v and S impact the prior sample variance assumed in estimation. This prior variance can have a large impact on the individual level posterior estimates, β (see e.g. Gelman, 2006). While some may see this as a disadvantage to Bayesian estimation, it does allow the analyst to control the amount of Bayesian shrinkage that occurs with the individual level posteriors. A higher prior variance results in more weight being placed on each individual's choice behavior, which reduces the amount of shrinkage that occurs and results in more dissimilar (across people) posterior estimates. A smaller prior variance results in more weight being placed on sample level behavior, increasing the amount of shrinkage and giving posteriors that are more similar across individuals. In Section 4, different prior variances and degrees of freedom were tested with the HB models to understand their impact on the results.

2.4 Modified Maximum Likelihood

As mentioned in Section 2.1, it is known that maximum likelihood estimates only exist in the case of data overlap and that models will otherwise fail to converge (Beggs et al., 1981; Albert and Anderson, 1984). The problem of separation is particularly likely when the number of observations is small, which is typically the case for ILMs. The modified maximum likelihood (MML) method proposed by Frischknecht et al. (2011a) offers a way to overcome these data separation challenges.

The approach relies on the principle of shrinkage which is accomplished by estimating parameters based on maximum penalized likelihood. The penalty function adopted in the MML method corresponds to a Bayesian approach designed to overcome the challenge of finite samples, which is to augment the limited data with prior beliefs about the behaviour of the data (see e.g. Geweke, 2005). As an example, with a completely flat prior, the method involves adding in equally likely choices of all alternatives for all choice sets, which is essentially random noise but guarantees convergence. An important issue is the weight attached to these artificial choices - the best weight can be determined through a hold out sample[3]. Frischknecht et al. (2011b) showed through Monte Carlo simulations that even with data sets as small as the minimum degrees of freedom, MML can improve out-of-sample performance and have superior parameter recovery and out of sample prediction compared to conventional maximum likelihood.

2.5 Models using Ordinary Least Squares/Weighted Least Squares

Louviere et al. (2008) and Frischknecht et al. (2011b) provide guidance on how OLS and

[3] I.e. a sample not used in estimation.

WLS[4] estimation can be used for ILMs. For OLS, the $0-1$ choice indicators are directly used as the dependent variable, while, for WLS, regression models can be estimated from the expected choice frequencies associated with the partial BW rankings or the full ranking of the choice options in each set. As an example, in the case of BW with four alternatives, Louviere et al. (2008) and Frischknecht et al. (2011b) suggest using ln(8) for best choice, ln(1) for worst choice and ln(3) for both middle ranks. For full ranking, they suggest taking values of ln(8), ln(4), ln(2) and ln(1), respectively. As noted by Frischknecht et al. (2011b) in the case of OLS and WLS, the model estimates are on the wrong "scale" to correctly predict the choice probabilities. The parameter estimates are too small (or large) and systematically under (or over)-predict the observed choice probabilities. In order to correct the estimates, Louviere et al. (2008) suggests a correction approach which is further in Frischknecht et al. (2011b). It equates to estimating an OLS (or WLS) model for a single person and then using this model to predict the observed dependent variable for each choice option in each choice set of interest. A next step calculates the associated residual mean squared error (MSE) for each person, and the OLS (or WLS) model estimates for an individual are then multiplied by a correction factor given by $\frac{1}{MSE}$.

2.6 Individual Bayesian Models with Priors Based on Sample Level Results

Finally, we introduce a new individual level model estimated using a Bayesian framework (ILM-B). For this model, in contrast with sample level Bayesian estimation, we treat each individual's data as independent in estimation (i.e. using a separate model per individual), where we however use a sample level model to inform the assumptions of b_0, v and S - in our case a MMNL model. This differs from the approach laid out by Frischknecht et al. (2011a) in that our priors used are informed by sample level behavior. This mitigates some of the issues faced by sample level HB models where an analyst needs to assume a prior variance which can have major effects on the individual posteriors estimates. The mean and variance estimates from a classically estimated mixed logit where chosen to serve as the informative priors in the model estimation. As with standard HB models, the variance of the prior controls the amount of fitting to the individual level that occurs. To test the effect of changing the prior variances on the ILM-B models, a scalar multiplicative factor was introduced. These scalar factors adjust the prior variances up or down while still maintaining the relative differences in variance for each parameter. Similar to the HB models, a larger multiplicative factor will result in more emphasis being placed on individual choice behavior while smaller factors would result in individual Bayesian models that do not differ far from the mean. An advantage of ILM-B over classically estimated ILMs is that, by drawing on sample level information, individual level results can still be obtained even for respondents where separation problems arise.

[4] See e.g. Greene (2011).

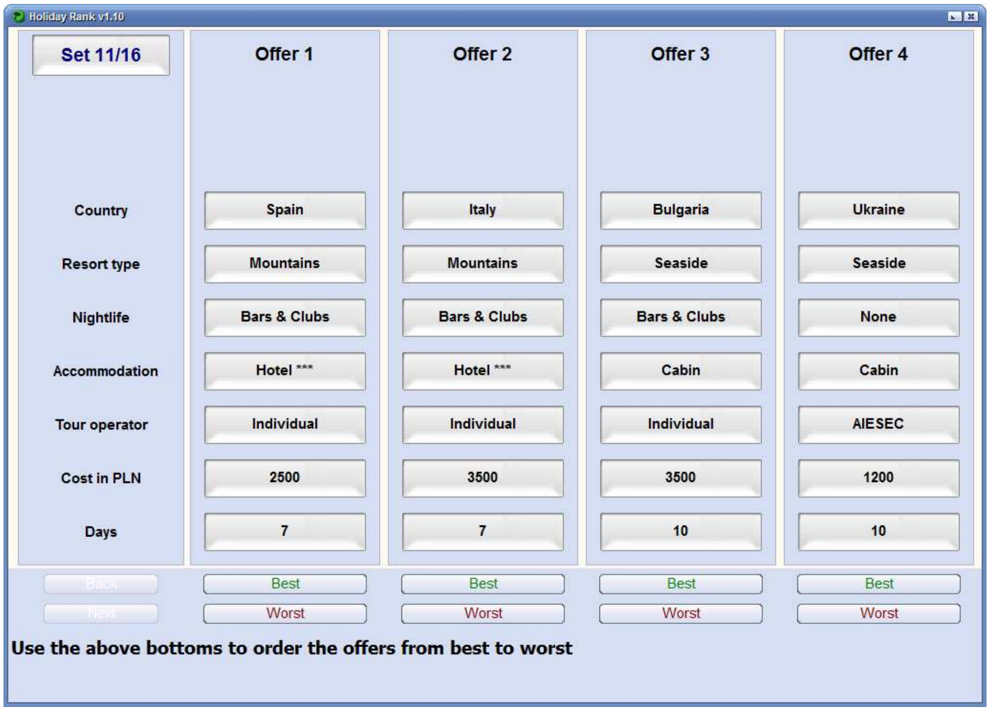

Figure 1: Example of choice scenario

3 Data

The data used in this paper comes from a study conducted on a sample of 822 students at the University of Warsaw in 2012. The students were requested to participate in nine independent sessions of the same stated choice survey about holiday destinations, where each session presented respondents with sixteen choice tasks. This gives a total of 144 tasks per person. To the best of our knowledge, only the paper by Brazell and Louviere (1998) makes use of a similarly long panel. In each choice task, four holiday options of the same type were presented to the respondent, where three types of holiday were used across the nine sessions; general, at the seaside, and active (culture & nature). Figure 1 provides an example of the trade-off experiment presented to the respondent. Each holiday type was described by five overlapping attributes; price, length, tour operator, nightlife, accommodation, in addition to a number of holiday specific attributes. The full list of attributes and their corresponding levels is presented in Table 1, along with the labels used in the remaining tables and figures in the paper. The presence of a price attribute allows us to estimate willingness-to-pay (WTP) measures for the other components in our empirical analysis.

The elicitation format focused on the full preference ordering over the four alternatives, and respondents were asked to provide this in three different choice formats, namely 1) best-worst-best, 2) best-best-best, 3) free to determine (any combination of best and worst)[5]. The

[5]In best-worst-best, the respondents were asked in alternating order to indicate the best and the worst alternatives. The process was sequential, i.e. the chosen best (or worst) alternative was then removed from

Table 1: Data characteristics

Attributes	Levels
Overlapping attributes	
Price	1000, 1500, 2000, 2500 (PLN)
Length (DAYS)	7, 10 days
Organisation (AISEC)	Individual (0), Organized by students' organization (1)
Nightlife (NIGHT)	None (0), Numerous discos and bars (1)
Accommodation (ACCOM)	Camping (0), Hotel (1)
General holiday	
Country	Bulgaria (BG), Italy (IT), Spain (SP), Ukraine (UK)
Location	Seaside (SEA), Mountains (MOUNTAINS)
At the seaside	
Beach type (BEACH)	stone (0), sand (1)
Walking distance to the beach in kilometres (BEACH-KM)	0.1, 0.5, 1.5 km
Water transparency in metres (CLEAR)	1.5, 3, 10 m
Crowding at the beach (CROWD)	Low (0), Medium (1), High (2) (photos used)
Active	
Historical and Cultural Sites (CULTR)	Locally (1), Country known (2), Internationally known (3)
Distance to cultural attractions in kilometres (CULTR-KM)	1, 10, 50 km
Nature attractions (NATUR)	Locally (1), Country known (2), Internationally known (3)
Distance to nature attractions in kilometres (NATUR-KM)	1, 10, 50 km

choice format alternated across sessions and each respondent saw each format three times across the nine survey sessions, with each format being used with each holiday type, and with the orders being randomised. The same design, composed of 48 choice tasks, was used for each of the 3 choice formats. Within the same holiday type, respondents saw three different sets of 16 choice tasks for a total of 48 choice tasks. Data collection was computer based, and a Bayesian d-efficient design optimized for MNL model was used, with priors taken from the pilot study conducted on a sample of 120 students in which each respondent was faced with 4 choice sets (4 alternatives each).

4 Empirical work

We start our analysis by reporting results for two sample level models in Table 2, one MNL model, and the upper level (i.e. sample level) model of a HB structure. Both models were estimated on the sample of 700 respondents making trade-offs involving all attributes[6].

subsequent questions. In best-best-best, the full preference order was obtained through a sequence of 'best' sequential questions as in the standard Ranking approach. In the final approach, respondents were free to determine the elicitation format, and the full preference ordering was obtained through any combination of best and worst; again the process was sequential.

[6]Non-traders are those respondents who did not use specific attributes in making their choices. These respondents were removed to avoid issues with separation and quasi-separation in the ILM-ML models.

Table 2: Estimation results for sample level models

	MNL		HB: sample level model			
Respondents	700		700			
Observations	100,800		100,800			
Final LL	-100,197		-69,411			
par.	20		230			
adj. ρ^2	0.2828		0.5016			
	est.	t-rat.	mean	t-rat	std. dev.	t-rat
ASC_2	12.9970	3.29	15.6268	2.97	9.1582	13.43
ASC_3	12.0640	2.94	12.5106	2.33	9.6369	13.15
ASC_4	1.5874	0.38	2.5671	0.43	9.9623	12.26
ACCOM	-258.2600	-17.77	-233.9588	-16.14	138.5360	16.60
AISEC	-51.8320	-4.74	-62.7921	-4.77	112.1556	15.84
BEACH	-350.8500	-18.35	-311.4717	-19.98	151.2714	14.42
BEACH-KM	224.5900	21.08	222.1589	22.32	58.1602	14.41
CLEAR	-12.6800	-8.62	-12.9808	-4.97	4.5203	17.64
COST	-0.0023	-44.63	-0.0047	-56.00	0.0041	14.44
CROWD	124.0700	15.56	108.6465	15.44	28.2514	13.99
CULTR	-198.5900	-19.56	-188.0427	-17.29	74.5124	15.72
CULTR-KM	4.6275	18.86	5.0779	2.48	2.9224	18.62
DAYS	-75.6980	-30.69	-76.0011	-23.18	6.8027	17.43
NATUR	-113.4900	-12.43	-97.0410	-11.11	45.6564	14.34
NATUR-KM	2.8200	10.62	2.8983	1.41	2.9656	18.57
NIGHT	-169.8300	-14.62	-203.3722	-14.44	134.0458	16.33
SEA	-374.6500	-16.77	-325.5011	-14.90	318.0600	15.55
BG	-223.7500	-14.16	-209.4899	-17.72	67.7571	10.38
IT	-479.4800	-20.09	-419.9530	-22.31	210.0695	12.73
SP	-568.2800	-22.88	-494.3252	-24.63	241.0464	13.68

For these models, we use a willingness-to-pay (WTP) space specification and assume Normal distributions for the random coefficients in the HB model, where we used the industry default of a prior variance of 2 for all parameters (Sawtooth (2009)) with off-diagonal covariance elements set to zero. Normal distributions were assumed for the random parameters due to the exploratory nature of this work, while working in WTP space meant that no issues with undefined moments for WTP measures would arise. It should also be noted that when looking at individual specific conditional means, the sign violations are much smaller than with the sample level distributions. As an example, even with the assumption of a Normal for price sensitivity, fewer than 1% of the sample had posterior price sensitivities that were positive. There is clearly a possibility that the findings might differ somewhat if the same analysis was conducted in preference space, for example due to a differential impact by the priors. We leave such an extension for future work as it would also require different distributional assumptions in the WTP space application for reasons of comparability.

The results show first of all a far superior model fit for the HB model, indicating substantial levels of heterogeneity in sensitivities across individual respondents. For the MNL model, point values are reported, while for the HB model, we report the means and standard deviations across respondents. Given that the models are estimated in WTP space, except for the cost coefficient, all values reported relate to monetary valuations of the associated attribute. In this case, a positive value means that a respondent is willing to pay extra money (expressed in PLN) for reducing the associated attribute (e.g. for reducing distances and/or crowding) while a negative value means that a respondent is willing to pay extra money for that attribute (e.g. beach instead of mountain holiday, sand instead of stone beach, Spain instead of UK).

In terms of estimates, we see that all holiday characteristics matter to respondents, with some, like duration and destination, mattering much more than others, such as who organises the holiday. In the HB model, we see some fluctuations in mean WTP measures compared to the MNL model, but they remain rather stable overall. However, we see substantial variations across individuals in their valuations for all characteristics.

We next move to our analysis looking at focussing on individual respondents. We start with an analysis on the full data before looking at subsets and discussing prediction performance. While the majority of existing studies have estimated ILM models using ranking or BW data, our main focus is on comparing models estimated on best choice answers only. This is made possible by having 144 choices per person, as discussed in Section 3. We further justify this reliance on best choices only by results in Giergiczny et al. (2013) which show, across different datasets, that sensitivities from BW data reveal the same or possibly even larger inconsistencies across stages than standard ranking, going beyond the differences in scale commonly taken into account[7]. The only exception we make to using data on best choices only is for ILM models estimated using OLS/WLS, where we look at estimates for three variants, namely best choice only, BW choices and full ranking[8].

ILM models estimated using maximum likelihood were estimated in Matlab, OLS/WLS ones in STATA, and the HB models were estimated using the R package RSGHB. To estimate ILMs using MML, we used software available from the CENSOC webpage[9], where we made use of two available options to identify a suitable prior weight, specifying a flat conjugate prior or an empirical Bayes conjugate prior. The MML method requires dividing the data for each individual into an estimation sample and a validation sample of equal sizes. This procedure was repeated ten times to create ten estimation and validation sample pairs, while the number of prior weight values to search was set to 15. This follows the recommendations in Frischknecht et al. (2011a).

[7]In all the datasets, significant and substantial differences between marginal utility estimates and implied monetary valuations across stages were found. The obtained results consistently show that pooling all stages and estimating joint models results in substantial bias in willingness-to-pay (WTP) values for numerous attributes.

[8]We make this exception because it is a standard practice in applications that analyze individuals' choices using OLS/WLS.

[9]http://www.censoc.uts.edu.au/resources/software.html

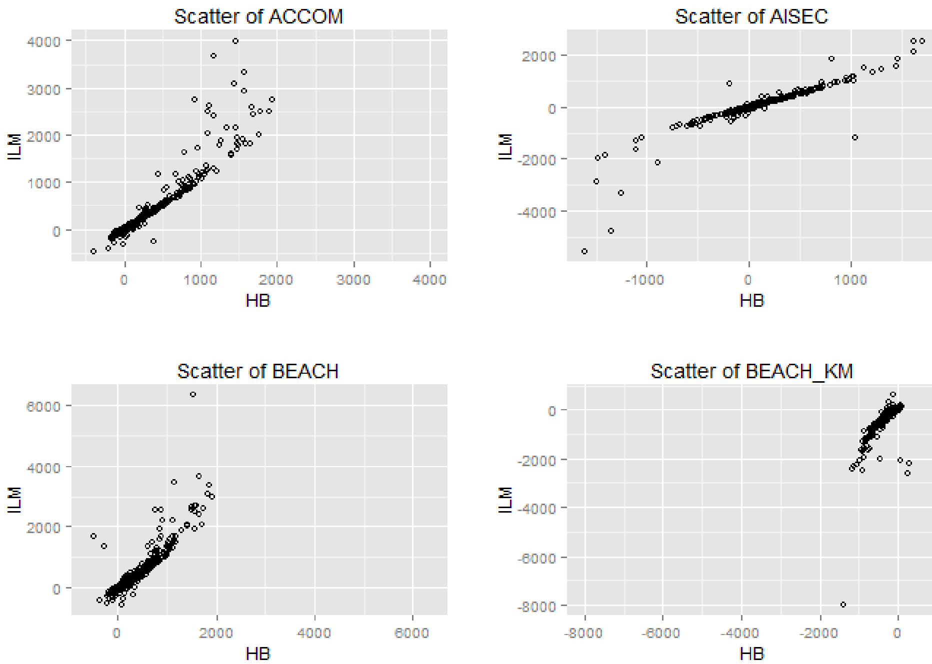

Figure 2: Scatterplots for selected WTP measures: analysis on full set of choices

4.1 Analysis on full set of choices

We next conduct a comparison between the results of the HB model at the individual level (i.e. the lower level model using the conditional distributions) and the individual models estimated using maximum likelihood (ILM-ML). The log-likelihood calculated at the means of the conditional distributions for the HB model is -57,406.9 while the combined log-likelihood across the ILM-ML models is -50,968.05. With no distributional assumptions imposed on the ILM-ML models, this difference in model fit is to be expected.

The key interest in this part of the analysis is in differences in the actual estimation results. Comparisons were made using scatterplots (Figure 2) and densities of individual-specific mean values (Figure 3) for selected WTP measures (WTP for hotel vs camping, students' organisation vs individual, sand vs stone beach, and beach distance), as well as Kolmogorov-Smirnov Tests for differences in distributions of all WTP measures (Table 3). The scatter plots and densities demonstrate at least qualitatively that the HB conditional means and ILM-ML point estimates are very consistent with each other for the four attributes shown here. The scatterplots reveal some outliers, mainly for ILM, where there would be more expectation of overfitting to the individual respondents.

The density function of the WTP for the beach (sand vs stone) attribute is also slightly more peaked for the individual posterior means from the HB model compared to the ILM-ML estimates. The results of the Kolmogorov-Smirnov tests provide more quantitative evidence to suggest that the majority of the WTP distributions are not statistically different between the

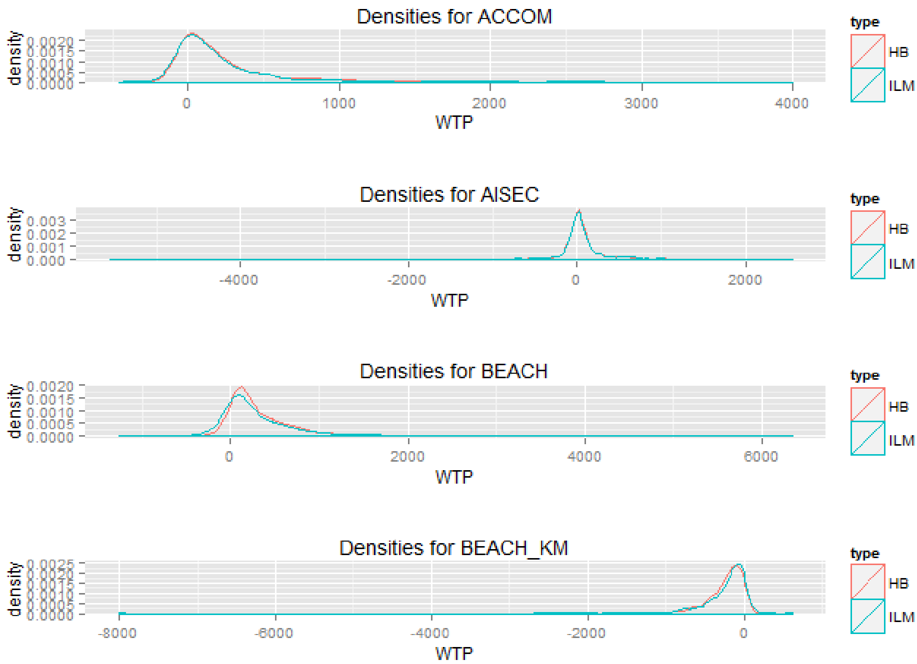

Figure 3: Selected densities for analysis on full set of choices

two approaches. Nevertheless, the results indicate that the distribution of the WTP values for beach type (sand over stone), crowding and Bulgaria (BG) holidays are statistically different between the ILM-ML and HB specifications.

4.2 Analysis on subsets of data

In practice, it is of course extremely rare to have data with as many as 144 choice tasks per individual. For the next step in our analysis, we thus reduced the number of choice tasks to 48 by focusing on a single holiday type, in this case the nature and culture holidays (active). The attributes used in this case are the five overlapping attributes from Table 1, i.e. price, length, type, nightlife and accommodation, along with the four holiday type specific attributes, namely cultural sites, distance to these sites, nature sites, and distance to these sites. Working again with traders only (across the 48 tasks), this yielded a sample size of 741 respondents, where this number is larger than when working with the full data, as more issues with non-trading arose in the other types of holidays.

We repeated the same analysis from Section 4.1 using the WTP space utility specification but added in OLS relying on best choice only as a further estimation technique. Figure 4 again provides some selected WTP densities comparing the results of the three models (type of accommodation, organisation, cultural sites and distance to cultural sites). Qualitatively, the ILM-OLS models show the biggest differences across the distributions presented, while HB and ILM-ML are more similar. This is further supported by looking at the Kolmogorov-

Table 3: Kolmogorov-Smirnov tests for differences between HB and ILM-ML for analysis on full set of choices

	Statistic	p-value*
ACCOM	0.034	0.81
AISEC	0.026	0.97
BEACH	0.103	0.00
BEACH-KM	0.066	0.10
CLEAR	0.034	0.81
CROWD	0.121	0.00
CULTR	0.031	0.88
CULTR-KM	0.050	0.35
DAYS	0.044	0.50
NATUR	0.069	0.07
NATUR-KM	0.034	0.81
NIGHT	0.030	0.91
SEA	0.034	0.81
BG	0.127	0.00
IT	0.057	0.20
SP	0.060	0.16

* The Kolmogorov-Smirnov test is a test of statistical difference.

Smirnov tests in Table 4, which show significant differences between HB and ILM-OLS for all WTP measures and for the majority of them when comparing ILM-ML and ILM-OLS. On the other hand, ILM-ML and HB remain more similar.

4.3 Model performance in prediction

The analysis thus far has focused solely on a comparison of model estimates. We now turn to the equally important issue of prediction performance. Individual level estimates are commonly used in forecasting especially in the context of market share simulators (Orme (2010)), and it is thus also crucial to establish any risks of overfitting to the estimation data with the different approaches.

We perform these tests in terms of forecasting performance on a set of within-sample hold out tasks. To do this, we further subset the active and culture holiday data by randomly selecting hold out samples of 12, 24 and 36 choice tasks from the set of 48 choice tasks for each respondent, leaving 36, 24 and 12 choice tasks, respectively, for model estimation. With this further reduction in estimation sample size, issues with separation arise for many respondents, and we thus avoid the estimation of ILM-ML models given the resulting convergence issues[10].

A larger number of models were estimated in this analysis to allow for further insights, where we now use estimation in preference space rather than WTP space, which thus also

[10] As an illustration, with 48 tasks, we can estimate ILM models for all 741 respondents, but this drops to 601 respondents with estimation on 36 tasks, 306 with 24 tasks, and only 81 with 16 tasks.

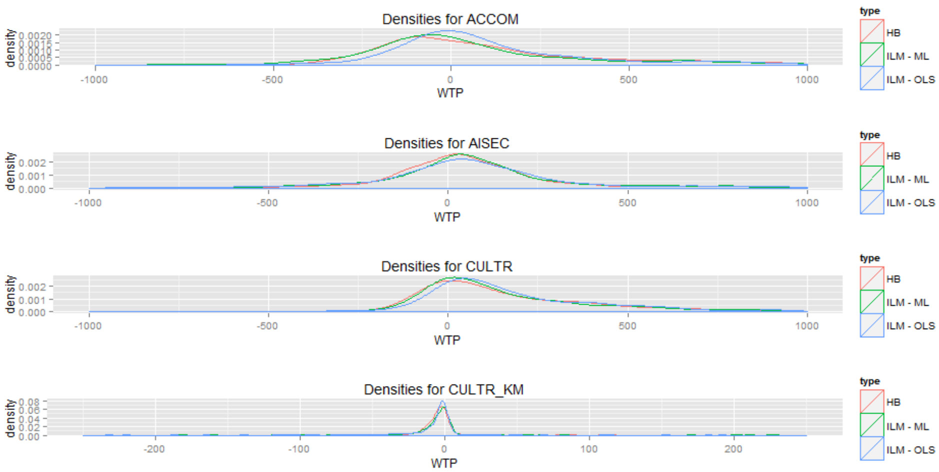

Figure 4: Selected densities on subset of 48 choices for nature and culture holidays

Table 4: Kolmogorov-Smirnov tests for differences between HB, ILM-ML and ILM-OLS on subset of 48 choices for nature and culture holidays

	p-values		
	HB v ILM-ML	HB v ILM-OLS	ILM-ML v ILM-OLS
ACCOM	0.06	0	0
AISEC	0.07	0.01	0.52
CULTR	0.31	0	0
CULTR-KM	0.03	0	0.04
DAYS	0	0	0
NATUR	0.17	0	0
NATUR-KM	0.08	0	0
NIGHT	0.15	0.01	0.14

allows us to generalise our findings beyond the WTP space context.

In terms of sample level models, we included MNL and classically estimated (through maximum simulated likelihood) MMNL models (MMNL-MSL), along with MMNL models estimated using Hierarchical Bayes (MMNL-HB), with different settings for prior variance and degrees of freedom. For MMNL-MSL and MMNL-HB, we also computed respondent specific conditional means for the various parameters. For MMNL-HB, we tested specifications with different prior variances (PV) and degrees of freedom (DF). In both MMNL-MSL and MMNL-HB, Normal distributions were once again used for all parameters and the full parameter covariance matrix was estimated. The earlier provisos about Normal distributions still arise, but once again, no WTP values were computed on the basis of ratios of coefficients and we were solely concerned with prediction performance.

In terms of ILMs, we estimated models on the data for best choices only, using ordinary least squares (ILM-OLS), modified maximum likelihood (ILM-MML) and our new proposed approach using Bayesian estimation of ILM models with priors from a sample level MMNL model (ILM-B) estimated on the full sample of respondents. Finally, we also included two models estimated using ordinary least squares on the extended set of individual choices, namely on best-worst choices (ILM-OLS-BW) and on the full ranking of choices (ILM-OLS-RANK). This is for completeness and to compare our models with those typically used. To compare the performance of the different models in forecasting, the sample level log-likelihood was calculated for each model on each of the three hold out samples, using the estimates from the corresponding estimation sample (i.e. 12 estimation tasks for the hold out sample of 36 tasks).

Table 5 shows the results when using a sample of 12 tasks in estimation and the remaining 36 in prediction. In general, the ILM models provide better fit on the estimation sample compared to the sample level models. However, with only 12 observations for estimation and 12 parameters, there is strong potential for non-trading on attributes of the alternatives, leading to a risk of overfitting of ILM models. As a result, while the ILM techniques do better in terms of fitting the estimation data, the better model fit does not necessarily translate into better forecasting accuracy with the exception of the ILM-B model using a variance scalar of 0.1. As one might suspect given the inclusion of the additional data (providing more information about an individual's preferences), the ILM-OLS-BW and ILM-OLS-RANK models outperformed the ILM-OLS model in predicting the hold out choices when looking at first choices only. However, when comparing across all models, their performance on the hold out sample was still only average. The best model in terms of prediction performance (of those tested) was the ILM-B using a variance scalar of 0.1 - the use of a lower variance reduces overfitting on the estimation data. While these are individual level models, they are better described as "hybrid" models as they borrow heavily from sample level behaviour via the assumed priors.

For sample level models, the classically estimated mixed logit model performs the best in prediction on the hold out sample by about 700 points of log-likelihood in comparison with the best MMNL-HB model. Except for the MMNL-HB model with the lowest prior variance, the use of conditionals/posteriors results in poorer performance on the hold out tasks than when using the same level model. This is true even though conditionals outperform the sample level models in explaining the estimation sample. The poor performance is particularly evident for HB models when increasing the prior variance which results in similar over-fitting issues that we see in the ILM models. This is a clear warning as to the continued reliance on conditionals in forecasting given the high risk of overfitting to the estimation sample when using modest numbers of tasks per person.

Table 6 provides the results of the forecasting on 24 hold out tasks while using the remaining 24 tasks for model estimation. Again, we see that the ILM models generally do significantly better in fitting the estimation data but that better fit does not translate into better performance on the hold out sample. Of the ILM models estimated, the ILM-B again performs best in explaining the hold out choices, and when using a small variance, performs best of all the models.

In this case, the sample level models again generally outperform the ILM (though not ILM-B) in terms of hold out prediction with MMNL-HB with a low prior variance offering the best performance when working with sample level parameters, where this now comes

Table 5: Results using estimation sample of 12 observations per individual

Model	k	Estimation Log-likelihood	Hold out Log-likelihood	Estimation LL with conditionals	Hold out LL with conditionals	Improvement in Hold out with conditionals
MNL	12	-8922	-27070	N/A	N/A	
MMNL-HB (PV 0.01 & DF 1)	90	-7865	-22249	-6253	-22070	0.8%
MMNL-HB (PV 2 & DF 5)	90	-8268	-25501	-4245	-28615	-12.2%
MMNL-HB (PV 10 & DF 5)	90	-9597	-33518	-3356	-42258	-26.1%
MMNL-MSL	90	-7624	-21546	-4508	-21605	-0.3%
ILM-OLS	8892	-4535	-28442	N/A	N/A	
ILM-OLS-BW	8892	-6319*	-25476*	N/A	N/A	
ILM-OLS-RANK	8892	-6380*	-24678*	N/A	N/A	
ILM-MML (flat prior)	8892	-3779	-28388	N/A	N/A	
ILM-MML (empirical Bayes)	8892	-3821	-28179	N/A	N/A	
ILM-B (Scalar=0.1)	8892	-4155	-21471	N/A	N/A	
ILM-B (Scalar=1)	8892	-1994	-28092	N/A	N/A	

* These likelihoods were calculated using the first choice only, but estimation was carried out on the full data.

Table 6: Results using estimation sample of 24 observations per individual

Model	k	Estimation Log-likelihood	Hold out Log-likelihood	Estimation LL with conditionals	Hold out LL with conditionals	Improvement in Hold out with conditionals
MNL	12	-17868	-18098	N/A	N/A	
MMNL-HB (PV 0.01 & DF 1)	90	-14599	-14677	-11299	-13080	-15.76%
MMNL-HB (PV 2)	90	-15702	-15752	-10203	-14467	8.2%
MMNL-HB (PV 10)	90	-17402	-17499	-9882	-15566	11.0%
MMNL-MSL	90	-14657	-14578	-10270	-13153	9.8%
ILM-OLS	8892	-10308	-16872	N/A	N/A	
ILM-OLS-BW	8892	-13748*	-16347*	N/A	N/A	
ILM-OLS-RANK	8892	-14044*	-16283*	N/A	N/A	
ILM-MML (flat prior)	8892	-7497	-16009	N/A	N/A	
ILM-MML (empirical Bayes)	8892	-7666	-15834	N/A	N/A	
ILM-B (Scalar=0.1)	8892	-9643	-13243	N/A	N/A	
ILM-B (Scalar=1)	8892	-7064	-15419	N/A	N/A	

* These likelihoods were calculated using the first choice only, but estimation was carried out on the full data.

close to ILM-B. An interesting contrast arises in relation to the performance with conditionals when comparing the results to those obtained with an estimation sample of just 12 tasks. Indeed, with 24 choice tasks, we now see that forecasting with conditional/posterior outperforms forecasting with the sample levels models, with the MMNL-HB model with the lowest prior variance and the MMNL-MSL now doing marginally better than ILM-B when using conditionals. This suggests that with 24 tasks, there is now far reduced risk of overfitting to the estimation data, compared to the sample with just 12 choices. Nevertheless, 24 tasks is more than what most studies have available to them, and forecasting with conditionals thus remains risky.

Table 7 provides the results of the forecasting on the 12 hold out tasks while using 36 tasks for model estimation. Except for the ILM-OLS-BW and ILM-OLS-RANK models, the ILMs still outperform the sample level models in fit on the estimation data[11]. We now see for the first time similar performance in hold out prediction between the ILM models and sample level models, a reflection of the fact that with a larger estimation sample, the risk of overfitting is reduced. ILM-B agains performs best of all models when not using conditionals from MMNL. However, with conditionals, the sample level MMNL-HB (PV 0.01 & DF 1) performs best overall, slightly ahead of ILM-B.

5 Conclusions

This paper has sought to add empirical evidence and critical discussion in the context of growing interest in modelling decisions at the level of individual agents. In particular, we have contrasted the performance obtained with sample level models and individual level models in both estimation and prediction. To this extent, we have made use of a dataset containing 144 choices per individual respondent but which is structured in such a way as to easily allow for the use of subsets of the data. Along with comparing the performance of existing approaches, we have also proposed the use of a new "hybrid" method, the ILM-B approach.

Our results on the full set of choices suggests that in the presence of substantial data at the individual level, the individual-specific conditional parameters obtained from sample level models (e.g. through HB estimation) are overall quite similar to those from individual level models, where the latter has a slightly larger tendency to produce outliers. When working with smaller subsets of the data, the differences between sample level and individual level models increase, where, in the former, the lower level of information at the person level gives more weight to the sample level results. The differences are also larger when comparing HB with OLS estimates than with ILM-ML estimates. This raises some questions as to the quality of estimates obtained with OLS in current practice, especially when such estimates are obtained from data where ILM-ML estimation is not possible.

When working with sample sizes at the individual level that are more representative of the majority of applied studies, we find that sample level models can be as good or even outperform those individual level models currently estimated in practice, when looking at out-of-sample prediction. There appears to be some promise in estimating individual level models

[11]ILM-OLS-BW and ILM-OLS-RANK are likely to do less well in explaining first choices only as they are estimated on the entire ranking.

Table 7: Results using estimation sample of 36 observations per individual

Model	k	Estimation Log-likelihood	Hold out Log-likelihood	Estimation LL with conditionals	Hold out LL with conditionals	Improvement in Hold out with conditionals
MMNL-HB (PV 0.01 & DF 1)	90	-21231	-7728	-14856	-6244	19.2%
MMNL-HB (PV 2)	90	-22997	-8119	-13278	-6600	18.7%
MMNL-HB (PV 10)	90	-25485	-8859	-12895	-6837	22.8%
MMNL-MSL	90	-21690	-7765	-16390	-6386	17.8%
ILM-ML	8892	-10390	-20774	N/A		
ILM-OLS	8892	-17436	-7781	N/A	N/A	
ILM-OLS-BW	8892	-22373*	-8170*	N/A	N/A	
ILM-OLS-RANK	8892	-22397*	-8074*	N/A	N/A	
ILM-MML (flat)	8892	-12901	-7204	N/A	N/A	
ILM-MML (empirical Bayes)	8892	-12740	-7253	N/A	N/A	
ILM-B (Scalar=0.1)	8892	-15077	-6365	N/A	N/A	
ILM-B (Scalar=1)	8892	-12486	-7108	N/A	N/A	

* These likelihoods were calculated using the first choice only, but estimation was carried out on the full data.

using Bayesian techniques mixed with priors that are drawn from sample level behaviour, where this ILM-B approach produces the best prediction performance across all sample sizes compared to using sample level MMNL estimates or traditional HB models. Only in the case of large estimation samples do the conditionals from HB or MMNL models perform better. Here, it is clear that the benefit of using conditionals from sample level models increases with the amount of information collected at the individual level.

While those ILM models commonly used at the moment do have better model fit on the estimation dataset across the analyses, this does not translate necessarily into better prediction performance given the tendency to overfit on the estimation sample. It should be acknowledged again that the majority of the work on ILMs partly addresses the issue of limited data at the person level by making use of data not just on the most preferred alternative in each task but through a fuller preference ordering, generally elicited through a best-worst approach. However, recent work by Giergiczny et al. (2013) has questioned the reliability of such data, and our results when using such data in the present paper in OLS estimation shows that it does not necessarily translate into a greater ability to predict individual choices in the holdout sample.

In closing, with the likely growing interest in individual level approaches also in a transport context, care must be taken to not overfit on estimation data using currently used ILM methods. Conditionals from sample level models (whether classical or HB) offer promise, but require large numbers of choices per person for good performance. The hybrid ILM-B approach put forward in this paper shows promising performance even with more modest numbers of tasks. Indeed, with limited data per person, sample level models are necessary to fill in the gaps at the individual level, and this is an inherent characteristic of that approach. Testing this approach on other data is an important avenue for future work.

References

Albert, A., Anderson, J. A., 1984. On the existence of maximum likelihood estimates in logistic regression models. Biometrika 71 (1), 1–10.

Altman, M., Gill, J., McDonald, M. P., 2003. Numerical Issues in Statistical Computing for the Social Scientist. Wiley-Interscience.

Beggs, S., Cardell, S., Hausman, J., 1981. Assessing the potential demand for electric cars. Journal of Econometrics 16, 1–19.

Ben-Akiva, M., Morikawa, T., Shiroishi, F., 1991. Analysis of the reliability of preference ranking data. Journal of Business Research 23 (3), 253–268.

Brazell, J. D., Louviere, J. J., 1998. Length effects in conjoint choice experiments and surveys: an explanation based on cumulative cognitive burden. Department of Marketing, The University of Sydney, Sydney, Australia.

Brownstone, D., Bunch, D. S., Train, K., 2000. Joint Mixed Logit models of stated and revealed preferences for alternative-fuel vehicles. Transportation Research Part B 34 (5), 315–338.

Burgess, L., Street, D., Viney, R., Louviere, J., 2006. Design of choice experiments in health economics. Edward Elgar, Cheltenham.

Chapman, R., 1984. An approach to estimating logit models of a single decision maker's choice behavior. Advances in Consumer Research 11, 656–661.

Daziano, R. A., Bolduc, D., 2013. Incorporating pro-environmental preferences towards green automobile technologies through a bayesian hybrid choice model. Transportmetrica A: Transport Science 9 (1), 74–106.

Dumont, J., Keller, J., Carpenter, C., 2012. Rsghb: Functions for hierarchical bayesian estimation: A flexible approach.
URL http://cran.r-project.org/web/packages/RSGHB/

Ebling, C., Frischknecht, B., Louviere, J., 2010. Keep it simple: Easy ways to estimate choice models for single consumers. Australian and New Zealand Marketing Academy Conference, Christchurch, New Zealand.

Flynn, T., Louviere, J., Peters, T., Coast, J., 2007. Best-worst scaling: What it can do for health care research and how to do it. Journal of Health Economics 26, 171–189.

Frischknecht, B., Eckert, C., Geweke, J., Louviere, J., 2011a. Incorporating prior information to overcome complete separation problems in discrete choice model simulation, centre for the Study of Choice, University of Technology.

Frischknecht, B., Eckert, C., J., L., 2011b. Simple ways to estimate choice models for single consumers, centre for the Study of Choice, University of Technology.

Gelman, A., 2006. Prior distributions for variance parameters in hierarchical models. Bayesian Analysis 1 (3), 515–533.

Geweke, J., 2005. Contemporary Bayesian Econometrics and Statistics. John Wiley & Sons.

Giergiczny, M., Czajkowski, M., 2011. Investigating preference heterogeneity through individual-level modelling. International Choice Modelling Conference. Leeds 2011.

Giergiczny, M., Dekker, T., P.K., C., S., H., 2013. Consistency (or lack thereof) between choices in best-worst surveys. International Choice Modelling Conference. Sydney 2013.

Green, P., 1984. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. Journal of the Royal Statistical Society 46 (2), 149–192.

Greene, W., 2011. Econometric Analysis. MacMillan Publishing Company.

Greene, W. H., Hensher, D. A., 2013. Revealing additional dimensions of preference heterogeneity in a latent class mixed multinomial logit model. Applied Economics 45 (14), 1897–1902.

Hausman, J., Ruud, P., 1987. Specifying and testing econometric models for rank-ordered data. Journal of Econometrics 46, 83–103.

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.

Ishaq, R., Bekhor, S., Shiftan, Y., 2014. A latent class model with fuzzy segmentation and weighted variables. Transportmetrica A: Transport Science 10 (10), 878–893.

Lancsar, E., Louviere, J., 2008. Estimating individual level discrete choice models and welfare measures using best worst choice experiments and sequential best worst mnl, centre for the Study of Choice, University of Technology.

Levinson, D., 2010. Equity effects of road pricing: A review. Transport Reviews 30 (1), 33–57.

Louviere, J., Woodworth, G., 1983. Design and analysis of simulated consumer choice or allocation experiments: An approach based on aggregate data. Journal of Marketing Research 20, 350–367.

Louviere, J. J., Street, D., Burgess, L., Wasi, N., Islam, T., , Marley, A. A. J., 2008. Modeling the choices of individual decision-makers by combining efficient choice experiment designs with extra preference information. Journal of Choice Modelling 1 (1), 128–163.

Marley, A., 2009. The best-worst method for the study of preferences: theory and applications. Proceedings of XXIX international congress of psychology.

Marshall, D., Chan, S., Curry, J., 2011. A head-to-head comparison of the traditional (top-down) approach to choice modeling with a proposed bottom-up approach. Sawtooth Software Conference, Newport Beach, CA, USA.

Orme, B., 2010. Getting Started with Conjoint Analysis: Strategies for Product Design and Pricing Research. Research Publishers, LLC.

Revelt, D., Train, K., 1998. Mixed Logit with repeated choices: households' choices of appliance efficiency level. Review of Economics and Statistics 80 (4), 647–657.

Rose, J., 2014. Interpreting discrete choice models based on best-worst data: A matter of framing. Working paper.

Rossi, P., Allenby, G., McCulloch, R., 2005. Bayesian Statistics and Marketing. John Wiley and Sons.

Sawtooth, 2009. The cbc/hb system for hierarchical bayes estimation version 5.0 technical paper.
URL http://www.sawtoothsoftware.com/support/technical-papers/hierarchical-bayes-estimation/cbc-hb-technical-paper-2009

Street, D., Burgess, L., Louviere, J., 2005. Quick and easy choice sets: Constructing optimal and nearly optimal stated choice experiments. International Journal of Research in Marketing 22, 459–470.

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