Performance evaluation and comparison

This section comprises the results we managed to achieve in the exploration of different performance metrics and provides insights on the functioning of the discussed mathematical models in a given context. As we have seen in the previous part, where the effects' estimates were provided, all of the models are able to provide some estimates for the retaliate utility function parameters in different discrete choice set-ups. The most simple models performed well on the dataset defined by the homogeneous preferences in the population for environmental attributes, underestimating the effects in the presence of preference heterogeneity. In the same time the more complex MMNL model performed sufficiently well in both behavioural set-ups, although it demonstrated some potential problems with the algorithmic implementation.

Overall precision

First of all we focus our attention on the general performance metrics, describing how well the estimated models fit the predicted outcomes over an original dataset. As we have discussed earlier we use only some of the available measures in an attempt not to make this work too cumbersome. The retained performance metrics are: accuracy, describing the overall goodness of fit over observed choices of the subjects; and more complex KDL measure, which compares the distributions instead of more simple metrics, which use only the information available in the confusion matrix.

We can observe the values of these general performance measures, describing overall performance of a given classifier in the table 1. The table regroups the metrics' values for all the estimated models.

Table 1: General performance measures

	Fixed effects			*Random effects*		
	MNL	MMNL	CNN	MNL	MMNL	CNN
Overall measures						
Accuracy	0.863	0.863	0.723	0.725	0.863	0.721
Probabilistic measures						
KLD	0.623	0.623	0.328	0.349	0.625	0.317

As we have underlined earlier we observe quite natural situation when the best model in terms of overall performance is the model, which was used in the data generation step. This situation perfectly demonstrates the potential bias, which is explained by our choice of the artificial data-generation algorithm. Nevertheless, it should be noted, that the MNL and MMNL models perform equally well on the fixed effects dataset, where the preferences for the environmental attributes are homogeneous. This fact supports our initial hypothesis that an implementation of a more complex model is preferred when the real effects are unknown to the researcher.

Focusing our attention on the CNN model observe that the *Adam* algorithm did not outperform the *BFGS* procedure. This observation may be explained by the data-generation set-up, where the generative algorithm favoured the MNL model, rather than *Adam*. The latter not supporting the fine tuning over the error distribution.

We can observe the results for the resources efficiency we managed to obtain, which are regrouped in the table 2. Even though we present all the time values, we are mostly interested with the "user" and "system" time values. The first one indicates the CPU time charged for the execution of user instructions of the calling process, while the second one stand for the CPU time spent for execution by the system on behalf of the calling process.

Table 2: Ressources efficiency

	Fixed effects			*Random effects*		
	MNL	MMNL	CNN	MNL	MMNL	CNN
User	20.910	452.414	17.433	18.722	2066.934	16.806
System	0.153	1.712	0.714	0.004	16.112	0.415
Total	21.068	454.192	8.412	18.726	2083.221	7.604

The more advanced *Adam* algorithm easily bypasses the algorithms available in the *mlogit* package, although this boost in efficiency goes at the cost of lower overall performance and goodness of fit. At the same time, the MMNL implementation is far less efficient and takes 128 times more time, than CNN model. This situation clearly illustrates us how the precision and flexibility come at higher costs.

Alternative specific metrics

We proceed with a look at some more specific measures. The table 3 regroups response specific metrics, that describe the precision of model in predicting only one target class of the dataset. These metrics are mostly used when we are interested in some in-depth insight into the model performance and allow to identify the models which perform the best over a single class of interest. Given the context of Michaud, Llerena, and Joly (2012) study we are interested in identifying the algorithm which predicts the best "buy" (A and B alternatives) against "not buy" (C) alternative, providing at the same time some information about the alternative chosen. In order to evaluate the performance at this dimension we use Geometric mean and the F-measure performance estimators.

Table 3: Variable specific performance measures, fixed effects data

	Fixed effects			*Random effects*		
	C	A	B	C	A	B
Geometric mean						
MNL	0.454	0.848	0.868	0.432	0.696	0.693
MMNL	0.454	0.849	0.867	0.452	0.848	0.867
CNN	0.443	0.697	0.698	0.447	0.697	0.700
F-measure						
MNL	0.318	0.834	0.873	0.282	0.666	0.704
MMNL	0.318	0.834	0.873	0.316	0.833	0.873
CNN	0.291	0.665	0.706	0.294	0.665	0.707

In the table 3 we are interested with the entries in the columns corresponding to the "No buy"

alternative (C). For the dataset with fixed effects across the population, the MNL and MMNL models perform identically according to both of the selected measures. The CNN model falls behind the econometrics models on the fixed effects dataset, although situation changes in the presence of heterogeneous effects. In the more complex case scenario, when the individuals have varying across population preferences towards one or another attribute, the CNN model outperforms the simple MNL model in detecting "No buy" decisions for given choice sets, which is rather interesting, because the overall model's performance is still inferior to the MNL, as it was shown in table 3.

Willingness to pay and premiums

Here we should present the most important results comparing the estimates for the WTP, as well as the premiums for particular attributes derived for different models. The Premium to pay for a rose's particular attribute as it was described previously can be represented as:

$$Premium = \frac{\frac{\delta V}{\delta X_k}}{\frac{\delta V}{\delta Price}} \tag{1}$$

At the same time, the WTP for a rose may be seen as the ratio of two corresponding coefficients of dummy variable and price. The table 4 presents the estimated WTP and premiums for the models, which output fixed coefficient estimates, without taking into account the randomness of the individual effects. In other words, this table regroups the results, which do not require bootstrapping for confidence interval estimation.

Table 4: WTP and Premiums obtained with MNL and CNN

	Fixed effects		*Random effects*		*Target*
	MNL	CNN	MNL	CNN	
WTP	1.421	1.377	0.747	0.751	1.401
Label	1.731	1.737	1.445	1.442	1.731
Carbon	4.091	4.101	3.679	3.669	4.086
LC	4.112	4.129	3.378	3.352	4.110

For the estimation of the WTP and the premiums for more complex models (the MMNL in our case) we use the same procedure, as was implemented by Michaud, Llerena, and Joly (2012). Because the random parameters are assumed to be correlated in the MMNL model's specification, the estimated standard deviations and confidence intervals are obtained using the Krinsky and Robb parametric bootstrapping method (**???**). This procedure consists of generating of multiple random draws from a multivariate normal distribution and using the obtained results to obtain the confidence interval estimates. Exactly as in the original study we generate 1000 draws from a multivariate normal distribution ($MNV(\mu, \Sigma)$), with the coefficient estimates as means μ and the estimated variance-covariance matrix of the random parameters as Σ.

The obtained results are then summarised as follows in the table 5

Comparing the estimates to the input values we observe that the variance of the WTP and Premiums estimates, estimated over a fixed effects dataset, do not potentially affect the conclusion one can derive from the results. The values stay positive with the 75% interval within 0.2€ of the mean

Table 5: WTP and Premiums obtained with MMNL

	Statistics					
	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Fixed effects						
WTP	1.416	0.058	1.233	1.377	1.455	1.613
Label	1.732	0.019	1.672	1.720	1.745	1.791
Carbon	4.097	0.103	3.730	4.026	4.166	4.434
LC	4.116	0.098	3.741	4.051	4.182	4.421
Random effects						
WTP	1.360	1.887	−4.239	0.073	2.662	7.893
Label	1.243	1.667	−3.867	0.104	2.330	6.638
Carbon	3.467	2.323	−4.026	1.880	5.043	11.671
LC	3.036	3.240	−7.430	0.908	5.160	14.259
Target						
WTP	1.418	1.973	−4.474	0.058	2.798	6.706
Label	1.735	1.611	−2.652	0.653	2.849	6.709
Carbon	4.076	2.134	−1.774	2.608	5.543	11.217
LC	4.106	3.379	−6.304	1.913	6.439	14.612

Note: The estimates are obtained with 1000 draws from MNV distribution

estimate. Assuming the model is not re-estimated and adjusted after the insignificant estimators are obtained for Choleski matrix elements, the results remain valid.

We may conclude, that given sufficiently large dataset the implementation of more complex model is preferable, because it will allow to control for unknown parameters without adding a risk of obtaining biased results. The more simple models, should be preferred in a more restricted context. They allow to obtain the valid results only in the case of correct theoretical assumptions, biasing the estimates in other conditions. Consequently, in the presence of uncertainty about the presence of heterogeneity in the customer choice modelling questions there is a strong interest to implement a more complex model, readjusting it afterwards if needed.

Michaud, Celine, Daniel Llerena, and Iragael Joly. 2012. "Willingness to pay for environmental attributes of non-food agricultural products: a real choice experiment." *European Review of Agricultural Economics* 40 (2): 313–29. https://doi.org/10.1093/erae/jbs025.