Simulating individual choices

Based on the article of Michaud, Llerena, and Joly (2012) we generate a synthetic dataset assuming the utility function is as described in the paper with some minor changes and adjustments. We have already delimited the scope of study and delimited our area of interest to the exploration of different models performance given the theoretical structure of consumer preferences for the alternative specific attributes. For simplicity we relax some of the assumptions made in the paper in order to generate two different datasets. For the first dataset we assume that estimations made in the paper and the derived utility functions are correct and reflect the real world situation. For the second one, we relax some of the advanced assumptions and regenerate a simplified version, which will allow us to contrast the performances of different models in different choice context assuming different nature of choice functions.

In both situations the utility functions are determined as in paper: we use the exact means for the coefficients estimates, assuming they are correct. The relative utility's deterministic part for each individual is defined by the following function, which was presented in a more detailed way in previous section:

$$V_{ij} = \alpha_{i,Buy} + \beta_{Buy,Sex} Sex_i + \beta_{Buy,Age} Age_i + \beta_{Buy,Salary} Salary_i + \beta_{Buy,Habit} Habit_i + \\ + \gamma_{Price} Price_{ij} + \gamma_{i,Label} Label_{ij} + \gamma_{i,Carbon} Carbon_{ij} + \gamma_{i,LC} LC_{ij} \quad (1)$$

Where $LC = Label \times Carbon$. The random component of the relative utility U_{ij} is defined as identically and independently distributed random variable ϵ_{ij} issued from the Gumble distribution parametrised with $(0, 1)$. The mean effects for the components of the deterministic part are given as presented in the table 1a

Table 1: The assumed relative utility function parameters

(a) Mean effects

	Effects
	Means
Individual characteristics (β)	
Sex	1.420
Age	0.009
Salary	0.057
Habit	1.027
Alternatives' attributes (γ)	
Price	−1.631
Buy	2.285
Label	2.824
Carbon	6.665
LC	−2.785

(b) Variance-covariance structure

	Effects	
	Fixed	Random
Variance		
Buy	0	3.202
Label	0	2.654
Carbon	0	3.535
LC	0	2.711
Covariance		
Buy:Label	0	-0.54
Buy:Carbon	0	-4.39
Buy:LC	0	6.17
Label:Carbon	0	8.77
Label:LC	0	-2.33
Carbon:LC	0	-4.82

The only difference between the two generated datasets is in the specification of the randomness of these coefficients as they may vary or not across population. It means, that the first dataset is

generated assuming the variance-covariance matrix for correlated random coefficients is composed with 0's only and the resulting multivariate normal distribution produces exact means for the coefficients. The second dataset is generated using the exact estimates of the variance-covariance matrix as provided in the article. The assumed parameters for effects distributions are represented in the table 1b.

Additionally we impose some supplementary constraints to our data due to the limitations of the simulation tool. Particularly, the individual characteristics are supposed to be not correlated, which can be explained by the fact that the context of a controlled experiment offers a possibility to control this particular feature. Obviously, this is not optimal decision, as naturally the age, sex, income and environmental habits of individuals should be correlated. Unfortunately, the original article does not provide information about the characteristics' variance-covariance matrix.

The targeted features and requirements to the resulting dataset are numerous and they make a contrast compared to the initial empirical dataset.

The simulated dataset allows us to explore significant number of choice sets for numerous artificial individuals, which ensures statistical validity for obtained results and permits us to use advanced estimation algorithms (such as neural networks, for example). It means that we generate a large sample with exhaustive number of choice sets, in which all the possible combinations of alternative attributes are represented. Here by *attributes* we understand the binary factors describing rose's labelling and carbon footprint and ignore the price, the latter being added afterwards using randomisation techniques. This choice is similar to the experimental design described in the Michaud, Llerena, and Joly (2012) work and is easily explained when we take a closer look at the number of choice sets for different specifications. In simulated datasets it is traditional to use Full-Factorial (FF) experimental design as it uncovers completely the full potential of simulation tools: it allows to observe all the possible combinations of factors affecting some process and fully explore their implications. In our case, a simple full factorial design for a binary choice context has 28 combinations of factors (seven levels of prices, two levels for eco-label and two levels for Carbon imprint), but a complete full factorial design for a choice context with two alternatives implies 784 different combinations (as we have two alternatives each having 28 possible variants), which is unrealistic in a standard experimental study context and risks to be too demanding in terms of calculation times.

The dataset should be equilibrated with relatively identical number of choices for all three alternatives. In the field experiment the authors managed to achieve satisfying result with 67.5% of "Buy" choices and 32.5% for "Not to buy" choices, although the "A" and "B" alternatives showed different properties. The resulting observed descriptive statistics derived from the data proposed by Michaud, Llerena, and Joly (2012) are presented in table 2. The table focusses on the choice "Buy" descriptive statistics, ignoring the "No buy" option, for which all the attributes are considered to be equal to 0. The p-values are the results of the two subsets ("A" and "B") comparison[1].

Of particular interest in the table 2 to us is the unbalanced structure of the resulting dataset. The *Carbon* imprint of the different alternatives has not identical properties, which leads to different *Choice* statistics, where the alternative with higher carbon imprint is chosen less frequently. In the original study such difference was not dangerous, because only the "Buy" option was compared against "No Buy" one. However, in case of the NN modelling such unbalanced dataset may lead to erroneous results, where the more popular alternative will always have a higher choice probability. The distribution inside the "Buy" group for different alternatives ("A" and "B") should be quasi-

[1] χ^2 test is used for discrete variables, while *Anova* is implemented for continuous ones.

Table 2: Alternatives' descriptive statistics by group, correlated random effects

	A (N=1186)	B (N=1186)	Total (N=2372)	p value
Choice				< 0.001
Mean (SD)	0.517 (0.500)	0.159 (0.366)	0.338 (0.473)	
Range	0.000 - 1.000	0.000 - 1.000	0.000 - 1.000	
Price				0.418
Mean (SD)	2.990 (0.881)	3.020 (0.893)	3.005 (0.887)	
Range	1.500 - 4.500	1.500 - 4.500	1.500 - 4.500	
Carbon				< 0.001
Mean (SD)	0.167 (0.373)	0.832 (0.374)	0.500 (0.500)	
Range	0.000 - 1.000	0.000 - 1.000	0.000 - 1.000	
Label				0.837
Mean (SD)	0.502 (0.500)	0.497 (0.500)	0.500 (0.500)	
Range	0.000 - 1.000	0.000 - 1.000	0.000 - 1.000	

identical, producing equally distributed three groups of choices each nearing 33.3%. Even if this property is not as important for a traditional MNL model, we are interested to observe the same choice structure in our artificial dataset, because it may highly affect the performance of more advanced models, such as NN for example.

Generated dataset presentation

In this section we will discuss the resulting datasets simulated under the listed above assumptions.

For our dataset we choose to generate 160000 observations, for 1000 individuals, each facing 16 different choice sets 10 times. The 16 choice sets include all the possible combinations of two roses ("A" and "B") described by two environmental attributes, while prices are randomly assigned within the choice sets. The prices are assumed to be uniformly distributed over the choice sets, following a discrete uniform distribution. The prices vary among the different replications. This procedure resulted in sufficiently large dataset, which in the same time was not difficult to treat without implementation of Big Data specific techniques.

The original experimental design used to generate the choice sets assumed no branding for the alternatives to avoid any undesired bias in the results. Theoretically this design should have provided an equilibrated dataset with no correlation between different attributes, although the size of the final dataset might have affected the results. In our case we assume that individuals have no additional information about the roses in choice sets except the three observed attributes. As in the original work we assign insignificant labels "A" and "B" to the roses within choice sets, which is done mostly for convenience and has no impact on the individuals' decisions.

It is interesting to explore the statistical properties of the resulting datasets: the original one (Original), gathered by Michaud, Llerena, and Joly (2012) and made available in anonymised format by Iragaël Joly; and the two generated artificial datasets, assuming homogeneous (Generated FE) and heterogeneous (Generated RE) preferences respectively of the individuals for the environmental attributes. First of all, we may observe the individuals descriptive statistics for three datasets in the table 3.

Table 3: Individuals' characteristics descriptive statistics by dataset

	Fixed Effects (N=1000)	Random Effects (N=1000)	Target (N=102)	p value
Sex				0.851
Mean (SD)	0.506 (0.500)	0.515 (0.500)	0.490 (0.502)	
Range	0.000 - 1.000	0.000 - 1.000	0.000 - 1.000	
Habit				0.182
N-Miss	0	0	1	
Mean (SD)	0.683 (0.466)	0.657 (0.475)	0.604 (0.492)	
Range	0.000 - 1.000	0.000 - 1.000	0.000 - 1.000	
Salary				< 0.001
Mean (SD)	2.750 (1.476)	2.671 (1.438)	2.147 (1.222)	
Range	1.000 - 6.000	1.000 - 6.000	1.000 - 6.000	
Age				0.255
Mean (SD)	41.862 (13.685)	42.161 (13.820)	39.755 (18.895)	
Range	18.000 - 84.000	18.000 - 84.000	18.000 - 85.000	

Even though the p-values show no evident differences between the simulated datasets and the original one, except for the *Age* variable, we observe the differences in the means. This is explained by the implemented dataset generation procedure. The variables in the original dataset are integers, assuming continuous nature of the real world variables. When synthesizing the dataset, we assume the quasi continuous variables, such as *Age* and *Salary* (denoted as *Income* in original work) to be issued from normal distribution with parameters as figuring in the descriptive statistics for the original dataset, and only afterwards we convert the resulting values to integers. The binary variables *Sex* and *Habit* are generated with random draws from Bernoully distribution and consequently produce more realistic results. This procedure leads to potential biases in the resulting datasets, which is true not only for the individual variables, but for the alternatives' attributes as well.

Table 4: Alternatives' descriptive statistics by dataset

	Fixed Effects (N=320000)	Random Effects (N=320000)	Target (N=2372)	p value
Price				0.002
Mean (SD)	2.936 (0.958)	2.936 (0.958)	3.005 (0.887)	
Range	1.500 - 4.500	1.500 - 4.500	1.500 - 4.500	
Carbon				0.999
Mean (SD)	0.500 (0.500)	0.500 (0.500)	0.500 (0.500)	
Range	0.000 - 1.000	0.000 - 1.000	0.000 - 1.000	
Label				0.999
Mean (SD)	0.500 (0.500)	0.500 (0.500)	0.500 (0.500)	
Range	0.000 - 1.000	0.000 - 1.000	0.000 - 1.000	

Secondly, we may as well observe the alternative specific descriptive statistics. They are presented in

table 4. In this table we present the cumulative statistics for the "Buy" option, including both rose "A" and rose "B" properties, while 160000 entries (1186 entries for the original dataset) describing the "No buy" alternative are omitted, because their attributes are reduced to zeros in order to achieve identifiability of the models (a complete presentation of descriptive statistics par dataset and stratified by alternative may be found in Appendix C). The distributions of *Carbon* footprint and Eco-*Label* attributes follows perfectly the ones inside the original dataset, although the prices differ. This particular divergence, may be explained by the procedure implemented to assign prices to the alternatives inside choice sets, because the random generator algorithms different across statistical programs and potentially the procedures implemented in *R* and *SAS* are not identical.

What is more interesting, is the difference in the *Choice* statistics. We may be interested in comparing the statistics for different classes in our sample to ensure that they are not biased in favour of label "A" or label "B", as in this case it risks to bias the estimates. For the artificial dataset the ratio of choices per "Buy" alternative is higher than 40% and reaches 47.3% for the fixed effect utility (table 5), while for the random effects specification the numbers are lower, reaching only 42% in mean for two classes (table 6). This particular observation is rather interesting as it demonstrates how the heterogeneous effects for alternatives' features the consumer decisions.

We will start with a close examination of the fixed effects dataset, where we can see, that prices are not equally distributed among the different choices.

Table 5: Alternatives' descriptive statistics by group, fixed coefficients

	A (N=160000)	B (N=160000)	Total (N=320000)	p value
Choice				< 0.001
Mean (SD)	0.427 (0.495)	0.518 (0.500)	0.473 (0.499)	
Range	0.000 - 1.000	0.000 - 1.000	0.000 - 1.000	
Price				< 0.001
Mean (SD)	3.069 (0.979)	2.803 (0.917)	2.936 (0.958)	
Range	1.500 - 4.500	1.500 - 4.500	1.500 - 4.500	

The unbalanced prices potentially bias our dataset and we can see how the option with inferior mean prices is chosen less frequently. Even thought this differences do not affect the MNL and MMNL models, which calculate average effects for all the alternatives, there may be an impact over the performances of the NN models performances.

For the dataset with correlated random effects of the alternative specific variables, we observe an identical situation in table 6. The class with lower average prices is chosen more rarely by the consumers, while the overall choices are less frequent due to the presence of stochastic individual preferences for particular alternatives' attributes.

We may conclude the preliminary datasets study and comparison with the main impression that two artificial datasets may be assumed to be quasi-identical. The slight differences in prices, captured by statistical tests may be considered insignificant in comparison with the biases present in the original dataset. What is more, even if the biases were more significant, the models' specification, which assumes no variable specific coefficients for choice A and B would have lead to the correct estimates, exactly as it was done by Michaud, Llerena, and Joly (2012). The heterogeneous preferences result in less probable decisions to buy a rose in the population, which should definitely impact the

Table 6: Alternatives' descriptive statistics by group, correlated random effects

	A (N=160000)	B (N=160000)	Total (N=320000)	p value
Choice				< 0.001
Mean (SD)	0.382 (0.486)	0.462 (0.499)	0.422 (0.494)	
Range	0.000 - 1.000	0.000 - 1.000	0.000 - 1.000	
Price				< 0.001
Mean (SD)	3.069 (0.979)	2.803 (0.917)	2.936 (0.958)	
Range	1.500 - 4.500	1.500 - 4.500	1.500 - 4.500	

performances of our models. Now it rests to verify how well the number of selected models will be able to derive the target values for the relative utility function.

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.