Extending the Multiple Discrete Continuous (MDC) modelling framework to consider complementarity, substitution, and an unobserved budget

David Palma[1,2]*, Stephane Hess[2]

[1]Centre for Decision Research, Leeds Business School
[2]Choice Modelling Centre, University of Leeds

8th April 2022

Abstract

Many decisions can be represented as interrelated discrete and continuous choices, i.e. what and how much to choose from a set of finite alternatives (incidence and quantity of consumption). In the last twenty years, several models of Karush-Kuhn-Tucker demand systems have been developed and used to study these kinds of decisions. While strongly grounded in economic theory, most of these models have two limitations: they require specifying a budget, and usually omit any complementarity effects. In this paper, we propose two extensions to the Multiple Discrete Continuous (MDC) modelling framework: (i) an MDC model including explicit complementarity and substitution effects, and (ii) an MDC model with complementarity, substitution that requires no budget definition. Model (ii) relies on the hypothesis that total expenditure on the alternatives under consideration is small compared to the overall budget. This allows using a linear utility function for the numeraire good, leading to a likelihood function without the budget or numeraire good in it. The lack of a budget is specially useful when forecasting, as it avoids cascading errors due to an inaccurate budget specifications. The inclusion of complementarity and substitution effects enriches the interpretability of the models, while the resulting functional form avoids theoretical issues present in previous formulations. Alongside the derivation of the models, we discuss their main properties and propose an efficient forecasting algorithm for (ii). We also report four applications to datasets about time use, household expenditure, supermarket scanner data, and trip generation. Free estimation code for both models is made available online.

Keywords: multiple discrete continuous; MDC; budgetless; Kuhn-Tucker demand; complementarity; substitution

*D.Palma@leeds.ac.uk

1 Introduction

Many choices can be represented as multiple discrete continuous decisions. In these, a decision maker faces a finite set of alternatives, and must choose how much to "consume" of each one, potentially consuming none, one, or multiple alternatives. Examples of these situation include activities performed during a day, grocery shopping, investment allocation, etc. Traditional choice models are not well suited for these situations, as they only allow the choice of a single alternative. Continuous models, on the other hand, often underestimate the probability of zero consumption for individual alternatives, also known as the "corner solution". Joint models, where the continuous choice is conditional on the discrete one, usually lack a strong grounding in economic theory, though there are exceptions (Hausman et al., 1995).

The Karush-Kuhn-Tucker multiple discrete continuous (MDC) consumer demand models (Bhat, 2008, 2018; Chintagunta, 1993; Hanemann, 1978; Kim et al., 2002; Mehta and Ma, 2012; Phaneuf and Herriges, 1999; Song and Chintagunta, 2007; Wales and Woodland, 1983) attend to the issues mentioned in the previous paragraph. These models begin by explicitly formulating the consumer utility maximisation problem, assuming either a direct or indirect utility function with associated randomness. Then the optimal solution is derived through the use of Karush-Kuhn-Tucker conditions. Finally, the likelihood function of these conditions is written given the distributional assumptions on the utility function. Nowadays, one of the most popular models of this category is the Multiple Discrete Continuous Extreme Value (MDCEV) model (Bhat, 2008). It has been applied in different areas, such as transport (Jäggi et al., 2012), time use (Enam et al., 2018), social interactions (Calastri et al., 2017), alcohol purchase (Lu et al., 2017), energy consumption (Jeong et al., 2011), investment decisions (Lim and Kim, 2015), household expenditure data (Ferdous et al., 2010), price promotions (Richards et al., 2012), and tourism (Pellegrini et al., 2017).

In this paper, we propose two extensions to the MDC modelling framework. First, we propose a new non-additive functional form for the utility that includes **explicit complementarity and substitution effects**. Secondly, we present an MDC model **formulation that does not require the definition of a budget**, while still allowing for explicit complementarity and substitution. The second approach is a suitable approximation of a full MDC model for (the relatively common) situation where the expenditure on all alternatives that are included in the model (i.e. inside goods) is small compared to the overall budget, which allows us to drop the budget from the model likelihood. To allow for a tractable likelihood function, we do not include a stochastic error term in the marginal utility of the outside good in any of the two proposed models.

Substitution and complementarity define relationships between the demand for pairs of products. If the demand for one of them increases, then the demand for the other is reduced in the case of substitution and increased in the case of complementarity (Hicks and Allen, 1934). While the budget constraint naturally induces substitution between products due to income effects, this is only an indirect effect. The inclusion of complementarity and substitution is necessary for a more realistic representation of behaviour in applications as diverse as time use or grocery shopping. For example, in the first case, it could be that going to the cinema makes it more likely for individuals to also eat at a restaurant. In the second case, it could be that products such as pasta and tomato sauce are usually bought together. On the other hand, it could be that the more hours an individual works, the fewer hours they allocate to leisure activities; or purchasing more bread leads to a reduction in the consumption of biscuits.

Concerning the budget, while determining it can be easy in some applications, it can be challenging in others. For example, in purchase decisions, the budget will rarely be an individual's full income, as there is likely mental accounting and recurring expenses to account for, all of which are not observable. Investment decisions face a similar problem, as the total budget may expand or shrink as a function of expected performance of the investment alternatives. There are other scenarios where even the simple definition of a budget is problematic, for example when modelling the number of recreational trips during a year, or the number of activities performed by an individual during a week. The problem becomes more acute in forecasting. Any predictions from a model require a budget, and predicting the budget, e.g. the income of individuals in the future, is another problem in itself, and introduces cascading errors in the forecast values.

While other models including complementarity and substitution effects through non-additive separable utility functions have been proposed in the literature, they either require complementarity and substitution effects to add up to zero (Song and Chintagunta, 2007), or pose specific constraints on their parameters, making either estimation or model transferability difficult (Bhat et al., 2015; Mehta and Ma, 2012; Pellegrini et al., 2021a). Models with implicit (also called infinite) budget have also been proposed by Bhat (2018) and **?** for models with neither complementarity or substitution effects. A detailed comparison between the models in this paper and those already in the literature is presented in section 5.

The remainder of this document is structured as follows. The next section introduces the formulation, derivation, likelihood function and forecasting algorithm of the model with complementarity and substitution. Section 3.2 presents the same for the model with complementarity, substitution and an implicit budget. Section 4 discusses the identification of both model parameters, some constraints that theory and estimation imposes on them, and compares the forecasting performance of both models to each other. Section 5 compares the proposed models' formulation to that of similar models in the literature. Section 6 presents applications of the proposed models to four different datasets, dealing with time use, household expenditure, supermarket scanner data, and number of trips, respectively. The paper closes with a brief summary of the proposed model formulations capabilities and limitations.

2 An MDC model with complementarity and substitution

2.1 Model formulation

Consider the classical (consumer) utility maximisation problem, where an individual n must decide what products k to consume from a set of alternatives, by maximising his or her utility subject to a budget constraint (Eqn. 1).

$$\begin{aligned} Max_{x_n} \quad & u_0(x_{n0}) + \sum_{k=1}^{K} u_k(x_{nk}) + \sum_{k=1}^{K-1} \sum_{l=k+1}^{K} u_{kl}(x_{nk}, x_{nl}) \\ s.t. \quad & x_{n0}p_{n0} + \sum_{k=1}^{K} x_{nk}p_{nk} = B_n \end{aligned} \tag{1}$$

where $n = 1...N$ indexes individuals and $k = 1...K$ alternatives, $x_n = [x_{n0}, x_{n1}, ..., x_{nK}]$ is a vector grouping the consumed amount of each alternative (product), p_{nk} is the price of alternative k faced by individual n, and B_n is the total budget available to individual n. x_{n0} is an *outside* or *numeraire* good, i.e. a good that aggregates all consumption outside of the category of interest. For example, if the researcher is interested in modelling demand for food, $x_{n1}, ..., x_{nK}$ would represent consumption of different food categories (the *inside* goods), while x_{n0} would represent the aggregate consumption of housing, transport, leisure, etc. It is usually assumed that $p_{n0} = 1$, so that x_{n0} becomes the total expenditure on categories other than the one of interest. To simplify the notation, we use this convention henceforth. It is assumed that the numeraire good is always consumed, so $x_{n0} > 0$ always.

The formulation in eqn. 1 is consistent with a two-stage budgeting approach, where the individual first allocates expenditure to broad groups (e.g. food, utilities, transport, entertainment, etc.) based on price indices representative for each group, followed by independent within-group allocations to individual products. According to Edgerton (1997), such an approach is sensible and subject to only small approximation errors when (i) the preferences for groups are weakly separable, i.e. the utility provided by each group is not affected by the level of consumption of other groups; and (ii) the group price indices being used do not vary too greatly with the utility or expenditure level. The first condition can be satisfied as long as the inside goods are reasonably separable from excluded goods. Edgerton (1997) argues that empirical and theoretical arguments support the fulfilment of the second condition.

We assume the following functional forms for the different parts of the utility function.

$$u_0(x_{n0}) = \psi_{n0} \log (x_{n0}) \tag{2}$$

$$u_k(x_{nk}) = \psi_{nk}\gamma_k \log\left(\frac{x_{nk}}{\gamma_k} + 1\right) \tag{3}$$

$$u_{kl}(x_{nk}, x_{nl}) = \delta_{kl}\left(1 - e^{-x_{nk}}\right)\left(1 - e^{-x_{nl}}\right) \tag{4}$$

We take the definition of u_k from Bhat (2008). In this formulation, ψ_{nk} represents alternative k's *base utility*, i.e. its marginal utility at zero consumption. This parameter could be interpreted as the scale of the utility of product k. The γ_k parameters, on the other hand, relate mainly to consumption satiation, by altering the curvature of alternative k's utility function. In general, a higher γ_k indicates higher consumption of alternative k, when consumed. While a common interpretation is that ψ_{nk} and γ_k determine what and how much of alternative k to consume, respectively, this is not completely true. There is a level of interaction between these parameters, and in some circumstances a low value of ψ_{nk} can be compensated by a high value of γ_k (Bhat, 2008, 2018).

Parameters ψ_{nk} must always be positive, as they represent the marginal utility of alternatives at the point of zero consumption. We ensure this using the following definition.

$$\begin{aligned} \psi_{n0} &= e^{\alpha z_{n0}} \\ \psi_{nk} &= e^{\beta_k z_{nk} + \varepsilon_{nk}} \end{aligned} \tag{5}$$

where z_{n0} is a column vector of characteristics of the decision maker that are expected to correlate with that individual's marginal utility of the outside good (e.g. socio-demographics); α is a row vector of parameters representing the weights of those characteristics on the marginal utility of the outside good; z_{nk} are attributes of alternative k; β_k are vectors of parameters representing weights of those attributes on the alternative's base utility; and ε_{nk} is a random disturbance term. We only include random disturbances in the base utility of the inside goods, as this leads to a computationally tractable likelihood function. We discuss the inclusion of a random disturbance in the marginal utility of the outside good in Section 4.1.

The final component of the utility function, $u_{kl}(x_{nk}, x_{nl})$, captures the complementarity and substitution effects between inside goods. This particular functional form is inspired by the translog function, and previous formulations by Vásquez Lavín and Hanemann (2008) and Bhat et al. (2015). Figure 1 presents the behaviour of this component for a set of δ_{kl} parameters, and different values of x_{nk} and x_{nl}, which are assumed to be equal. If $\delta_{kl} > 0$, there is complementarity between alternatives k and l, as this component will increase the overall utility. If $\delta_{kl} < 0$, there is a substitution effect between alternatives k and l, as u_{kl} becomes more negative as x_{nk} and x_{nl} increase. If $\delta_{kl} = 0$, the consumption of both alternatives is independent of each other. The value

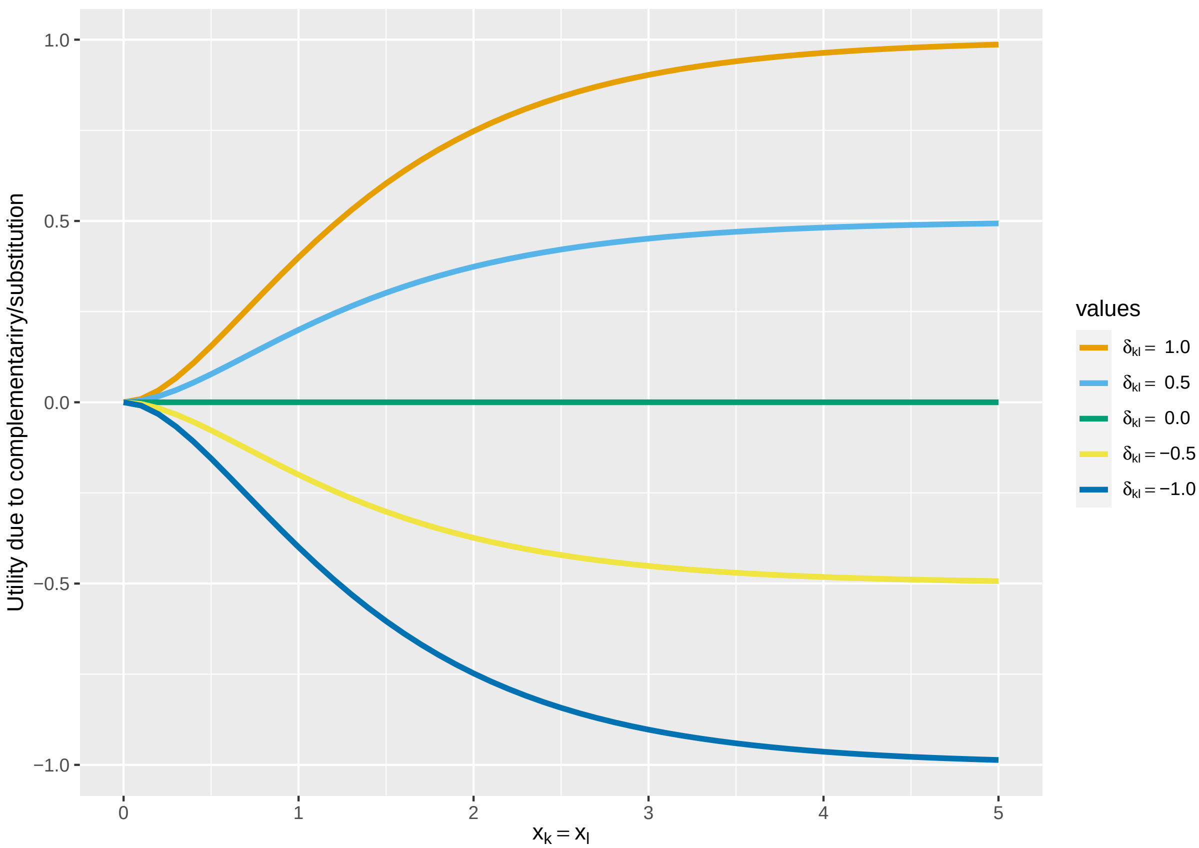

Figure 1: Complementarity/substitution component of the utility.

of u_{kl} is bounded to the interval $[0, \delta_{kl})$, ensuring transferability of estimated models to other datasets, a point we discuss in Section 4.2.

In summary, the proposed MDC model has two main characteristics. First, it contains no stochastic error in the marginal utility of the outside good, allowing for a tractable likelihood function. Second, its non-additive utility function allows for interaction (complementarity and substitution) among alternatives.

2.2 Model derivation

To solve the optimisation problem, we begin by writing its Lagrangian (Eqn. 6) and Karush-Kuhn-Tacker conditions of optimality (eqns. 7 and 8). We drop the n subindex to simplify the notation.

$$Lagr(x) = u_0(x_0) + \sum_{k=1}^{K} u_k(x_k) + \sum_{k=1}^{K-1} \sum_{l=k+1}^{K} u_{kl}(x_k, x_l) - \lambda \left(x_0 + \sum_{k=1}^{K} x_k p_k - B \right) \tag{6}$$

$$\frac{\partial Lagr}{\partial x_0} = 0 \quad : \quad \frac{\psi_0}{x_0} = \lambda \tag{7}$$

$$\frac{\partial Lagr}{\partial x_k} \leq 0 \quad : \quad \frac{\psi_k}{\frac{x_k}{\gamma_k} + 1} + e^{-x_k} \sum_{l \neq k} \delta_{kl} \left(1 - e^{-x_l}\right) \leq \lambda p_k \tag{8}$$

Eqn. 8 will be an equality when alternative k is consumed (i.e. $x^*_{nk} > 0$, with x^*_{nk} the consumption at the optimum, i.e. the observed consumption). Eqn. 8 will be an inequality when $x^*_{nk} = 0$. In other words, the marginal utility of any consumed product k at the optimum level of consumption will be λ scaled by the alternative's price p_{nk}. Instead, if the product is not consumed, its marginal utility will be lower. By combining eqns. 7 and 8, we obtain:

$$\frac{\psi_k}{\frac{x_k}{\gamma_k} + 1} + e^{-x_k} \sum_{l \neq k} \delta_{kl} \left(1 - e^{-x_l}\right) \leq \frac{\psi_0}{x_0} p_k \tag{9}$$

Replacing ψ_0 and ψ_k by their definitions (Eqn. 5), and isolating the random component ε_k, we obtain

$$\begin{aligned} \varepsilon_k &\leq -W_k \qquad (10) \\ W_k &= z_k \beta_k - \log\left(\frac{x_k}{\gamma_k} + 1\right) - \log\left(\frac{\psi_0}{x_0} p_k - e^{-x_k} \sum_{l \neq k} \delta_{kl} \left(1 - e^{-x_l}\right)\right) \end{aligned}$$

Now, if we assume all ε_k disturbances to follow identical and independent distributions, we only need to apply the Change of Variable Theorem from ε_k to x_k (only over the consumed alternatives) to obtain the likelihood function of the model. Then, if f and F are the density and cumulative distribution functions of ε_k, respectively, we can write the likelihood function as follows:

$$Like(x_k) = |J| \prod_{k=1}^{K} f(-W_k)^{I_{x_k > 0}} F(-W_k)^{I_{x_k = 0}} \tag{11}$$

$$
\begin{aligned}
J_{ii} &= \frac{1}{x_i+\gamma_i} + \frac{\frac{\psi_0}{x_0^2}p_i^2 + E_i}{\frac{\psi_0}{x_0}p_i - E_i} \\
J_{ij} &= \frac{\frac{\psi_0}{x_0^2}p_i p_j - \delta_{ij} e^{-x_i} e^{-x_j}}{\frac{\psi_0}{x_0}p_i - E_i} \\
E_i &= e^{-x_i} \sum_{l \neq i} \delta_{il}(1 - e^{-x_l})
\end{aligned} \tag{12}
$$

In this set of equations, $|J|$ is the value of the determinant of the Jacobian J of vector $-W_m$, where m indexes consumed alternatives. The elements of this Jacobian are defined in Eqn. 12 (i indexes rows, and j columns). No obvious compact form exists for this determinant. $I_{x_k>0}$ and $I_{x_k=0}$ are binary variables taking value 1 if $x_k > 0$ or $x_k = 0$, respectively, or zero in other case. If no alternative is consumed, the Jacobian drops out of Eqn. 11.

In the remainder of this paper, we assume all ε_k disturbances to follow identical and independent Normal distributions with mean fixed to zero and a standard deviation σ, which is estimated. Assuming other distributions is possible, where the use of Gumbel distribution leads to a closed-form likelihood, but has the disadvantage of generating a high rate of outliers during prediction, due to the thick tails of the distribution. The Normal distribution, on the other hand, has thinner tails and it is a natural choice due to the Central limit theorem, while being computationally tractable.

2.3 Forecasting

Once the model has been estimated, forecasting requires solving the original maximisation problem proposed in eqn. 1 several times, each time using different draws of ε_k from a Normal distribution with mean zero and standard deviation σ, and then averaging the result across these draws. This must be done separately for each observation in the sample. The optimisation problem can be solved using any algorithm, with the Newton or gradient descent algorithms being the most common type.

This forecasting procedure is demanding from a computational perspective, especially if a high number of draws are used for each individual. However, due to the forecast for each individual and draw being independent from one another, calculating them in parallel can significantly reduce the overall processing time. The software implementation in Apollo (ApolloChoiceModelling.com) uses parallel computing to speed up the forecasting.

3 An MDC model with complementarity, substitution and an implicit budget

In this section we introduce an extension of the model presented in section 2, such that it does not require defining a budget. The formulation and derivation of the model is very similar to that presented in the previous section, so in this section we only highlights the points where the two models differ.

3.1 Model formulation

Considering the classical consumer utility maximisation problem described in eqn. 1, we now assume a different utility formulation for the outside good, while all other definitions remain as in the previous section (i.e. as in eqns. 3, 4, and 5).

$$u_0(x_{n0}) = \psi_{n0} x_{n0} \tag{13}$$

We assume a linear utility function for the outside good (eqn. 13), as this will later on allow us to drop both the outside good consumption x_0 and the budget B from the final model formulation.

While a linear utility function does not comply with the law of diminishing marginal utility (a common assumption in demand models), it should be considered as an approximation of a function that does, when most of the budget is spent on the outside good, and only a relatively small amount is spent on the inside goods. In such a case, changes in the total expenditure of inside goods would lead to a relatively small change in the consumed amount for the outside good, and therefore a negligible change in the marginal utility of it.

More formally, we can write changes in the utility of the outside good using a second degree Taylor expansion as $u_0(x_0 + \Delta) \simeq u_0(x_0) + u'_0(x_0)\Delta + \frac{1}{2}u''_0(x_0)\Delta^2$, where u'_0 and u''_0 are the first and second derivatives of u_0, respectively, and Δ is a small change in the consumption of the outside good. If u_0 is continuous, monotonically increasing, and satisfies the law of diminishing returns, then $\lim_{x_0 \to +\infty} u'_0$ is a constant equal to or bigger than zero, because the slope must smoothly decrease as x_0 increases, without ever becoming negative. It then follows that $\lim_{x_0 \to +\infty} u''_0 = 0$. Therefore, for a large value of x_0, we can assume that $u''_0(x_0)$ is small, and approximate u_0 using a linear function, making $u'_0 \simeq \psi_0$.

Assuming a linear utility function for the outside good does not necessarily imply that all individuals have the same marginal utility for it, nor that absolutely no information on the budget can be included in the model. The proposed formulation allows for parameterisation of the ψ_0 parameter. The modeller could make ψ_0 a function of socio-demographics, or other proxies of the budget. For example, ψ_0 could be explained by an individual's full income, occupation, or their level of education.

3.2 Model derivation

Proceeding in the same way as in section 2.2, we first find a difference when calculating the derivative of the Lagrangean (Eqn. 6) with respect to the outside good, as follows.

$$\frac{\partial Lagr}{\partial x_0} = 0 : \psi_0 = \lambda \tag{14}$$

which combined with Eqn. 8 leads to the Eqn. 15

$$\frac{\psi_k}{\frac{x_k}{\gamma_k}+1} + e^{-x_k} \sum_{l \neq k} \delta_{kl} \left(1 - e^{-x_l}\right) \leq \psi_0 p_k \tag{15}$$

Replacing ψ_0 and ψ_k by their definitions (Eqn. 5), and isolating the random component ε_k, we obtain

$$\begin{aligned} \varepsilon_k &\leq -W_k \\ W_k &= z_k \beta_k - \log\left(\frac{x_k}{\gamma_k}+1\right) - \log\left(\psi_0 p_k - e^{-x_k} \sum_{l \neq k} \delta_{kl} \left(1 - e^{-x_l}\right)\right) \end{aligned} \tag{16}$$

Assuming all ε_k disturbances follow identical and independent distributions, and applying the Change of Variable Theorem from ε_k to x_k for the consumed alternatives, to obtain the likelihood function of the model, as described in eqn. 11, except this time the definition of the Jacobian elements is as in eqn. 17, with E_i the same as in eqn. 12.

$$\begin{aligned} J_{ii} &= \frac{1}{x_i + \gamma_i} + \frac{E_i}{\psi_0 p_i - E_i} \\ J_{ij} &= \frac{-\delta_{ij} e^{-x_i} e^{-x_j}}{\psi_0 p_i - E_i} \end{aligned} \tag{17}$$

Just as with the model with observed budget, we assume all ε_k disturbances to follow identical and independent Normal distributions with mean zero and a standard deviation σ to be estimated.

3.3 Forecasting

Once the model has been estimated, forecasting requires solving the original maximisation problem proposed in Eqn. 1 several times, each time using different draws of ε_{nk} from a Normal(0,σ) distribution, and then averaging the result across these draws.

To solve the optimisation problem we once again use the Lagrangian in Eqn. 6 and the KKT conditions in eqns. 14 and 8, leading us to Eqn. 15. Assuming an equality and isolating x_k, we obtain

$$x_k = h(x_k) = \gamma_k \left(\frac{\psi_k}{\psi_0 p_k - E_k} - 1 \right) \tag{18}$$

where the definition of E_k can be found in eqn. 17, and where it depends on the value of all x_n. Eqn. 18 is a fixed point problem, i.e. a problem of the form $x = h(x)$. According to the Existence and Uniqueness theorem, as the right part of Eqn. 18 is continuous in x_n over the closed interval $[0, \frac{B_n}{p_{nk}}]$, at least one solution to the problem exists. However, we cannot ensure that the solution is unique. We solve Eqn. 18 through the following iterative approach:

1. Set $r = 0$ and $x^{(r)} = [x_1^{(r)}, ..., x_K^{(r)}]$ to zero.
2. For each $k \in \{1, 2, ..., K\}$
   - 2.1. Set $s = 0$ and calculate $E_k^{(r)}$.
   - 2.2. Set $x_k^{(r)(s)}$ to a random starting value.
   - 2.3. Make $x_k^{(r)(s+1)} = h(x_k^{(r)(s)})$.
   - 2.4. If $|x_k^{(r)(s+1)} - x_k^{(r)(s)}| > \tau$ and $s < S$, go to step 2.3.
   - 2.5. If $x_k^{(r)} < 0$ or $|\frac{\partial U}{\partial x_k} - \frac{\partial U}{\partial x_0}| > \tau$, or $|x_k^{(r)(s+1)} - x_k^{(r)(s)}| > \tau$ make $x_k^{(r)} = 0$, otherwise make $x_k^{(r)} = x_k^{(r)(s+1)}$
3. If $|x^{(r)} - x^{(r)}| > \tau$ and $r < S$ go to 2.

where S is the maximum number of iterations allowed, and τ indicates the convergence tolerance parameter, which can be set to the desired precision. This procedure must be performed multiple times for each observation, each time with a different set of draws for the ε_k disturbances. Then results for each set of draws must be averaged.

As this model assumes a very large budget, in practice, there is no bound on the magnitude of the forecast consumption. Therefore, we recommend only forecasting for values of the explanatory variables in a reasonable vicinity of the values observed in the estimation dataset. What defines reasonable is difficult to quantifiy, but, for example, if an explanatory variable $z_1 \in [0, 1]$ in the estimation dataset, forecasting for $z_1 = 10$ could lead to unreasonably high consumption levels. This is similar to how linear models are usually valid only in the vicinity of values on which they were estimated.

4 Model properties

In this section, we discuss some of the most relevant properties of the model, namely the identifiability of its parameters, including the possibility of using random coefficients; some theoretical constraints on its parameters; and the performance of the model with implicit budget as compared to the model with observed budget.

4.1 Identification of parameters

When estimating the proposed models, the modeller should consider the following six points regarding identifiability of parameters.

First, observations who do not consume any inside good should **not** be excluded from the sample. Even though these observations do not provide any information on the value of ψ_k, they do provide information of the value of ψ_0 in relation to the inside goods.

Second, there should be no constant (intercept) in the definition of ψ_0, i.e. z_0 should not contain an element equal to 1 for every individual. As utility does not have any meaningful units, we require setting a base against which all other utilities are measured. To do this, we recommend setting the intercept of the outside good to zero. Any variable that changes across observations can be included in z_0, even if they are not centred around zero. We recommend populating z_0 with characteristics of decision makers, such as socio-demographics.

In the case of the model with implicit budget (see section 3) we recommend including the individual's income in z_0. Including income in this way does not imply that the budget is equal to the income, but only that the marginal utility of the outside good depends on it. We would expect a negative coefficient for income if included in ψ_0, as an increase of income usually leads to increased overall consumption, and therefore a smaller marginal utility of the outside good. In general, a negative coefficient α indicates that an increase in the corresponding explanatory variable leads to increased consumption. The opposite is true for a positive coefficient.

Third, just as most other MDC models, the two formulations presented in this paper are not scale-independent. This means that the magnitude of the dependent variable influences the results of the model. For example, expressing the dependent variable in grammes or kilogrammes might lead to different forecasts and marginal rates of substitution. This is due to the non-linear nature of the utility functions used in the models. We recommend testing different scalings of the dependent variable, favouring those making the dependent variable range between zero and five, so as to match the range of maximum variability of the transformation in u_{kl}, which is mostly flat for values $x_k > 5$ (see figure 1).

Fourth, in the case of the model with implicit budget, complementarity and substitution effects can be confounded with income effects. In the model with implicit budget, all interactions between the consumption of alternatives are captured by the δ_{kl} parameters. The cause of interaction could be complementarity or substitution, but it could also be due to income effects. For example, a restricted budget could induce increased demand for an inexpensive product while decreasing the demand for an expensive one. This could be captured by the model as substitution between the two products. This problem will be attenuated if the budget is large in comparison with the expenditure on the inside good.

Fifth, concerning the number of complementarity and substitution parameters (δ_{kl}), while the model formulation defines one parameter per pair of products, the modeller can easily impose restrictions to reduce the number of parameters to estimate. For example, if alternatives can be grouped into non-overlapping sets, the modeller could impose all δ_{kl} parameters to be the same within each group, and across the same pair of groups. Alternatively, the modeller could perform a Principal Component Analysis on the dependent variables, identifying the most important interactions between alternatives, and then estimating only those δ_{kl} parameters and fix all others to zero (as done in section 6.2). These or other strategies are recommended when the number of alternatives is large.

Finally, as recommended by Manchanda et al. (1999), the proposed models allow for complementarity, substitution, and *coincidence* effects, both in a deterministic and random way. Complementarity and substitution effects are captured by the δ_{kl} parameters. *Coincidence* effects are shocks to demand influencing either one or multiple alternatives at the same time, and they can be captured by either ψ_0 (common shocks to all alternatives), or ψ_k and γ_k (independent shocks). All of these parameters allow for deterministic heterogeneity, for example defining δ_{kl} as a function of socio-demographic characteristics. It is also possible to incorporate random heterogeneity in ψ_k and γ_k by using simulated maximum likelihood techniques (Train, 2009), but we do not recommend including such heterogeneity in ψ_0 nor δ_{kl} as it could lead to violations of eqns. 23 and 24 (see section 4.2).

To test identifiability of the model through simulation, we created 50 datasets using the generation process of the model with observed budget, and another 50 datasets using the generation process of the model with implicit budget. We then estimated the corresponding model on each generated dataset to check if we were able to recover the parameters used during data generation. All datasets were composed of 500 observations with four alternatives each. All models shared the specification described in eqn. 19, but with the value of their parameters randomly drawn on each occasion from the distributions defined in table 1. The range of parameters was influenced by other models estimated in section 6 and considerations discussed in section 4.2. All explanatory variables (z, x, y) followed a U(0,1) distribution, except for $z_1 \sim Bernoulli(0.5)$. Prices were drawn from a U(0.1, 1) distribution, while the budget was set to 10 for the models with observed budget.

$$\begin{aligned} \psi_0 &= e^{\alpha_1 z_1 + \alpha_2 z_2 + \varepsilon_0} \\ \psi_k &= e^{\beta_{k0} + \beta_{k1} x_{k1} + \beta_{k2} x_{k2} + \varepsilon_k} \\ \gamma_k &= \gamma'_k + \gamma'_{k1} y_{k1} + \gamma'_{k2} y_{k2} \end{aligned} \tag{19}$$

Table 1: Distributions used to draw parameters from when simulating datasets.

	Observed budget	Implicit budget
α_1	U(0.1, 1.0)	U(0.1, 0.2)
α_2	U(-1.0, -0.1)	U(-0.2, -0.1)
β_k	U(-1.0, 1.0)	U(0.1, 1.0)
β_1	U(0.1, 1.0)	U(0.1, 0.5)
β_2	U(-1.0, -0.1)	U(-0.5, -0.1)
γ'_k	U(5.0, 10.0)	U(0.1, 1.0)
γ'_1	U(2.0, 5.0)	U(0.1, 0.5)
γ'_2	U(2.0, 5.0)	U(0.1, 0.5)
δ_{kl}	U(-0.1, 0.1)	U(-0.03, 0.03)
σ	U(0.5, 1.0)	U(0.25, 0.1)

Figures 2 and 3 summarise the true and estimated parameter for the model with observed and implicit budget, respectively. In the graphs, the horizontal axis indicates the true value of the parameter, while the vertical axis indicates the estimated value. In these graphs, a perfect recovery of a parameter is represented by a dot along the identity line (in blue). The graph also contains the 95% confidence interval for each estimated parameter. Both figures offer a similar perspective: while all parameters are recovered correctly, α and β parameters are recovered more precisely, while γ and δ parameters (specially the latter) are harder to recover.

4.2 Constraints on estimated parameters

The derivation of the likelihood function relies on the assumption of the utility function being monotonically increasing with decreasing marginal returns of consumption. In other words, it

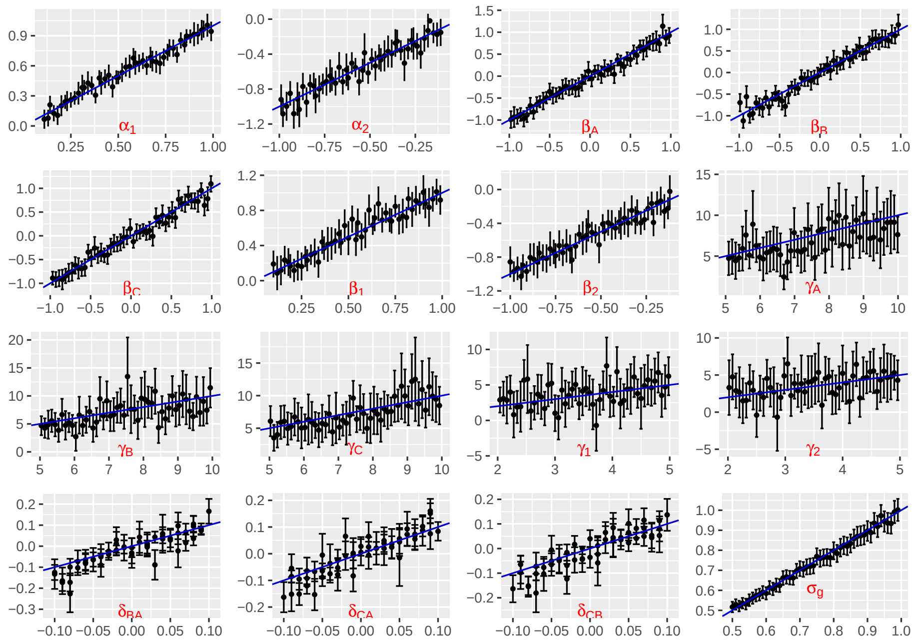

Figure 2: Recovery of parameters for the model with observed budget.

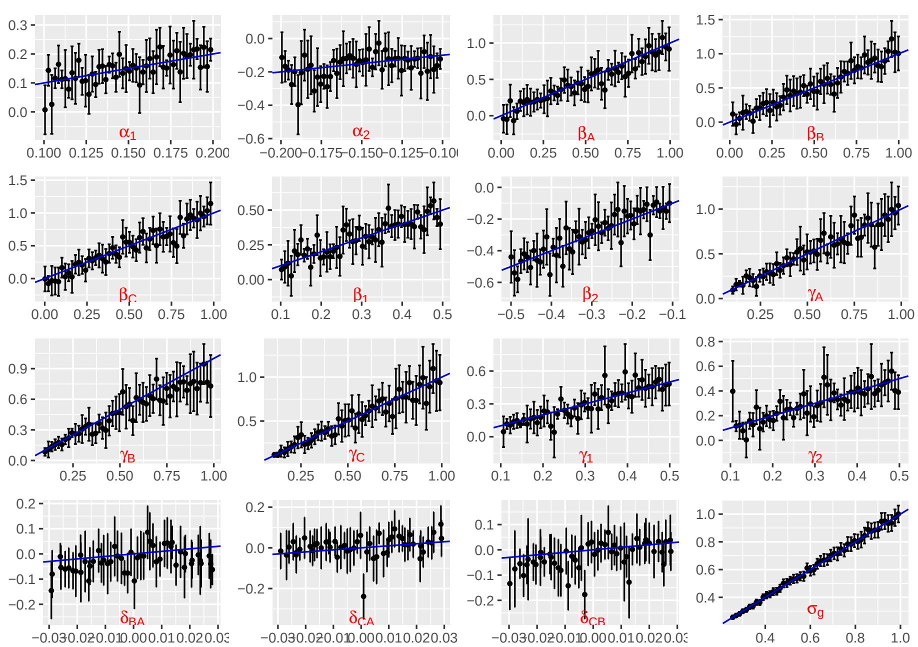

Figure 3: Recovery of parameters for the model with implicit budget.

assumes $\frac{\partial U}{\partial x_k} > 0$, where U is the global utility. Failing to comply with this assumption renders the likelihood function invalid, as second order derivatives on the Lagrangean would have to be checked to make sure the critical point is not a minimum. Furthermore, it could lead to the existence of multiple local critical points, i.e. the solution may not be unique, which is once again contrary to the assumptions made during the derivation of the likelihood function. The marginal utility of the outside good is always positive in both models proposed in this paper. But the marginal utility with respect to an inside good will only be positive when the inequality in Eqn. 20 is fulfilled.

$$\frac{\psi_k}{\frac{x_k}{\gamma_k}+1} + e^{-x_k} \sum_{l \neq k} \delta_{kl} \left(1 - e^{-x_l}\right) > 0 \tag{20}$$

Additionally, the argument of the logarithm inside W_k must be larger than zero, so as to avoid undefined operations. In the case of the model with observed budget, this translate into the inequality in Eqn. 21. And in the case of the model with implicit budget, it implies Eqn. 22 must be satisfied.

$$\frac{\psi_0}{x_0} p_k - e^{-x_k} \sum_{l \neq k} \delta_{kl} \left(1 - e^{-x_l}\right) > 0 \tag{21}$$

$$\psi_0 p_k - e^{-x_k} \sum_{l \neq k} \delta_{kl} \left(1 - e^{-x_l}\right) > 0 \tag{22}$$

These conditions are functions of x_k, making their fulfillment dependent on the particular dataset at hand. We would like to instead derive dataset-independent conditions. This is possible by noting that the impact of x_k in both conditions is bounded by its exponential transformation to the interval $0 \leq e^{-x_k} \leq 1$ (because $x_k \geq 0$). This allows us to derive more general conditions than Eqns. 20, 21 and 22 by analysing the extreme cases $x_k = 0$ and $x_k = \infty$, as the value of the conditions for all other x_k values will fall between these. These extreme cases have the benefit of removing x_k from the conditions. Table 2 summarises the results from this analysis.

All conditions in table 2 with zero on the right hand side are always fulfilled because ψ_k, γ_k, p_k, Δ^- and Δ^+ are all equal or bigger than zero. Eqn. 20 for $x_k = \infty$ will also always be true as zero is approached from the right (i.e. from positive values). Among the remaining conditions, $\psi_k > \Delta^-$ implies $\psi_k + \Delta^+ > \Delta^-$, just as $\frac{\psi_0}{x_0} p_k > \Delta^+$ implies $\frac{\psi_0}{x_0} p_k + \Delta^- > \Delta^+$ and $\psi_0 p_k > \Delta^+$ implies $\psi_0 p_k + \Delta^- > \Delta^+$. Therefore, the sufficient conditions for the model with observed budget can be summarised as in eqn. 23

$$-\psi_k < \sum_{l:\,\delta_{kl}<0} \delta_{kl} < \sum_{l:\,\delta_{kl}>0} \delta_{kl} < \frac{\psi_0}{x_0} p_k \;\forall k \tag{23}$$

Table 2: Constraints on proposed model parameters for extreme levels of consumption

x_k	$x_{l:\delta_{kl}>0}$	$x_{l:\delta_{kl}<0}$	Eqn. 20	Eqn. 21	Eqn. 22
0	0	0	$\psi_k > 0$	$\frac{\psi_0}{x_0} p_k > 0$	$\psi_0 p_k > 0$
0	0	∞	$\psi_k > \Delta^-$	$\frac{\psi_0}{x_0} p_k + \Delta^- > 0$	$\psi_0 p_k + \Delta^- > 0$
0	∞	0	$\psi_k + \Delta^- > 0$	$\frac{\psi_0}{x_0} p_k > \Delta^+$	$\psi_0 p_k > \Delta^+$
0	∞	∞	$\psi_k + \Delta^+ > \Delta^-$	$\frac{\psi_0}{x_0} p_k + \Delta^- > \Delta^+$	$\psi_0 p_k + \Delta^- > \Delta^+$
∞	any	any	$0^+ > 0$	$\frac{\psi_0}{x_0} p_k > 0$	$\psi_0 p_k > 0$
		Where:	$\Delta^- = \sum_{l:\delta_{kl}<0} \lvert\delta_{kl}\rvert$; $\Delta^+ = \sum_{l:\delta_{kl}>0} \delta_{kl}$		

And the sufficient conditions for the model with implicit budget are summarised in eqn. 24.

$$- \psi_k < \sum_{l:\, \delta_{kl}<0} \delta_{kl} < \sum_{l:\, \delta_{kl}>0} \delta_{kl} < \psi_0 p_k \; \forall k \tag{24}$$

Conditions in eqns. 23 and 24 are based on extreme cases, so they represent sufficient but not necessary conditions for the validity of the parameters. In other words, estimated parameters need only to comply with eqn. 20, and with eqn. 21 or 22, but satisfying eqn. 23 or 24 guarantees that those conditions are met.

If individuals in the dataset behave rationally and in accordance with economic theory, then the estimated parameters should naturally comply with eqn. 23 or 24. At the time of writing, we have not experienced any issues of running into inconsistent parameters, nor have we had to impose parameter constraints during estimation to enforce compliance with these equations.

4.3 Suitability of a linear utility for the outside good

In the model with implicit budget, we propose a linear utility for the outside good as an approximation of the case where expenditure on the inside goods (i.e. considered alternatives) is small compared to that on the outside (numeraire) good. In these cases, we expect only very small changes to the marginal utility of the outside good due to changes in the consumption of the inside goods. For example, consider consumption of the yoghurt product category. The expenditure on yoghurt will be small compared to the total expenditure on food, and even smaller compared to the entire disposable income of the household. By using the model with implicit budget, the modeller does not need to determine what the correct budget is, but only needs to know that total expenditure in the category of interest is small compared to the budget, whatever that may be.

If our interpretation is correct, then the forecast of the model with implicit budget should approach that of the model with observed budget when the expenditure on the outside good is large compared to that on the inside goods. We tested this assumption through simulation. We first created 30 different datasets of 500 observations each, assuming a data generation process with observed budget, i.e. using the model presented in section 2. Besides having an outside good, each dataset had four inside goods that were always available. The base utility of the outside good was set to zero, while the base utility of the inside goods was composed of a single constant, each drawn from $U(-2, 0)$, i.e. a uniform distribution between -2 and 0. Satiation parameters γ_k were drawn from $U(0.5, 1.5)$, δ_{kl} were drawn from a $U(-0.01, 0.01)$, while price p_k followed a $U(0.1, 1)$, and the budget was set to 10 for every observation. We measured the fit of each model on each dataset using the Root Mean Squared Error (RMSE) of the forecast aggregate demand in the whole sample. Results are exhibited in figure 4.

As figure 4 shows, the fit of the model with implicit budget approaches that of the model with observed budget as the expenditure on the outside good increases. This indicates that the model with implicit budget is an appropriate approximation when the expenditure on the outside good is large relative to the expenditure on inside goods.

5 Comparison with other MDC formulations

The MDC models presented in this paper are not the first to include complementarity, substitution or an implicit budget in the literature. In this section, we discuss other MDC models with these properties, and compare them to the models proposed in this paper. We begin with a very brief review of models without complementarity or substitution (other than income effects), which form the basis for more flexible models.

5.1 No complementarity or substitution, and an observed budget

One of the most popular models in this category is the MDCEV model by (Bhat, 2008). It is derived from the same consumer optimisation problem proposed in eqn. 1, but using a different functional form for the utility components. While there are several possible formulations, the most common one is the *alpha-gamma* formulation, due to it allowing for an efficient forecasting algorithm (Pinjari and Bhat, 2011). In this case, the utility takes the form described in eqn. 25, where α can either tend towards zero during the estimation process, or the modeller can fix it *a priori*.

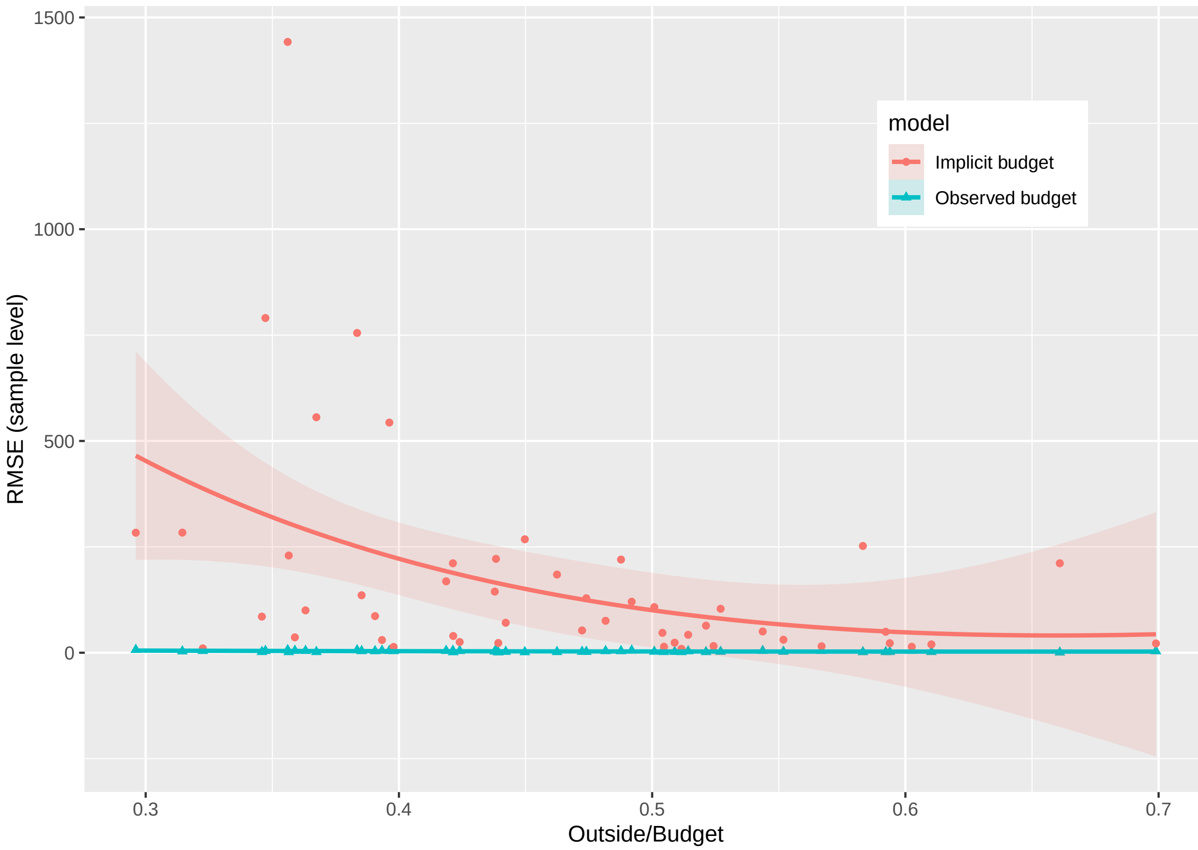

Figure 4: Compared fit of models with observed and implicit budget, on data generated assuming a generation process with observed budget

$$
\begin{aligned}
u_0 &= \frac{1}{\alpha}\psi_0\left((x_0+1)^{\alpha}-1\right) && \xrightarrow[\alpha\to 0]{} \psi_0 \log\left(x_0+1\right) \\
u_k &= \frac{\gamma_k}{\alpha}\psi_0\left(\left(\frac{x_0}{\gamma_k}+1\right)^{\alpha}-1\right) && \xrightarrow[\alpha\to 0]{} \gamma_k\psi_k \log\left(\frac{x_k}{\gamma_k}+1\right) \\
u_{kl} &= 0 && \xrightarrow[\alpha\to 0]{} 0
\end{aligned}
\tag{25}
$$

Parameter interpretation in the MDCEV model is essentially the same as in the models described in this paper, except for two differences. First, the outside good's marginal utility contains no covariates, but only a stochastic error term, i.e. $\psi_0 = e^{\varepsilon_0}$. Second, α measures satiation across the whole choice set in MDCEV, and not the influence of covariates in the outside good's marginal utility as in the models proposed in this paper. And while it is possible to introduce explanatory variables into the base utility of the outside good in MDCEV models (either directly, or by including them with the same coefficient in all inside goods' base utility), it is not commonly done in practice.

By setting $u_{kl} = 0$, the MDCEV model does not allow for pure complementarity or substitution effects, though product substitution can still take place due to income effects. Also, the form of u_0 requires the value of x_0, and therefore the budget, to be observed.

Kim et al. (2002) use a similar utility function to the MDCEV model, but assume that the random disturbances follow a multivariate normal distribution. While more flexible, this distribution makes the model much more computationally demanding. Von Haefen and Phaneuf (2005) also present a similar model to MDCEV, but without an error term in the marginal utility of the outside good. Other models in this category include Habib and Miller (2008) and Habib and Miller (2009), who present models similar to that by Von Haefen and Phaneuf (2005).

5.2 Introducing complementarity and substitution through new functional forms

Vásquez Lavín and Hanemann (2008) propose a model formulation allowing for complementarity and substitution using a non-additively separable utility function and an observed budget. This formulation was later refined by Bhat et al. (2015), who called it the *NASUF* model. Beginning from the consumer optimisation problem set in eqn. 1, the utility components are defined as described in eqn. 26.

$$\begin{aligned} u_0 &= \psi_0 \log(x_0 + \gamma_0) \\ u_k &= \psi_k \gamma_k \log\left(\frac{x_k}{\gamma_k} + 1\right) \\ u_{kl} &= \theta_{k,l}\left(\gamma_k \log\left(\frac{x_k}{\gamma_k} + 1\right)\right)\left(\gamma_l \log\left(\frac{x_l}{\gamma_l} + 1\right)\right) \end{aligned} \tag{26}$$

The definition of u_{kl} makes the NASUF utility function non-additive, effectively introducing complementarity and substitution effects. A positive value of θ_{kl} is indicative of complementarity, while a negative one represents substitution, and $\theta_{kl} = 0$ implies no complementarity or substitution. Yet, this formulation has three main drawbacks.

The first drawback is that the utility function is valid only for some values of θ_{kl}. Just as in the case of the models proposed in this paper, and as discussed in section 4.2, the derivation of the likelihood function assumes $\frac{\partial U}{\partial x_k} > 0$. For this to be true, the inequality in eqn. 27 must be satisfied.

$$\frac{\partial U}{\partial x_k} = \psi_k + \sum_{l \neq k} \theta_{kl} \gamma_l \log\left(\frac{x_k}{\gamma_k} + 1\right) > 0 \; \forall k, l \tag{27}$$

While it is possible to bound the value of parameters during estimation, the problem with the condition in eqn. 27 is that it depends on the value of x_k. As the logarithm is not a bounded function, whether or not this condition is satisfied will depend on the level of consumption x of each individual, making it impossible to assess the correctness of a model without associating it to a particular dataset. This hinders model transferability from one dataset to another, and jeopardises forecasting, as only scenarios that fulfil the condition above should be permissible forecasts.

If all individuals in the dataset behave in accordance with economic theory, then the parameters should automatically fulfill eqn. 27. Yet, this does not prevent the estimation algorithm from trying parameter values violating eqn. 27 during the parameter value search. Furthermore, calculating the likelihood of the model requires calculating the logarithm of the expression in eqn. 27, leading to an error if the expression is less or equal than zero.

The second issue with the solution proposed by Bhat et al. (2015) is that the stochasticity is introduced midway through the derivation of the model in the Karush-Kuhn-Tacker conditions, and not in the initial formulation of the model. While this is merely a formal issue, it does imply that the origin of the randomness is not clear, and it is not possible to easily associate it with unobserved variables or measurement errors, as would be the case in more traditional econometric models.

The third issue is that γ parameters have a role both in satiation and in the interaction term (i.e. complementarity and substitution) of the utility, making their interpretation difficult.

Pellegrini et al. (2019) refine the model proposed in Bhat et al. (2015) by proposing a different interaction term in the utility function. While this new formulation leads to an improved fit and provides a clear interpretation of γ parameters, it retains at least the first issue associated to the formulation of Bhat et al. (2015). Pellegrini et al. (2021a) further expand the NASUF model by allowing for two budget constraints in an application where both time and monetary constraints are considered jointly.

A similar formulation was proposed by Lee and Allenby (2009), but using a quadratic function to incorporate satiation, complementarity, and substitution. This model only considers inside goods, defining the global utility as $U = \sum_k \psi_k x_k - \frac{1}{2} \sum_k \sum_l \theta_{kl} \psi_k x_k \psi_l x_l$ (we assume only one product per category to simplify the analysis). Note that θ_{kk} is not restricted to zero in this case, as is in the models proposed in this paper. The validity of the formulation rests on the condition $\frac{\partial U}{\partial x_k} = \left(1 - \sum_l \theta_{kl} \psi_k x_k\right) \psi_k > 0$, which depends on the value of x_k, leading to the same issue already discussed in the context of the *NASUF* model.

Finally, Lee et al. (2010) propose a model allowing for asymmetric complementarity and substitution among categories of product. However, the formulation of the model does not satisfy the principle of weak complementarity (Maler, 1974), i.e. that an individual's utility is not influenced by the attributes of non-consumed goods or, in other words, that goods provide utility only through their use. This is a reasonable assumption in cases where non-use values are believed to be absent or small (see von Haefen (2004) for a more detailed discussion).

5.3 Introducing complementarity and substitution through the indirect utility function

While in this paper we derived MDC models from the direct utility function of consumers, it is also possible to make assumptions on the indirect utility instead, and then calculate the optimal consumption using Roy's identity, as described in section 3.1 of Chintagunta and Nair (2011).

Song and Chintagunta (2007) propose an MDC model following the indirect utility approach, considering not only a set of alternatives, but grouping them into categories, and assuming that at most one alternative inside each category is consumed. Furthermore, this model imposes a symmetry constraint on its complementarity and substitution parameters, as described in eqn. 28.

$$\sum_{l=0}^{M} \theta_{kl} = 0 \, \forall k \tag{28}$$

where θ_{kl} represents the complementarity and substitution parameters (originally called β in Song and Chintagunta (2007)). Eqn. 28 forces that, for each product, the amount of complementarity and substitution with other products needs to add up to zero. But there are no theoretical reasons for this to necessarily be the case in any given application. This requirement prevents, for example, for a product to only have complementarity with one other product, while not having substitution with any other product.

Mehta and Ma (2012) propose a model with a similar formulation to that of Song and Chintagunta (2007), but without the symmetry constraint. However, it requires the matrix of complementarity and substitution parameters (whose elements are $\theta_k l$) to be positive semi-definitive. Additionally, the likelihood function does not have a closed functional form, requiring multiple-dimension integration; and the number of parameters increases geometrically with the number of alternatives.

5.4 Introducing complementarity and substitution through correlation in utility functions

An alternative way to introduce complementarity and substitution into an MDC model is by introducing correlation across the utility of alternatives. This can be done in two ways: (i) by directly correlating the random error term ε in the utility function of each alternative across multiple alternatives, or (ii) by adding new random error terms common to the utility of multiple alternatives. Pinjari and Bhat (2010) use the first approach, using extreme value distributions to nest alternatives together into mutually exclusive subsets, allowing for perfect substitutes but not for complementarity. This approach was generalised by Pinjari (2011), by allowing for overlapping non-exclusive nests, but still limiting its applicability to complementarity. Bhat et al. (2013) makes ε follow a multivariate normal distribution across alternatives, allowing for flexible correlation patterns. Calastri et al. (2020a) follows the second approach, by using random intercepts and coefficients (β in our notation) correlated across alternatives.

As Pellegrini et al. (2021a) discuss, the main limitation of introducing complementarity and substitution through correlation in the utility functions of different alternatives is that of confounding effects. Indeed, using this approach it is impossible to discriminate between correlation due to common heterogeneity in preferences, from correlation due to complementarity and substitution. For example, two utilities could be positively correlated due to them sharing unobserved attributes, but not because the alternatives are complementary.

5.5 Two stage approaches to unobserved budgets

The necessity to observe the budget can lead to two separate issues. The first one is during estimation, in the case when the budget is not observed. This forces the modeller to assume some value for the budget before even estimating and MDC model. A common solution to this problem in past work has been to use the total expenditure as the budget. This is a strong assumption, as it implies that the total expenditure will not change as a function of prices or other attributes of the products. For example, it implies that consumers will spend the same amount regardless of the level of discount offered.

The second problem due to the necessity of an observe budget in MDC models manifests during forecasting. Forecasting for any future scenario requires exogenously defining a budget. Any errors in the forecasting of the budget will cascade down to the MDC model, as shown in section 6.2.

In the literature, these problems have been addressed mostly through two-stage procedures, where in the first stage, a model is used to estimate (and predict) the budget, and in the second stage, a traditional MDC model with observed budget is used to allocate the budget to the different alternatives.

Pinjari et al. (2016) proposes a two-stage approach. In the first stage, they use either a stochastic frontier or a log-linear regression to estimate the expected budget, and in the second stage they use the expected budget in an MDCEV model. They compare the performance of both approaches against arbitrarily determined budgets. When using the stochastic frontier method, they assume the budget to be an unobservable characteristic of decision makers, defined as the maximum amount they are willing to spend. This implies that the expected budget under this approach tends to be bigger than the total expenditure. The log-linear regression, on the other hand, attempts to predict total expenditure, so it leads to expected budgets that are of the same magnitude as the total expenditure. While both approaches offer similar performance, and both outperform the arbitrarily determined budget, the stochastic frontier approach leads to bigger expected budgets, therefore allowing for more variability in the forecast, as the total expenditure has room to grow if the attributes of the alternatives improve. This approach is also used by Pellegrini et al. (2021b).

Dumont et al. (2013) propose a different two-step approach to estimate the budget. In the first step, they estimate a Structural Equation Model (SEM) where the budget is a latent variable, whose structural equation has socio-demographics as explanatory variables. The budget can have several indicators, such as average expenditure in the category during the last three months, expected expenditure in the future, and ownership of goods from the same category. Income is also considered a latent variable, with at least stated income as indicator. More formally, the latent budget B_n and latent income I_n relate as follows :

$$B_n = Z_n\zeta_z + \zeta_I I_n + \eta_n \tag{29}$$

$$I_n = \xi_n \tag{30}$$

$$y_{nj} = \lambda_j B_n + \sigma_j \varepsilon_{nj} \tag{31}$$

$$S_n = \lambda_s I_n + \sigma_s \varepsilon_{ns} \tag{32}$$

where Z_n are socio-demographics of individual n, y_{nj} is indicator j of the budget, S_n is the stated income, $\eta_n, \xi_n, \varepsilon_{nj}$ and ε_{ns} are standard normal error terms, and $\zeta_z, \zeta_I, \lambda_j, \sigma_j, \lambda_s$ and σ_s are parameters to be estimated. As expected, authors report lower log-likelihoods when using the SEM approximation to the budget than when using maximum expenditure, but they also do note an improvement in the MDC parameters significance levels. They do not report changes in forecast performance, making it difficult to evaluate the performance of the proposed approach.

5.6 Other MDC models with implicit budget

Other models in the literature have also used linear utility functions for the outside good, in the same way that in the models proposed in this paper. This functional form leads to a likelihood function that does not depend on the budget, effectively allowing for unobserved budgets.

In the context of the MDCEV model and its derivations, Bhat (2018) was the first one to propose using a linear utility function for the outside good. This functional form, however, was not motivated by the need to drop the budget from the model formulation, but it was used to allow for more separability between the parameters that determine the discrete choice (i.e. *what* to choose), from those that determine the continuous choice (i.e. *how much* to choose). Therefore, this property of the model is hardly explored in that paper.

More recently, Saxena et al. (2022) discussed the consequences of using a linear utility for the outside good in models with additively separable utility functions. Such a configuration leads to models that do not consider complementarity, substitution, nor income effects, therefore making demand from one product independent from another, unlike the model proposed in this paper (though it does allow for parameterising ψ_0). Similarly to our own advice, they recommend using a linear utility function for the outside good only when the total expenditure in the inside goods is no more than 35% of the budget (or more strictly, less than 5%). If the expenditure in inside goods is higher than those values, they find bias in the model estimates and poor forecasting performance. While we did not find evidence of biased parameters in the proposed model (see figure 3), we did find evidence of poor forecast performance (see figure 4). The absence of parameter bias in the proposed model could be due to it including complementarity and substitution effects, and the fact that the error term follows a Normal distribution instead of a Gumbel distribution.

6 Model application and comparison

In this section we apply the proposed models to four different datasets. The first dataset records time use, where all participants face the same budget (24 hours a day), and all alternatives (in this case, activities) have the same price (one unit of time). This dataset allows us to measure how much fit is lost when using the model with implicit budget when the budget is known, as well as compare the proposed models against a model without complementarity nor substitution. The second dataset deals with household expenditure, where budgets vary between different households, but consumption is aggregated to categories, so prices are still unitary (one unit of money). This dataset helps us illustrate how the fit of the model with observed budget degrades when the budget is misspecified, a case particularly relevant in forecasting. The third dataset contains scanner data from a supermarket, where both budgets and prices vary from one observation to the next. This dataset allows us to compare the sensitivity to price of the models with observed and implicit budget. The last dataset reports the number of trips performed by travellers for different purposes. This dataset is a case where the very definition of a budget is problematic, as there is no evident limit on the number of trips during a day.

6.1 Fixed budget and fixed prices: time use dataset

The first dataset records time use of 447 individuals across 2,826 days in total. Details about the data collection can be found in Calastri et al. (2020b), and an application to time use analysis using this data can be found in Calastri et al. (2019) and Palma et al. (2021). Only out-of-home activities are registered in the dataset, which we aggregate to six plus the outside good, as described in table 3.

Table 3: Main descriptive statistics of the time use database

		Consumption (H)		Correlation			
	Engagement	Total	Average†	Work	School	Shopping	Private B.
Home*	100.00%	51467	18.21				
Work	40.30%	8170	7.17	1.00			
School	3.01%	299	3.52	-0.06	1.00		
Shopping	27.71%	1408	1.80	-0.08	-0.03	1.00	
Private B.	18.93%	1253	2.34	-0.09	0.00	-0.01	1.00
Leisure	41.54%	5227	4.45	-0.17	-0.01	-0.01	-0.04

** outside good; † when engaged*

We estimated three different models using the Time Use data. First we estimated a traditional *MDCEV* model (Bhat, 2008), which has an observed budget and no complementarity. We also estimated the first model proposed in this paper (*eMDC1*), with an observed budget, complementarity and substitution. Finally, we estimate the second model proposed in this paper (*eMDC2*), with an implicit budget, complementarity and substitution.

In the case of time use, the budget is observed (24 hours a day for everyone), and remains unchanged in forecasting scenarios, giving a clear advantage to the *MDCEV* and *eMDC1* models. Nevertheless, we are interested in exploring the consistency of results across the models with observed budget, as well as the loss of fit in the *eMDC2* model (which uses an implicit budget) with respect to the others. We estimated the models using 70% of the sample, and forecast for the remaining 30%. Table 4 presents the estimated parameters, likelihood and root mean squared error (RMSE) of the forecast consumption at the aggregate sample level for each model.

The parameter estimates point towards consistent effects across models. And while parameters across models change in magnitude, their signs remain unchanged. Parameter interpretation is equivalent across models, except for α. In the *MDCEV* model α measures satiation across all alternatives. Instead, in the proposed *eMDC* models α represents the impact of the associated explanatory variable (z_0) on the marginal utility of the outside good (ψ_0). In the proposed models,

Table 4: Comparison of the proposed extended MDC and a traditional MDCEV models on a time use dataset

	MDCEV		eMDC1		eMDC2	
	Estimate	t-ratio*	Estimate	t-ratio*	Estimate	t-ratio*
α Constant	0.036	20.77				
α Female			-0.102	-1.44	-0.044	-1.83
β Work	-3.351	-34.15	-3.789	-22.35	-0.237	-3.36
x Full time	0.880	7.66	1.257	7.23	0.494	6.36
x weekend	-1.830	-9.77	-2.883	-11.22	-1.115	-8.60
β School	-5.672	-18.52	-7.298	-21.31	-1.578	-8.85
x 30 or younger	1.440	5.01	1.741	5.01	0.634	4.60
β Shopping	-3.363	-60.27	-4.175	-39.19	-0.496	-11.19
β Private	-3.643	-47.91	-4.762	-38.19	-0.716	-10.05
β Leisure	-3.106	-63.07	-3.661	-36.72	-0.282	-7.13
x weekend	0.115	1.89	0.283	2.64	0.183	4.49
γ Work	9.186	8.43	3.323	9.64	7.426	8.38
γ School	5.414	4.56	3.380	4.57	8.003	5.25
γ Shopping	0.804	8.05	0.443	7.80	2.452	4.32
γ Private	1.081	5.68	0.751	4.99	4.012	4.74
γ Leisure	3.811	8.16	1.713	8.63	5.619	6.29
δ Work - School			-0.021	-2.52	-0.208	-4.30
δ Shopping - Private business			0.011	2.69	0.107	3.65
δ Shopping - Leisure			0.017	4.97	0.108	4.42
δ Private business - Leisure			0.023	7.24	0.172	6.60
σ	0.661	13.758	1.932	17.19	0.709	9.79
Parameters		16		20		20
Loglikelihood		-10446.59		-10577.74		-10706.19
RMSE		115		48		96

** Robust t-ratio*

$\alpha > 0$ ($\alpha < 0$) implies a positive (negative) effect of z_0 on ψ_0, therefore an increased (decreased) consumption of the outside good, and a decreased (increased) consumption of the inside goods when z_0 grows. In this particular application, the negative sign of $\alpha_{\texttt{female}}$ indicates that, after controlling for other variables, women on average perform more out-of-home activities than men.

Concerning the β parameters, all of them are negative because all "inside" activities are less common than the "outside" activity (staying at *home*, see table 3). These parameters become more negative as the engagement with their corresponding activity decreases, except for *leisure* and *work* in *eMDC1*, probably due to the effect of interactions. As expected, working full time increases the chance to engage in *work* activities, while the weekend decreases it but increases the chance of engaging in *leisure* activities; and being 30 years old or younger increases the probability of engaging in *school* activities. γ parameters follow a similar trend, with higher values associated with activities performed for longer periods of time. The only exception is *school*, which has a large γ parameters despite being consumed for shorter periods than *leisure*, probably to compensate for its small $\psi_{\texttt{school}}$.

Only the *eMDC* models provide information on complementarity and substitution through their δ parameters, which are fairly consistent across *eMDC1* and *eMDC2*. As expected, there is substitution between *work* and *school*, because few people work and study concurrently. On the other hand, we observe complementarity between *shopping*, *private business* and *leisure*, probably because all of these activities are often performed at the city centre, and therefore easier to chain into a single trip. As table 3 shows, correlations between time consumption are negative for all pairs of activities, because of the fixed budget and competing nature of the activities. Yet we do observe that correlations with a magnitude smaller than 0.05 tend to be associated with complementarity effects. In section 6.3, we again compare correlations and complementarity/substitution parameters, but in a dataset where the budget constraint is less strenuous, finding a much stronger connection between them.

Concerning fit, the *eMDC1* model achieves the lowest RMSE of the three models, followed by *eMDC2* and *MDCEV*. We expected the *eMDC1* achieving the best fit, as it uses all the available information, including the total consumption or budget, and it includes complementarity and substitution effects. On the other hand, it was hard to predict which of the other two models would achieve the second best fit, as the *MDCEV* model omitts complementarity and substitution, while the *eMDC2* model does not use information about the budget. In this particular case, the *eMDC2* model fit better than *MDCEV*, but this is probably a dataset-dependent result, and may change in other study scenarios. The loglikelihood is not comparable across models, as they have different formulations, making the RMSE a better indicator of fit. In summary, when the budget is known, and will be known in future scenarios when forecasting is relevant, then we recommend using the *eMDC* model with observed budget.

6.2 Variable budget and fixed prices: expenditure dataset

The second dataset records expenditure during a fortnight for 10,460 Chilean households, aggregated to a dozen categories: *food*, *alcoholic beverages*, *clothing*, *bills* (rent and utilities), *homeware*, *health*, *transport*, *communications* (IT), *leisure*, *education*, *restaurants*, and *other*. This data comes from the 7^{th} Chilean Expenditure Survey (Bilbao, 2013). We use the expenditure in *bills* as the outside good, because all households in the sample pay rent or utilities and as this is -on average- the biggest expenditure of most households. Table 5 presents a summary of the data in thousands of Chilean pesos (kCLP, around 1.1 EUR).

Table 5: Main descriptive statistics of the expenditure data

	Fraction of the sample who bought	Total consumption (kCLP)	Average consumption when bought (kCLP)	Average consumption when bought (fraction of budget)
Food	99.6%	1532154	147.04	19.8%
Alcohol	53.8%	132523	23.55	2.8%
Clothing	53.5%	378139	67.57	5.8%
Bills	100.0%	2703618	258.47	32.7%
Homeware	88.1%	585591	63.55	5.3%
Health	72.7%	543183	71.39	5.9%
Transport	92.2%	1421741	147.50	11.7%
Communications	80.8%	418142	49.51	5.3%
Leisure	84.0%	580010	65.99	5.7%
Education	56.9%	641883	107.95	8.7%
Restaurants	69.8%	364162	49.91	4.2%
Others	93.9%	742852	75.62	6.7%

We estimated four different models with the available data. *eMDC1-100* is an *eMDC* model with observed budget equal to each household total expenditure, i.e. using the true (correct) budget. We estimated two additional eMDC models with observed budget: one assuming only 80% and another 120% of the true budget, which we call *eMDC1-80* and *eMDC1-120*, respectively. We also estimated one *eMDC* model with implict budget, which we called *eMDC2*. All models use the same formulation, including both intercepts and explanatory variables in both the base utilities and satiation parameters (i.e. $\psi_k = e^{\beta_k+\beta_{k,z}z+\varepsilon_k}$ and $\gamma'_k = \gamma_k + \gamma_{k,z}z$). The base utility of the outside good does not include an intercept to avoid identification issues, as discussed in section 4.1. Only the most relevant complementarity/substitution parameters ($\delta_k l$) identified through a Principal Component Analysis of the consumption data were included in the model. Non significant parameters were removed from the final formulation. The expenditure was expressed as hundreds of thousands of CLP. Parameter estimates and maximum log-likelihood values for *eMDC1-100* and *eMDC2* are presented in Table 6. Parameter estimates of *eMDC1-80* and *eMDC1-120* followed similar trends, and are available from the authors.

α, β and γ parameters follow a similar trend in models *eMDC1-100* and *eMDC2*. Results indicate that having a female or older household head both increase the marginal utility of the outside good (i.e. decrease expenditure in the inside goods), while a more educated household head has the opposite effect. These effects can be explained by the low female participation in the labour market (Contreras and Plaza, 2010), higher levels of education among younger individuals (for Economic Co-operation and Development, 2009), and a strong correlation between level of education and income among the Chilean population (Bilbao, 2013). Among β parameters, we observe that a higher number of adults, children, elders, workers and students per household increase the chance of spending money on alcohol, clothing, health, transport and education, all of which are reasonable effects. Furthermore, the estimates of the γ parameters indicate that more populous households tend to spend more on food, transport, communications, leisure, education and others, but not necessarily on alcohol, clothing, homeware, health, and restaurants, as these categories are more discretionary.

Complementarity and substitution parameters δ are particularly different between the model with observed and implicit budget (*eMDC1-100* and *eMDC2*, respectively). While the model with observed budget captures substitution between multiple pairs of categories, the model without it is dominated by complementarity. This is because when the budget is not controlled for, all categories of consumption seem to increase or decrease in tandem, because a higher (lower) income implies a higher (lower) expenditure across all categories. In other words, the income effect is confounded with complementarity in the model with implicit budget, as discussed in section 4.1.

Our main objective with this dataset was to analyse how errors in the definition of the budget lead to different forecast errors in models with observed budget. To do this, we first estimated the models using 70% of the full sample (training dataset), and then forecast demand on the remaining 30% of observations (validation dataset) multiple times, assuming a different value of the budget in each occasion. We repeated this for each of the *eMDC1* models we estimated. Different budgets lead to different forecasts in the *eMDC1* models, but not in *eMDC2* model. Figure 5 presents the results of this exercise. We used the root mean squared error (RMSE) of the aggregate predictions in the validation sample as an indicator of error in the forecast.

As Figure 5 shows, the forecast performance of the model with implicit budget (*eMDC2*) does not change as a function of the budget. Instead, the *eMDC1* models achieve a better forecast performance when the forecast budget is close to the estimation budget, but their error grows in a quadratic way with the budget misspecification. It does not seem to be very important how the estimation budget is defined in *eMDC1* models. For example, the estimation budget could be defined as the total income of the household or just the total expenditure on the inside goods plus one. However, once a budget has been used during estimation, it is very important to accurately and consistently predict the budget for any forecasting scenario, otherwise the forecast error can

Table 6: Comparison of model with observed and implicit budget on expenditure dataset

	eMDC1-100		eMDC2	
	Estimate	t-ratio*	Estimate	t-ratio*
α Household (hh) head is female	0.1029	7.21	0.1743	8.00
α hh head's age (years) †	0.4229	43.06	0.3925	23.17
α hh head's years of education ‡	-0.0902	-7.21	-0.4712	-18.99
β Food	4.2627	35.19	4.0531	26.43
β Alcohol	0.6105	14.70	-0.0700	-1.53
x number of adults	0.1579	13.00	0.2116	16.15
β Clothing	0.7081	23.69	0.4573	3.93
x number of children	0.1477	11.65	0.0912	6.31
β Homeware	2.1172	60.92	1.6430	16.66
β Health	1.4062	43.99	0.9394	12.18
x hh head over 60 years old	0.1655	9.49	0.2726	14.16
β Transport	2.0564	52.22	1.6049	39.11
x Number of workers in hh	0.2727	17.17	0.3156	3.79
β Communications	1.8652	56.51	1.4337	14.55
β Leisure	1.9036	58.63	1.4225	14.06
β Education	0.0000	(fixed)	0.0000	(fixed)
x Number of students	0.9261	47.22	0.7825	29.13
β Restaurants	1.3683	45.52	0.8445	9.14
β Others	2.5457	65.21	2.0959	31.47
γ Food	0.0171	8.36	0.0147	3.95
x hh size	0.0172	8.37	0.0159	4.84
γ Alcohol	0.1146	43.33	0.1204	19.82
γ Clothing	0.2889	35.37	0.3001	33.10
γ Homeware	0.0760	32.00	0.0942	20.42
γ Health	0.1436	33.62	0.1743	20.74
γ Transport	0.0946	16.77	0.1104	2.64
x hh size	0.0245	5.01	0.0215	0.50
γ Communications	0.0855	28.53	0.1075	11.74
x hh size	0.0218	10.15	0.0248	2.15
γ Leisure	0.0756	20.47	0.0885	5.59
x hh size	0.0267	9.27	0.0315	1.84
γ Education	0.3491	24.98	0.3695	8.35
x hh size	-0.1286	-24.91	-0.1360	-8.29
γ Restaurants	0.1265	37.74	0.1504	23.76
γ Others	0.0408	18.75	0.0483	17.30
x hh size	0.0224	12.61	0.0274	5.32
δ Leisure – Restaurants	0.1096	5.54	0.9390	9.02
δ Alcohol – Homeware	-0.3583	-8.04	0.3679	3.08
δ Alcohol – Health	-0.4486	-7.35	0.1250	0.28
σ	1.0044	141.86	1.0295	88.12
Number of parameters		39		39
Loglikelihood		-54929.18		-69141.89

** Robust t-ratio †log transform ‡log(1 + x) transform*

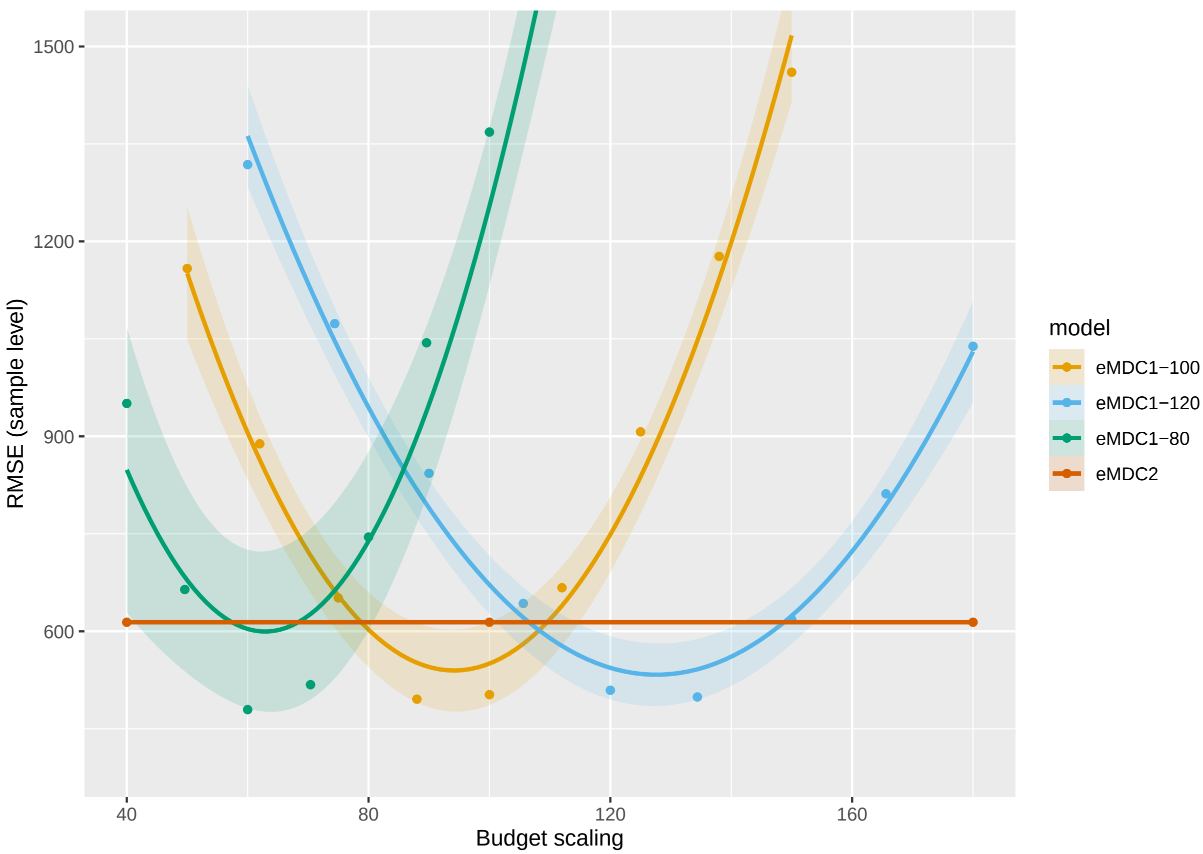

Figure 5: Comparison of forecast precision of model with implicit and observed budget, when the budget is wrongly specified in the latter.

increase rapidly.

These results reveal that in contexts where the forecasting of the budget implies even mild uncertainty, the proposed model with implicit budget can ensure a bounded level of error in the forecast.

6.3 Variable budget and variable prices: supermarket scanner dataset

The third application deals with scanner data from a chain of supermarkets (Venkatesan, 2014). After dropping all records of transactions from households with missing socio-demographic characteristics, and limiting the analysis to only four product categories, the dataset contains 4,002 purchase baskets from 656 households. All the considered product categories are fresh fruits: oranges, peaches, pears, and pineapples. Each fruit can be purchased in packs of different weights, but to simplify the analysis, we calculated the average price per Kg of each product, and expressed the amount purchased in Kg. Table 7 summarises consumption in the dataset.

Table 7: Main descriptive statistics of the supermarket scanner data

	Fraction of sample who bought (%)	Consumption					
		Total	Avg. when bought		Correlation		
		(Kg)	(Kg)	(% budget)	Oranges	Peaches	Pears
Oranges	24.0	758	0.79	51.1	1.00		
Peaches	28.0	988	0.88	49.9	-0.04	1.00	
Pears	20.7	645	0.78	44.5	-0.05	0.13	1.00
Pineapples	43.4	1406	0.81	51.1	-0.08	-0.16	-0.09

Our objective with this dataset was to compare the model with observed and implicit budget in terms of their sensitivity to changes in price. We estimated two models on the supermarket dataset: *eMDC1* is the model with observed budget, which we set to the observed consumption plus one; the second model (*eMDC2*) assumes an implicit budget. The parameter estimates and log-likelihood at convergence of these models are shown in Table 8. Non significant parameters were not removed from the model formulation. To compare their sensitivity to price, we changed the price of oranges between 70% and 130% of their original price, and calculated both models' aggregated forecast demand on the training dataset. Figure 6 plots the demand forecast by each model, for different prices.

As can be seen in Figure 6, both models predict a similar demand for the product whose price changes (oranges), but offer different predictions for the other products, whose prices remain constant. This is because of the income effect only being present in the model with observed budget, pushing for a much more dramatic reassignment of consumption when price changes. On the other hand, the model with implicit budget assumes a large unobserved budget, inducing smaller reassignment effects caused only by the δ parameters. Assuming a larger budget in *eMDC1* would decrease the sensitivity of the forecast demand among the products whose price does not change, making it more similar to the forecast of the *eMDC2* model (not reported). Based on the available data we cannot determine which of the two predictions is more accurate, as we are forecasting for unobserved prices.

The complementarity and substitution (δ_{kl}) parameters are significantly different across models. While *eMDC1* captures only complementarity, *eMDC2* captures both complementarity and substitution. This is because the δ parameters in *eMDC2* are not only capturing the complement-

Table 8: Parameters estimates of model with observed and implicit budget on the supermarket scanner dataset

	eMDC1 (observed budget)		eMDC2 (implicit budget)	
	Estimate	t-ratio*	Estimate	t-ratio*
α Household (hh) size	0.004	0.33	-0.010	-0.75
α Age of hh head	0.028	1.47	0.015	0.90
β Oranges	0.934	12.41	0.922	12.40
β Peaches	0.841	11.22	0.873	12.96
β Pears	0.789	10.45	0.751	10.21
β Pineapples	0.824	11.06	0.922	14.24
β Discount	0.061	4.64	0.321	15.35
γ Oranges	8.874	4.04	1.329	14.86
γ Peaches	12.461	5.35	1.654	14.86
γ Pears	10.610	3.76	1.679	15.93
γ Pineapples	5.454	12.56	1.199	17.63
δ Oranges – Peaches	0.382	5.06	-0.538	-2.47
δ Oranges – Pears	0.295	2.66	-0.266	-1.96
δ Oranges – Pineapples	0.188	2.74	-0.892	-4.02
δ Peaches – Pears	0.798	7.86	0.300	2.24
δ Peaches – Pineapples	0.014	0.26	-1.037	-7.73
δ Pears – Pineapples	0.011	0.14	-0.614	-8.86
σ	0.254	35.19	0.367	21.01
Number of parameters		18		18
Log-likelihood		-714.6124		-9214.29
RMSE		41.62		64.76

** Robust t-ratio*

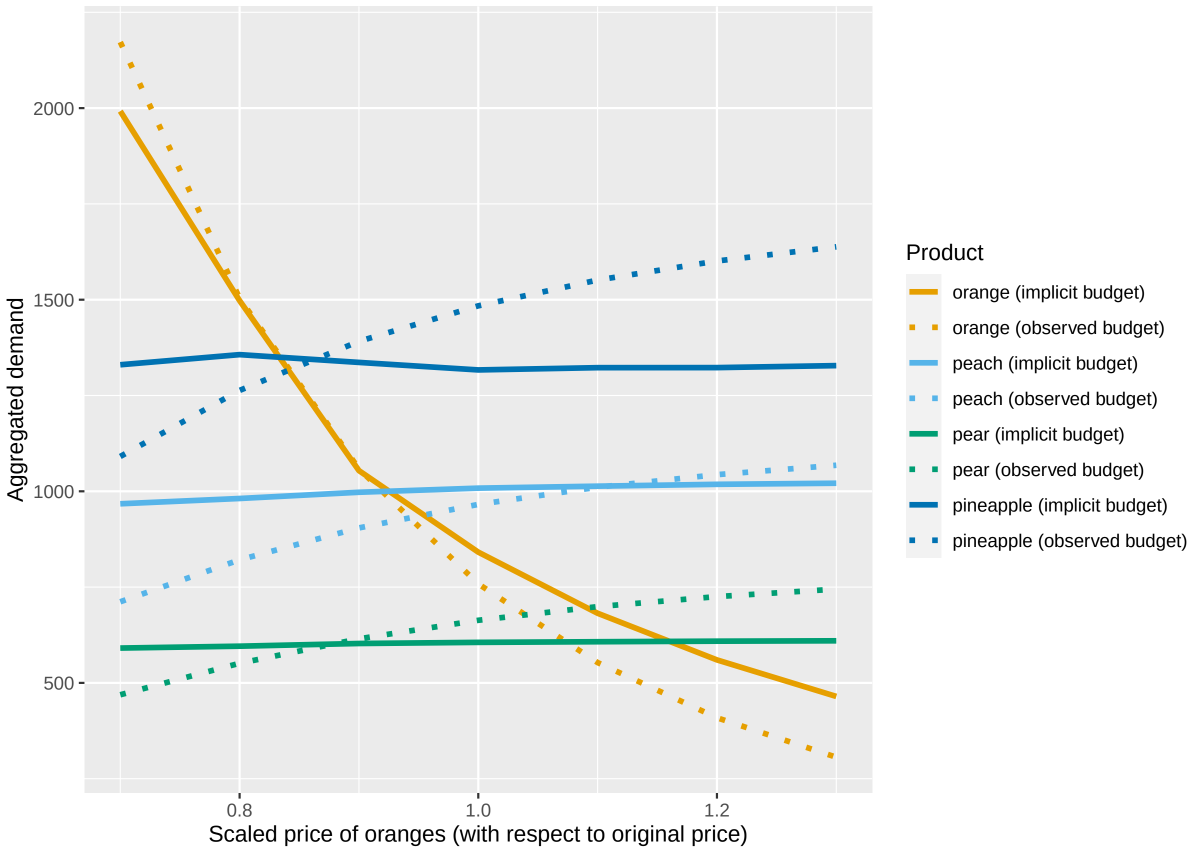

Figure 6: Relative aggregated sample demand forecasted by the traditional and extended MDCEV models for variations in the price of oranges. The black line indicates unity (i.e. original demand).

arity and substitution effects, but are also confounded with the income effect. This is apparent as the sign of δ parameters in *eMDC2* mirror those of the correlation of demand in the dataset (see table 7). This also explains why the δ parameters in *eMDC2* have higher t-ratios, as they are used to capture any interaction between the demand of different products, be it due to complementarity, substitution, or income effects. Larger budgets (as compared to expenditure in inside goods) will reduce the size of income effects, making the model with implicit budget more suitable for such scenarios.

Table 9: Main descriptive statistics of the number of trips database

		Number of trips							Homes
		Work	Study	Per. B.	Shopping	Leisure	Ret. home	All	
Number of vehicles	0	1.16	0.78	0.82	0.54	0.12	3.05	6.46	6475
	1	1.46	0.91	1.17	0.52	0.17	3.5	7.73	3508
	≥ 2	2.11	1.08	1.81	0.67	0.41	4.35	10.44	944
Household income	Low	0.61	0.74	0.97	0.6	0.12	2.7	5.75	3691
	Mid	1.34	0.92	0.93	0.51	0.12	3.34	7.16	3605
	High	2.04	0.89	1.16	0.52	0.23	3.86	8.7	3631
Total		1.34	0.85	1.02	0.54	0.16	3.3	7.21	10927

6.4 Unknown budget: Number of trips by purpose dataset

The last application deals with number of trips generated by a household, split across different purposes: *work, study, personal business, leisure* and *return home*. Data comes from the 2012 Origin-Destination survey of Santiago, Chile (Observatorio Social, 2014). The database contains observations for a single day from 10,927 households. Table 9 summarises the average number of trips per purpose by households' number of vehicles and income.

Our objective with this dataset is to compare out-of-sample forecast performance between the proposed models with explicit and implicit budget (*eMDC1* and *eMDC2*, respectively) when the definition of the budget is arbitrary. In theory, the budget in our dataset should be the maximum amount of trips a household could generate during a day, but this value is very difficult to determine. Defining the budget as any lower (but more reasonable) value would be an arbitrary decision. A common approach in situations without an evident budget is to use the observed total consumption as the budget (Bhat and Sen, 2006). We follow this approach when estimating *eMDC1*, assuming the budget to be equal to the observed total number of trips plus one, so that the "outside good" is always consumed. However, this strategy poses a problem when predicting out of sample, as the budget needs to be predicted using an auxiliary model. To reproduce this situation, we estimate our models using only 70% of the whole sample, and predict for the remaning 30%. In the case of *eMDC1* we predict the budget using a linear regression on the training data. In the case of *eMDC2* we have no need to make assumptions on the budget nor using an auxiliary model for out-of-sample prediction, as the budget is not needed during estimation nor forecasting.

In both *eMDC1* and *eMDC2* we use a linear function with the same socio-demographics to explain the base utility of the outside good (ψ_0). The base utility of the inside good and their satiation is described by a single constant each. The linear regression used to predict the budget has the same socio-demographics as explanatory variables than the discrete-continuous models. Table 10 presents the coefficients of each model estimated with the training dataset (70% of the whole sample), and their forecast performance when predicting on the validation dataset (remaining 30% of the sample). Table 11 presents the complementarity/substitution (δ) parameters of both *eMDC1* and *eMDC2*.

Table 10: Parameter estimates and forecast performance for models on number of trips dataset

	Number of trips (linear regression)		eMDC1 (observed budget)		eMDC2 (implicit budget)	
	Estimate	t-ratio*	Estimate	t-ratio*	Estimate	t-ratio*
α Intercept	1.7474	20.02				
α Household size	1.6564	61.11	-0.00104	-22.85	-0.0275	-8.65
α Number of vehicles	1.0022	17.83	-0.00063	-9.79	-0.0152	-6.82
α Bicycle availability	0.2168	2.96	-0.00023	-1.98	-0.0063	-2.60
α Household income	0.1879	3.41	-0.00011	-5.15	-0.0025	-2.06
α Number of workers	0.1035	2.24	-0.00024	-7.01	-0.0068	-3.53
β Work			-0.00077	-1.86	-0.0149	-1.95
β Study			-0.00081	-3.58	-0.0347	-6.23
β Personal business			0.00008	0.88	-0.0535	-10.58
β Shopping			0.00269	10.81	-0.0227	-1.49
β Leisure			-0.00467	-15.85	-0.0706	-5.75
β Return home			0.00136	8.75	0.0786	4.32
γ Work			336.12	10.89	11.3760	5.39
γ Study			296.23	22.62	10.7013	7.11
γ Personal business			325.12	27.45	15.5757	5.83
γ Shopping			184.46	36.04	9.2767	4.48
γ Leisure			355.13	16.87	9.4037	7.37
γ Return home			569.84	22.17	14.6113	5.47
σ			0.0035	25.82	0.0863	7.80
Number of parameters				33		33
R^2 / Loglikelihood		0.469		-2250.37		-56276.52
RMSE † indiv. level		3.11		1.07		1.08
RMSE † sample level		14.12		433.72		248.42

** Robust t-ratio. † Calculated based on out-of-sample prediction*

Establishing parallels between the parameters of both models is difficult. In the model with observed budget (*eMDC1*) the effect of socio-demographics has two components: their effect on the budget prediction, and their effect on the multiple discrete continuous model itself. On the other hand, the model with implicit budget (*eMDC2*) does not have this complexity. The sign of

Table 11: Complementarity/substitution (δ_{kl}) parameters in trips dataset.

	Work	Study	Personal B.	Shopping	Leisure	Return H.
Work		-0.0401	-0.1069	-0.1070	-0.0266	0.0948
Study	-0.0013		-0.0333	-0.0532	-0.0333	-0.0198
Personal B.	-0.0049	-0.0013		-0.0394	-0.0017†	0.0393
Shopping	-0.0050	-0.0023	-0.0020		0.0062	-0.0180†
Leisure	-0.0016	-0.0018	-0.0001†	0.0000		-0.1411
Return H.	0.0045	-0.0021	-0.0006	-0.0045	-0.0031	

Lower (upper) triangular matrix exhibits δ_{kl} from eMDC1 (eMDC2).
† *Not significant at 95% confidence.*

the complementarity/substitution parameters (δ) are consistent across models, with the exception of the *Personal business* - *Return home* pair.

In term of forecast performance at the aggregate level, the model with implicit budget (*eMDC2*) is more precise than the one with observed budget (*eMDC1*), as reflected in the last line of table 10. This is probably due to the prediction of the budget not being precise enough (see figure 5). At the individual level, both models perform similarly, though these kinds of models are rarely used to forecast at the individual level. This shows once again that the model with implicit budget is preferable when there is significant uncertainty in the prediction of the budget.

7 Conclusions

Many decisions can be represented by interrelated discrete and continuous choices, i.e. choosing *what* (incidence) and *how much* (quantity) to choose from a set of finite alternatives. A few examples include purchase decisions at a retail store (what to buy and how much of it), time use (what activities to perform and for how long), investment decisions (what instruments to buy or projects to execute and how much to invest in each), energy matrix choice (what energy sources to use and how much of each), etc. Among other approaches, this kind of decisions have been modelled using Kuhn-Tucker demand systems, which derive econometric models directly from the consumer utility maximising problem. This provides a strong grounding in economic theory, but also implies the necessity to define a budget, and imposes limitations on the definition of the utility function, leading to the omission of relevant effects, notably complementarity and substitution, in most implementations.

In this paper, we proposed two extensions to the Multiple Discrete Continuous framework: a Kuhn-Tucker demand model that incorporates complementarity and substitution effects, and another that -additionally to these effects- does not require the analyst to define a budget. The inclusion of explicit complementarity and substitution effects enriches the interpretability and realism of the model (Manchanda et al., 1999), while its functional form avoids issues present in previous formulations proposed in the literature (see section 1). The second model, with its implicit budget, is particularly useful when forecasting as it avoids cascading errors due to inaccurate budget predictions (see section 6.2).

The model with implicit budget is based on the hypothesis that total expenditure on the alternatives under consideration is small compared to the overall budget. This hypothesis allows us to approximate the utility of the numeraire good by a linear function, hence removing the necessity to define a budget. This approximation comes at the cost of reduced fit, as compared to the model with observed budget. However, simulations show that the fit of both models converges when the hypothesis above is fulfilled (see section 4.3). Such an assumption is realistic in most daily consumption decisions, but should always be justified when using the model. In general, if the budget can be determined with a great degree of confidence in forecasting scenarios, then we recommend using the model with observed budget. But if there is significant uncertainty in the budget prediction, the model with implicit budget can be a useful alternative, as it makes the prediction error independent from the budget estimation.

A computational implementation of the proposed model is available for R, as an extension of the *Apollo* package (Hess and Palma, 2019). To download this extension and see examples, visit ApolloChoiceModelling.com.

The models proposed in this paper contribute to the literature on Kuhn-Tucker system demand models to study multiple-discrete choices. There are still several avenues for improvement and further investigation. New functional forms for the complementarity and substitution term in the direct utility function could be explored, with special emphasis on those leading to a compact form of the Jacobian in the likelihood function. More generally, including a random component in the marginal utility of the outside good would be a useful development, especially if it leads to a closed-form likelihood function. Alternative formulations based on indirect utility functions could be less restrictive, as they avoid assumptions on the shape of decision makers' direct utility functions. The model formulation could also be modified to incorporate multiple constraints, for example a monetary and a time budget, or a storage capacity. Of particular interest would be an approach that mixes constraints with an explicit and implicit budget. Finally, an empirical comparison of alternative formulations for the complementarity and substitution component of the utility, as well as the utility of the outside good, is of much interest specially given recent developments in Bhat (2018) and Pellegrini et al. (2021a).

8 Acknowledgements

The authors thank Abdul Pinjari, Luis Ignacio Rizzi and an anonymous reviewer for their valuable feedback on earlier formulations of the proposed model and manuscript. The authors also acknowledge the financial support by the European Research Council through the consolidator Grant 615596-DECISIONS.

References

Bhat, C.R., 2008. The multiple discrete-continuous extreme value (mdcev) model: role of utility function parameters, identification considerations, and model extensions. Transportation Research Part B: Methodological 42, 274–303.

Bhat, C.R., 2018. A new flexible multiple discrete–continuous extreme value (mdcev) choice model. Transportation Research Part B: Methodological 110, 261–279.

Bhat, C.R., Castro, M., Khan, M., 2013. A new estimation approach for the multiple discrete–continuous probit (mdcp) choice model. Transportation Research Part B: Methodological 55, 1–22.

Bhat, C.R., Castro, M., Pinjari, A.R., 2015. Allowing for complementarity and rich substitution patterns in multiple discrete–continuous models. Transportation Research Part B: Methodological 81, 59–77.

Bhat, C.R., Sen, S., 2006. Household vehicle type holdings and usage: an application of the multiple discrete-continuous extreme value (mdcev) model. Transportation Research Part B: Methodological 40, 35–53.

Bilbao, F., 2013. VII Encuesta de presupuestos familiares. techreport. Instituto Nacional de Estadísticas Chile.

Calastri, C., Hess, S., Daly, A., Carrasco, J.A., 2017. Does the social context help with understanding and predicting the choice of activity type and duration? an application of the multiple discrete-continuous nested extreme value model to activity diary data. Transportation Research Part A: Policy and Practice 104, 1–20.

Calastri, C., Hess, S., Palma, D., Crastes dit Sourd, R., 2019. Capturing relationship strength: a choice model for leisure time, frequency of interaction and ranking in name generators. Working paper.

Calastri, C., Hess, S., Pinjari, A.R., Daly, A., 2020a. Accommodating correlation across days in multiple discrete-continuous models for time use. Transportmetrica B: Transport Dynamics 8, 108–128.

Calastri, C., dit Sourd, R.C., Hess, S., 2020b. We want it all: experiences from a survey seeking to capture social network structures, lifetime events and short-term travel and activity planning. Transportation 47, 175–201. doi:10.1007/s11116-018-9858-7.

Chintagunta, P.K., 1993. Investigating purchase incidence, brand choice and purchase quantity decisions of households. Marketing Science 12, 184–208.

Chintagunta, P.K., Nair, H.S., 2011. Structural workshop paper—discrete-choice models of consumer demand in marketing. Marketing Science 30, 977–996.

Contreras, D., Plaza, G., 2010. Cultural factors in women's labor force participation in chile. Feminist Economics 16, 27–46.

Dumont, J., Hess, S., Daly, A., Ferdous, N., 2013. The use of the multiple discrete continuous extreme value modeling framework when the budget is latent: two consumer package good examples, in: 3rd International Choice Modelling Conference, Sydney.

for Economic Co-operation, O., Development, 2009. Reviews of national policies for education: Tertiary education in Chile. OECD/World Bank.

Edgerton, D.L., 1997. Weak separability and the estimation of elasticities in multistage demand systems. American Journal of Agricultural Economics 79, 62–79.

Enam, A., Konduri, K.C., Eluru, N., Ravulaparthy, S., 2018. Relationship between well-being and daily time use of elderly: evidence from the disabilities and use of time survey. Transportation 45, 1783–1810.

Ferdous, N., Pinjari, A.R., Bhat, C.R., Pendyala, R.M., 2010. A comprehensive analysis of household transportation expenditures relative to other goods and services: an application to united states consumer expenditure data. Transportation 37, 363–390.

Habib, K.M., Miller, E.J., 2008. Modelling daily activity program generation considering within-day and day-to-day dynamics in activity-travel behaviour. Transportation 35, 467–484.

Habib, K.M.N., Miller, E.J., 2009. Modelling activity generation: a utility-based model for activity-agenda formation. Transportmetrica 5, 3–23.

von Haefen, R.H., 2004. Empirical strategies for incorporating weak complementarity into continuous demand system models. Unpublished, Department of Agricultural and Resource Economics, University of Arizona .

Hanemann, W.M., 1978. A methodological and empirical study of the recreation benefits from water quality improvement. Department of Agricultural and Resource Economics, University of California.

Hausman, J.A., Leonard, G.K., McFadden, D., 1995. A utility-consistent, combined discrete choice and count data model assessing recreational use losses due to natural resource damage. Journal of Public Economics 56, 1–30.

Hess, S., Palma, D., 2019. Apollo: a flexible, powerful and customisable freeware package for choice model estimation and application. Journal of Choice Modelling , 100170.

Hicks, J.R., Allen, R.G., 1934. A reconsideration of the theory of value. part i. Economica 1, 52–76.

Jäggi, B., Erath, A., Dobler, C., Axhausen, K.W., 2012. Modeling household fleet choice as function of fuel price by using a multiple discrete–continuous choice model. Transportation Research Record 2302, 174–183.

Jeong, J., Kim, C.S., Lee, J., 2011. Household electricity and gas consumption for heating homes. Energy Policy 39, 2679–2687.

Kim, J., Allenby, G.M., Rossi, P.E., 2002. Modeling consumer demand for variety. Marketing Science 21, 229–250.

Lee, S., Allenby, G.M., 2009. A direct utility model for market basket data. Fisher College of Business Working Paper .

Lee, S., Kim, J., Allenby, G.M., 2010. A category-level model of asymmetric complements. Fisher College of Business Working Paper .

Lim, S., Kim, Y., 2015. How to design public venture capital funds: Empirical evidence from s outh k orea. Journal of Small Business Management 53, 843–867.

Lu, H., Hess, S., Daly, A., Rohr, C., 2017. Measuring the impact of alcohol multi-buy promotions on consumers' purchase behaviour. Journal of choice modelling 24, 75–95.

Maler, K.G., 1974. Environmental economics: a theoretical inquiry. Johns Hopkins University Press.

Manchanda, P., Ansari, A., Gupta, S., 1999. The "shopping basket": A model for multicategory purchase incidence decisions. Marketing science 18, 95–114.

Mehta, N., Ma, Y., 2012. A multicategory model of consumers' purchase incidence, quantity, and brand choice decisions: Methodological issues and implications on promotional decisions. Journal of Marketing Research 49, 435–451.

Observatorio Social, 2014. Actualización y recolección de información del sistema de transporte urbano, IX Etapa: Encuesta Origen Destino Santiago 2012. Encuesta origen destino de viajes 2012. Technical Report. Universidad Alberto Hurtado.

Palma, D., Enam, A., Hess, S., Calastri, C., dit Sourd, R.C., 2021. Modelling multiple occurrences of activities during a day: an extension of the mdcev model. Transportmetrica B: Transport Dynamics 9, 456–478. URL: https://doi.org/10.1080/21680566.2021.1900755, doi:10.1080/21680566.2021.1900755, arXiv:https://doi.org/10.1080/21680566.2021.1900755.

Pellegrini, A., Pinjari, A.R., Maggi, R., 2021a. A multiple discrete continuous model of time use that accommodates non-additively separable utility functions along with time and monetary budget constraints. Transportation Research Part A: Policy and Practice 144, 37–53.

Pellegrini, A., Sarman, I., Maggi, R., 2021b. Understanding tourists' expenditure patterns: a stochastic frontier approach within the framework of multiple discrete–continuous choices. Transportation 48, 931–951.

Pellegrini, A., Sarman, I., Scagnolari, S., 2017. Stochastic frontier estimation of holiday budgets for multiple discrete-continuous extreme value model: An application to tourist expenditure analysis, in: 5yh International Choice Modelling Conference.

Pellegrini, A., Saxena, S., Pinjari, A., Dekker, T., 2019. Alternative non-additively separable utility functions for random utility maximization based multiple discrete continuous models, in: 6th International Choice Modelling Conference.

Phaneuf, D.J., Herriges, J.A., 1999. Choice set definition issues in a kuhn-tucker model of recreation demand. Marine Resource Economics 14, 343–355.

Pinjari, A.R., 2011. Generalized extreme value (gev)-based error structures for multiple discrete-continuous choice models. Transportation Research Part B: Methodological 45, 474–489.

Pinjari, A.R., Augustin, B., Sivaraman, V., Imani, A.F., Eluru, N., Pendyala, R.M., 2016. Stochastic frontier estimation of budgets for kuhn–tucker demand systems: Application to activity time-use analysis. Transportation Research Part A: Policy and Practice 88, 117–133.

Pinjari, A.R., Bhat, C., 2010. A multiple discrete–continuous nested extreme value (mdcnev) model: Formulation and application to non-worker activity time-use and timing behavior on weekdays. Transportation Research Part B: Methodological 44, 562–583.

Pinjari, A.R., Bhat, C.R., 2011. Computationally efficient forecasting procedures for Kuhn-Tucker consumer demand model systems: Application to residential energy consumption analysis. Technical Report.

Richards, T.J., Gómez, M.I., Pofahl, G., 2012. A multiple-discrete/continuous model of price promotion. Journal of Retailing 88, 206–225.

Saxena, S., Pinjari, A.R., Bhat, C.R., 2022. Multiple discrete-continuous choice models with additively separable utility functions and linear utility on outside good: Model properties and characterization of demand functions. Transportation Research Part B: Methodological 155, 526–557. URL: https://www.sciencedirect.com/science/article/pii/S0191261521002228, doi:https://doi.org/10.1016/j.trb.2021.11.011.

Song, I., Chintagunta, P.K., 2007. A discrete–continuous model for multicategory purchase behavior of households. Journal of Marketing Research 44, 595–612.

Train, K.E., 2009. Discrete choice methods with simulation. Cambridge university press.

Venkatesan, R., 2014. The complete journey. techreport. Dunnhumby Source Files.

Von Haefen, R.H., Phaneuf, D.J., 2005. Kuhn-tucker demand system approaches to non-market valuation, in: Applications of simulation methods in environmental and resource economics. Springer, pp. 135–157.

Vásquez Lavín, F., Hanemann, W.M., 2008. Functional forms in discrete/continuous choice models with general corner solution .

Wales, T.J., Woodland, A.D., 1983. Estimation of consumer demand systems with binding non-negativity constraints. Journal of Econometrics 21, 263–285.