A Multiple Discrete Extreme Value Choice Model with Grouped Consumption Data and Unobserved Budgets

Chandra R. Bhat (corresponding author)
The University of Texas at Austin
Department of Civil, Architectural and Environmental Engineering
301 E. Dean Keeton St. Stop C1761, Austin TX 78712, USA
Tel: 1-512-471-4535; Email: bhat@mail.utexas.edu
and
The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

Aupal Mondal
The University of Texas at Austin
Department of Civil, Architectural and Environmental Engineering
301 E. Dean Keeton St. Stop C1761, Austin TX 78712, USA
Email: aupal.mondal@utexas.edu

Katherine E. Asmussen
The University of Texas at Austin
Department of Civil, Architectural and Environmental Engineering
301 E. Dean Keeton St. Stop C1761, Austin TX 78712, USA
Email: kasmussen29@utexas.edu

Aarti C. Bhat
The Pennsylvania State University
Department of Human Development and Family Studies
405 Biobehavioral Health Building, State College PA 16802, USA
Email: acb6009@psu.edu

ABSTRACT

In this paper, we propose, for the first time, a closed-form multiple discrete-grouped extreme value model that accommodates grouped observations on consumptions rather than continuous consumptions. For example, in a time-use context, respondents tend to report their activity durations in bins of time (for example, 15-minute intervals or 30-minute intervals, depending on the duration of an activity). Or when reporting annual mileages driven for each vehicle owned by a household, it is unlikely that households will be able to provide an accurate continuous mileage value, and so it is not uncommon to solicit mileages in grouped categories such as 0-4,999 miles, 5000-9,999 miles, 10000-14,999 miles, and so on. Similarly, when reporting expenditures on different types of commodities/services, individuals may round up or down to a convenient dollar value of multiples of 10 or 100 (depending on the length of time in which expenditures are sought). In some other cases, a product itself may be available only in specific package sizes (such as say, instant coffee, which is typically packaged in fixed sizes). In this paper, we use the so-called linear outside good utility MDCEV structure of Bhat (2018) to show how the model can be used for grouped consumption observations. Of course, this is also possible because the linear outside good utility does not need a continuous budget value, and allows for unobserved budgets. We discuss an important identification issue associated with this linear outside good utility model, and proceed to demonstrate applications of the proposed model to the case of weekend time-use choices of individuals and vehicle type/use choices of households.

Keywords: Multiple discrete-grouped choice models, MDCEV models, multiple discrete outcomes, linear outside good utility, grouped consumption, unobserved budgets, utility theory, time use, consumer theory.

1. INTRODUCTION

Many consumer choice situations are characterized by the choice of multiple alternatives (or goods) at the same time. These situations, referred to as "multiple discreteness" by Hendel (1999) in the literature, are usually also associated with the choice of a continuous dimension (or quantity) of consumption. Bhat (2005) proposed the label of "multiple discrete-continuous" (MDC) choice for such situations. Specifically, an outcome is said to be of the MDC type if it exists in multiple states that can be jointly consumed to different continuous extents. Starting with Wales and Woodland (1983), it has been typical to consider MDC models from a direct utility maximization perspective subject to a budget constraint associated with the total consumption across all alternatives. A particularly appealing closed-form model structure following the MDC paradigm is the MDC extreme value (MDCEV) model of Bhat (2005, 2008). Some recent applications of the MDCEV model and its many variants include the proportion of annual income spent on different transportation categories (such as vehicle purchase, gas costs, maintenance costs, air travel, etc.; see Ma et al., 2019), the holding and usage level of traditional fuel vehicles and different alternative fuel vehicle types (gasoline, diesel, hybrid, electric, fuel cell, etc.; see Shin et al., 2019), and the different types of activities (such as sleeping, reading, listening to music, playing games, talking with other passengers, working, etc.) an individual may pursue as part of multi-tasking during travel (Varghese and Jana, 2019).

The basic approach in a direct utility maximization framework for MDC choices is to employ a non-linear (but increasing and continuously differentiable) utility structure with decreasing marginal utility (or satiation). Doing so has the effect of introducing imperfect substitution in the mix, allowing the choice of multiple alternatives (see Wales and Woodland, 1983, Kim et al., 2002, von Haefen and Phaneuf, 2003, and Bhat, 2005). Bhat (2008) proposed a Box-Cox utility function form that is quite general and subsumes earlier utility specifications as special cases, and that is consistent with the notion of weak complementarity (see Mäler, 1974), which implies that the consumer receives no utility from a non-essential good's attributes if she/he does not consume it. Then, if a multiplicative log-extreme value error term is superimposed to accommodate unobserved heterogeneity in the baseline preference for each alternative, the result is the MDCEV model, which has a closed-form probability expression and collapses to the MNL in the case that each (and every) decision-maker chooses only one alternative.

In almost all of the MDC formulations thus far, especially in the context of the use of the MDCEV model and its variants, satiation effects are allowed in both the outside good as well as the inside goods. This results in a situation where the discrete and continuous consumption quantities become very closely tied to one another. Indeed, the discrete choice probability of a specific combination of consumption requires knowledge of the continuous consumption quantity of the outside good (which in turn requires the budget E to be specified, because the consumption quantity of the outside good is implicitly determined from the budget E and the continuous consumption values of all the inside goods). As discussed in detail by Bhat (2018), the tightness maintained by the traditional MDC model will typically lead to a situation where the continuous consumption amount is predicted well, but not the discrete choice (see also You et al., 2014 and Lu et al., 2017). This latter result is because, given that the same baseline parameters drive both the discrete and continuous consumption predictions in the traditional MDC model, it uses satiation in the outside good as an additional instrument to fit the continuous consumption values well (basically, the emphasis of the MDC model is to fit the continuous quantities of consumption well across all individuals, even if it is at the expense of poor fit for the discrete combination for many individuals). However, as shown by Bhat (2018), using a linear utility structure for the outside good removes the tight linkage between the continuous and discrete consumptions; in fact, using a linear utility structure for the outside good allows the explicit development of the probability of discrete consumption without any need (or knowledge) for the continuous consumption quantities or the budget. Additionally, while the resulting MDC model also focuses expressly on maximizing the likelihood of the continuous consumptions, the optimization procedure essentially "realizes" that its effort is better spent on predicting the zero continuous consumption values of the inside goods well even as its goal is to fit all inside good continuous consumptions well (because it has more limited ability to utilize the satiation in the outside good to fit the non-zero values well; it is true, however, that the traditional model can provide better continuous consumption predictions than the linear outside good utility structure used here despite its poor discrete consumption predictions).[1] Of course, having a flexible model such as that developed in Bhat (2018) that imposes a complete separation of the baseline preference for the discrete and continuous components over and beyond the linear utility

[1]A more detailed and systematic investigation of the performance of the traditional MDCEV model and the linear outside good utility model in terms of the continuous consumption value predictions is left as a direction for future research.

specification for the outside good can provide the best fit for both the discrete and continuous components. But doing so also leads to a proliferation of model parameters to be estimated (because the baseline preferences are parameterized as functions of exogenous variables).

1.1. The Linear Outside Good Utility MDCEV Model

A go-between the traditional MDC formulation (which ties the discrete and continuous consumptions very closely, and also requires the knowledge of the budget and continuous consumption values) and the Bhat (2018) formulation (which is proliferate in parameters) is to allow a linear utility specification on the outside good, but also maintain a single baseline preference for each good. The resulting model, which we will label as the "Linear Outside Good" MDCEV Model (also labeled as the $L\gamma$-profile MDCEV in Bhat, 2018), can be augmented as needed by specifying a rich structure for the satiation parameter so it varies across individuals to allow for a better fit of both the discrete and continuous components of choice.[2] This approach also allows estimation accommodating the case when the continuous consumptions of choice are not reported as such, but reported only in grouped categories, as well as when the budget constraint is unobserved, as we discuss next. Importantly, as alluded to but not explicitly stated in Bhat (2018), his Linear Outside Good MDCEV model (his $L\gamma$-profile model) immediately accommodates unobservable budgets within a continuous consumption context; in the current paper, we explicate that point while also accommodating grouped (instead of continuous consumption) data.[3]

[2] Importantly, it must be noted that the linear outside good MDCEV model is intrinsically an MDCEV model, except with the utility structure as specified in Bhat (2018) as opposed to as specified in Bhat (2008).

[3] Bhat (2008) developed a general utility formulation that subsumes earlier utility formulations for MDC situations as special cases. His general formulation includes two types of satiation parameters that he refers to as the α parameters (that engender satiation effects through exponentiating consumption quantities) or γ parameters (that create satiation by translating consumption quantities). He then proceeds to show why, in almost all empirical cases, the analyst will have to choose the α-profile (with free or "to-be-estimated" α satiation parameters after arbitrarily normalizing the γ parameters) or the γ-profile (with free or "to-be-estimated" γ satiation parameters after arbitrarily normalizing the α parameters). In most empirical contexts, the γ-profile comes out to be typically superior in data fit to the α-profile (see, for example, Bhat et al., 2016; Jian et al., 2017; Jäggi et al., 2013). Further, from a prediction standpoint, the γ-profile provides a much easier mechanism for forecasting the consumption pattern, given the observed exogenous variates, as explained in Pinjari and Bhat (2011). Thus, it is not uncommon today to use the label traditional MDCEV to refer to the utility profile with a γ-profile. In all subsequent references to the MDCEV model in this paper, it will be understood that the reference is being made to the γ-profile, except if expressly defined otherwise.

1.2. Grouped Consumption Data and Unobserved Budgets

The focus of Bhat's (2018) paper was to de-link the tight connection between the discrete and continuous consumptions of choice by (a) adopting a linear utility structure for the outside good, and (b) allowing separate baseline preferences dictating the discrete consumption choice and the continuous consumption choice. But even the use of only the first component of that de-linkage, while retaining a single baseline preference influencing the discrete and continuous choices, can be valuable in two specific circumstances (an issue that did not receive adequate attention in Bhat, 2018, even though his formulation is what allows us to address the two specific circumstances). The first situation is the case when the continuous consumption values are not observed by the analyst or are unlikely to be reported accurately by respondents. For example, as clearly evidenced by Bhat (1996) and many subsequent studies, in a time-use context, respondents report their activity time durations in bins of time, rounding to the nearest 15-minute or 30-minute duration mark. Or when reporting annual mileages driven for each vehicle owned by a household, it is unlikely that households will be able to provide a continuous mileage value, and so it is not uncommon to solicit mileages in grouped categories such as 0-4,999 miles, 5000-9,999 miles, 10000-14,999 miles, and so on. Similarly, when reporting expenditures on different types of commodities/services, individuals may round up or down to a convenient dollar value. In some other cases, a product itself may be available only in specific package sizes (such as say, instant coffee, which is typically packaged in specific sizes). In such instances, we say that the consumption quantities x_k^* (k being the index for a specific good or alternative) are observed in grouped form. We however assume that consumers make their utility-maximizing decisions based on a continuous value of each good. That is, the form of the multivariate stochasticity in x_k^* engendered by the presence of stochastic (due to unobserved heterogeneity across individuals) baseline preferences is still assumed to hold. Again, as will be discussed later, it is the linear utility profile for the outside good that enables a neat expression for model probabilities in the case when the consumed quantities are observed in grouped form, as opposed

to a continuous form. Our procedure would not be possible with Bhat's (2008) traditional MDC utility expression.[4]

The second situation where retaining a linear outside good utility profile for the outside good and a single baseline preference for the inside goods is when the budget E is not readily observed. In the case of the traditional MDC utility expression, the budget E is needed. This does create problems in the many MDC cases when this information is not readily available. For example, Bhat and Sen (2006) and Garikapati et al. (2014) assume the presence of an outside alternative that they label as the "non-motorized mode" to accommodate for the possibility that a household may not own any vehicles at all and to complete the specification of the budget E (in both studies case, E is the total annual miles driven by household vehicles plus the household annual non-motorized mileage). Their justification is that all households have to walk (and/or bicycle) for at least some non-zero distance over the course of an entire year. However, travel surveys do not always collect information on non-motorized mileage, and so both studies assign an arbitrary value of 0.5 miles/person/day × 365 days/year × household size as the non-motorized mileage to construct the budget. Many other time-use and consumption studies (see, for example, Born et al., 2014 and Castro et al., 2011) "skirt" the budget unobservability problem by focusing on specific types of sub-activities within a broader activity purpose (such as say focusing only on different types of out-of-home discretionary activities) and constructing a total budget simply as the aggregation of time spent on the specific types of sub-activities. Unfortunately, this has the problem that the budget is considered exogenous and thus the total allocation on the broader type of activity purpose has to remain fixed. A third possibility is to use a two-stage approach, such as that proposed by Pinjari et al. (2016), which uses a stochastic

[4]Another important issue is that we do consider the underlying consumption quantity as fundamentally divisible and continuous. That is, an individual can conceivably participate in an activity for a few seconds of time in a time-use model, but the self-reporting will involve a rounding off in windows of time in minutes. Similarly, a vehicle can be driven to any fraction of miles, but the reporting or recording may be done in grouped categories of miles. This situation is different from the earlier studies of Lee and Allenby, 2014 and Kuriyama and Hanemann, 2006, who focus on the case of fundamentally indivisible demand (where the underlying quantity can take only non-negative integers; sometimes referred to as count data). In addition, these earlier studies consider that there is no stochasticity in the baseline utility preference for the outside good, while we explicitly consider the more realistic case that there could be individual-level unobserved variations in the baseline preference for all goods, inside and outside. Indeed, there is certainly no reason that unobserved factors should enter only the utility preference for the inside goods, but not the outside good; and this is not simply an issue that can be waived on the grounds of the singularity issue engendered by the budget constraint, because there are real ramifications to the model structure by ignoring stochasticity in the baseline preference for the outside good; see Bhat (2008) Section 6 for a detailed discussion. Finally, similar to the MDCEV model, we use an extreme value distribution for the stochastic terms that leads to closed-form analytic structure for the consumption probability.

frontier approach to develop an expected estimate of the budget that is then used in a second stage MDC model. While an interesting approach, this is really a rather elaborate workaround with two stages that do not necessarily come together within a single unifying utility-theoretic framework. Our approach, on the other hand, retains the simplicity of the usual MDCEV model in terms of model formulation. As discussed in more detail later, there is no need for an explicit budget if a linear utility form is used for the outside alternative.[5] Of course, our approach may be viewed as a strict single stage utility-theoretic approach, which does not expressly consider potential exogenous variable effects on an overall budget that can then impact individual good consumptions. Rather, by defining the goods of interest as inside goods, changes in exogenous variables directly impact the consumptions of these inside goods (even if the true effect is an indirect impact through budget changes), co-mingling strict budget effects and strict allocation effects. Approaches to handle both an endogenous budget as well as consumption quantities separately but within a single unifying utility-theoretic framework have been elusive; additional investigations in this area are certainly an important direction for further research.

The rest of this paper is structured as follows: The next section lays out the statistical specification and the econometric modeling aspects of the multiple discrete-grouped model that we propose in this paper. In doing so, we revisit Bhat's (2018) $L\gamma$-profile MDCEV model, and discuss an important identification issue in the model that did not receive any attention in that paper. This is followed by the third section on forecasting methods that presents several approaches to forecast MDC models without an external budget and discusses forecasting techniques for multiple discrete-grouped consumptions. The fourth section provides two empirical application of the proposed method – one in the context of time-use and the other in the context of vehicle-use. Concluding remarks are provided in the fifth and last section.

[5]A related advantage of the linear utility form for the outside good is that the magnitude of the outside good consumption does not skew the results of the MDC model substantially. In particular, if the consumption of the outside good is very large (such as say in-home time investment in a time-use model), this creates problems in the traditional MDC model estimation because it will tend to drive the baseline preferences of the inside goods to very small values and also drive the satiation to be extremely high for these goods. This results in convergence problems and extremely small predicted time-investments in the inside goods. On the other hand, the use of a linear utility form for the outside good, because it focuses better on fitting the discrete probabilities and does not involve the appearance of the outside good consumption in the baseline preference for the inside goods handles such situations much better. Of course, it is possible that the traditional MDCEV model that explicitly considers the budget (with the logarithm of the outside good consumption appearing in the outside good utility) will perform better than the linear outside good MDCEV in the continuous consumption predictions (see Bhat, 2018 for a detailed explanation).

2. THE MDGEV (Multiple Discrete-Grouped Extreme Value) MODEL STRUCTURE[6]

Assume without any loss of generality that the essential Hicksian composite outside good is the first good. Following Bhat (2008) and Bhat (2018), the typical utility maximization problem (assuming the budget information is available and so is the continuous consumption values for an estimation sample) in the MDC model is written (using a gamma-profile, as discussed in Bhat, 2018) as:

$$U(\boldsymbol{x}) = \psi_1 x_1 + \sum_{k=2}^{K} \gamma_k \psi_k \ln\left\{\left(\frac{x_k}{\gamma_k}+1\right)\right\} \tag{1}$$

$$s.t. \sum_{k=1}^{K} p_k x_k = E,$$

where the utility function $U(\boldsymbol{x})$ is quasi-concave, increasing and continuously differentiable, $\boldsymbol{x} \geq 0$ is the consumption quantity ($\boldsymbol{x}$ is a vector of dimension $(K \times 1)$ with elements x_k), and ψ_k and γ_k are parameters associated with good k.[7] The constraint in Equation (1) is the linear budget constraint, where E is the total expenditure across all goods k (k = 1, 2,…, K) and $p_k > 0$ is the unit price of good k (with $p_1 = 1$ to represent the numeraire nature of the first essential good). The function $U(\boldsymbol{x})$ in Equation (1) is a valid utility function if $\psi_k > 0$, and $\gamma_k > 0$ for all k. As discussed in detail in Bhat (2008), ψ_k represents the baseline marginal utility, and γ_k is the vehicle to introduce corner solutions (that is, zero consumption) for the inside goods (k = 2, 3,…, K), but also serves the role of a satiation parameter (higher values of γ_k imply less satiation). There is no γ_1 term for the first good because it is, by definition, always consumed. Further, we use a linear utility profile (no satiation) for the outside good. Of course, the reader will note that there is an assumption of additive separability of preferences in the utility form of Equation (1), which immediately implies that none of the goods are *a priori* inferior and all the goods are strictly Hicksian substitutes (see Deaton and Muellbauer, 1980; p. 139). Further, as in

[6] This is not to be confused with the multiple discrete-continuous generalized extreme value (MDCGEV) model in Pinjari (2011) that uses a multivariate generalized extreme value distribution for the kernel error terms in the baseline preference of alternatives within the context of a multiple discrete-continuous (MDC) model rather than focusing on a multiple discrete-grouped (MDG) model. Of course, the model proposed here can be extended to an MDGGEV (multiple discrete-grouped generalized extreme value) model.

[7] The assumption of a quasi-concave utility function is simply a manifestation of requiring the indifference curves to be convex to the origin (see Deaton and Muellbauer, 1980, p. 30 for a rigorous definition of quasi-concavity). The assumption of an increasing utility function implies that $U(\boldsymbol{x}^1) > U(\boldsymbol{x}^0)$ if $\boldsymbol{x}^1 > \boldsymbol{x}^0$.

the traditional MDCEV, we maintain the assumption that there are no cost economies of scale in the purchase of goods; that is, we will continue to retain the assumption that the unit price of a good remains constant regardless of the quantity of good consumed.

2.1. Statistical Specification

To ensure the non-negativity of the baseline marginal utility, while also allowing it to vary across individuals based on observed and unobserved characteristics, ψ_k is usually parameterized as follows:

$$\psi_k = \exp(\boldsymbol{\beta}' \boldsymbol{z}_k + \varepsilon_k),\ k = 1, 2, ..., K, \tag{2}$$

where $\boldsymbol{z}_k$ is a set of attributes that characterize alternative *k* and the decision maker (including a constant), and ε_k captures the idiosyncratic (unobserved) characteristics that impact the baseline utility of good *k*. A constant cannot be identified in the $\boldsymbol{\beta}$ term for one of the *K* alternatives. Similarly, individual-specific variables are introduced in the vector $\boldsymbol{z}_k$ for (*K*–1) alternatives, with the remaining alternative serving as the base. As a convention, we will not introduce a constant and individual-specific variable in the vector $\boldsymbol{z}_1$ corresponding to the first outside good.

To find the optimal allocation of goods, the Lagrangian is constructed and the first order equations are derived based on the Karush-Kuhn-Tucker (KKT) conditions. The Lagrangian function for the model, when combined with the budget constraint, is:

$$L = U(\boldsymbol{x}) + \lambda\left(E - \sum_{k=1}^{K} p_k x_k \right), \tag{3}$$

where λ is a Lagrangian multiplier for the constraint. The KKT first order conditions for optimal consumption allocations (x_k^*) are as follows, given that $x_1 > 0$:

$$\psi_1 - \lambda = 0;$$

$$\left[\psi_k \left(\frac{x_k^*}{\gamma_k} + 1 \right)^{-1} \right] - \lambda p_k = 0 \text{ if consumption} = x_k^* \ (x_k^* > 0),\ k = 2, 3, \ldots, K, \tag{4}$$

$$[\psi_k] - \lambda p_k < 0 \text{ if } x_k^* = 0,\ k = 2, 3, ..., K.$$

Substituting $\psi_1 = \lambda$ into the latter two equations, using the statistical specification for the baseline preference functions from Equation (2), taking logarithms, and rewriting, we get:

$\eta_k = \tilde{V}_k$, $\eta_k = \varepsilon_k - \varepsilon_1$, $\tilde{V}_k = V_1 - V_k$ if consumption is equal to x_k^* ($k = 2, 3, \ldots, K$), where $x_k^* > 0$

$\eta_k < \tilde{V}_{k0}$, $\eta_k = \varepsilon_k - \varepsilon_1$, $\tilde{V}_{k0} = V_1 - V_{k0}$ if $x_k^* = 0$ ($k = 2, 3, \ldots, K$), where (5)

$V_k = \boldsymbol{\beta}' \boldsymbol{z}_k - \ln\left(\frac{x_k^*}{\gamma_k} + 1\right) - \ln p_k$, $V_{k0} = \boldsymbol{\beta}' \boldsymbol{z}_k - \ln p_k$ ($k = 2, 3, \ldots, K$), and $V_1 = \boldsymbol{\beta}' \boldsymbol{z}_1$.

As discussed at length in Bhat (2018), the linear utility form for the outside good is the reason that the expressions for V_1 above does not include x_1^*. If satiation is allowed in the outside good using the traditional specification of the sub-utility form as $\psi_1 \ln x_1$ for the outside good, $\ln x_1^*$ appears in the V_1 expression. With that, the probability expression for the observed consumption choice will require the consumption quantities for every good. However, with the linear specification, there is no need to have the consumption for the inside good (alternatively, no need for the observability of the budget *E*), as we discuss next.

2.2. Econometric Model

The econometric model is completed once assumptions are made regarding the joint distribution of the ε_k terms. As in the single discrete choice case, the two most commonly used joint distributions are the multivariate extreme value distribution or the multivariate normal distribution. Assume that the first *M* inside goods (*k*=2,3,…,*M*+1) are observed to be consumed. Assume also that the ε_k terms are independent and identically distributed with a Type-1 extreme value distributed with a scale parameter of σ. The probability that the first *M* of the inside goods are consumed ($M \geq 1$; $M < K-1$) at levels $x_2^*, x_3^*, \ldots, x_{M+1}^*$ (with zero consumption for the remaining goods) may be written as follows (see Bhat, 2008, 2018):

$$P\left(x_2^*, x_3^*, \ldots, x_{M+1}^*, 0, \ldots, 0, 0\right)$$

$$= |J| \left(\frac{M!}{\sigma^M} \times \frac{\prod_{k=2}^{M+1} e^{\frac{-(\tilde{V}_k)}{\sigma}}}{\left(1 + \sum_{k=2}^{M+1} e^{\frac{-(\tilde{V}_k)}{\sigma}} + \sum_{k=M+2}^{K} e^{\frac{-(\tilde{V}_{k0})}{\sigma}}\right)^{M+1}} \right), \text{ where } |J| = \left[\prod_{i=2}^{M+1} f_i\right], \; f_i = \left(\frac{1}{x_i^* + \gamma_i}\right). \quad (6)$$

[8] The determinant of the Jacobian as presented in Bhat (2018) has an extra $1/p_i$ term in the expression for f_i, which is incorrect. The expression given here is the correct one.

The right side of the expression above includes only the consumption quantities $x_2^*, x_3^*, ..., x_{M+1}^*$ as embedded in $\tilde{V}_k$. It does not include x_1^*, which also means that there is no need to observe the budget *E*. At the same time, it is easy now to use the above model to accommodate the case when the consumption quantities x_k^* for the consumed inside goods (*k*=2,3,…,*M*+1) are observed only in grouped form as opposed to in continuous form. Assume that what is observed in grouped form is $w_{kl} = a_{k,l-1} < x_k^* < a_{k,l}$ (*k*=2,3,…,*M*+1; *l*=1,2,…,*L*), where $a_{k,l}$ represents the upper bound for grouped category *l* for good *k* ($a_{k,0} = 0, a_{k,L} = \infty$). That is, if an individual chooses a specific grouped category *l*, it means that the continuous optimal quantity for consumption is between $a_{k,l-1}$ and a_{kl}. Let the actual observed grouped category for an individual for good *k* be c_k (that is, $w_{kl} = c_k$). Then, the probability of the consumption pattern for the case of $M \geq 1$ and $M < K-1$ may be written as follows:

$$
\begin{aligned}
&P\left(c_2, c_3, ..., c_{M+1}, 0, ..., 0, 0 \right) \\
&= P(a_{2,c_2-1} < x_2^* < a_{2,c_2},\ a_{3,c_3-1} < x_3^* < a_{3,c_3}, ..., a_{M+1,c_{M+1}-1} < x_{M+1}^* < a_{M+1,c_{M+1}}, x_{M+2}^*=0, x_{M+3}^*=0, ..., x_k^* = 0) \\
&= \int_{x_2^*=a_{2,c_2-1}}^{a_{2,c_2}} \int_{x_3^*=a_{3,c_3-1}}^{a_{3,c_3}} ... \int_{x_{M+1}^*=a_{M+1,c_{M+1}-1}}^{a_{M,c_{M+1}}} |J| \left(\frac{M!}{\sigma^M} \times \frac{\prod_{k=2}^{M+1} e^{\frac{-(\tilde{V}_k)}{\sigma}}}{\left(1 + \sum_{k=2}^{M+1} e^{\frac{-(\tilde{V}_k)}{\sigma}} + \sum_{k=M+2}^{K} e^{\frac{-(\tilde{V}_{k0})}{\sigma}} \right)^{M+1}} \right) dx_{M+1}^* ... dx_3^* dx_2^*
\end{aligned}
\tag{7}
$$

After some tedious but straightforward integration, the integral above collapses to a nice closed-form expression (see Appendix A). Specifically, define

$$
\begin{aligned}
&G_{K-1}(x_2^* < a_{2,c_2}, x_3^* < a_{3,c_3}, ..., x_{M+1}^* < a_{M+1,c_{M+1}}, x_{M+2}^* = 0, ..., x_{K-1}^* = 0, x_K^* = 0) \\
&= \boldsymbol{F}_{K-1}(\eta_2 < W_{2,c_2}, \eta_3 < W_{3,c_3}, ..., \eta_{M+1} < W_{M+1,c_{M+1}}, \eta_{M+2} < \tilde{V}_{M+2,0}, ..., \eta_{K-1} < \tilde{V}_{K-1,0}, \eta_K < \tilde{V}_{K,0}) \\
&= \left(1 + \sum_{k=2}^{M+1} e^{\frac{-(W_{k,c_k})}{\sigma}} + \sum_{k=M+2}^{K} e^{\frac{-(\tilde{V}_{k0})}{\sigma}} \right)^{-1},\ W_{k,c_k} = V_1 - \left(\boldsymbol{\beta}' \boldsymbol{z}_k - \ln\left(\frac{a_{k,c_k}}{\gamma_k} + 1 \right) - \ln p_k \right) = \tilde{V}_{k,0} + \ln\left(\frac{a_{k,c_k}}{\gamma_k} + 1 \right),
\end{aligned}
\tag{8}
$$

In the above equation, $W_{k,0} = \tilde{V}_{k,0}$ and $W_{k,L} = \infty$, and $\boldsymbol{F}_{K-1}(.)$ represents the multivariate logistic CDF that takes the general form:

$$
\boldsymbol{F}_{K-1}(\eta_2 < h_2, \eta_3 < h_3, ..., \eta_K < h_K) = \left(1 + \sum_{k=2}^{K} e^{\frac{-h_k}{\sigma}} \right)^{-1}. \tag{9}
$$

Based on the inclusion-exclusion probability law, and for all Fretchet class of multivariate distribution functions with given univariate margins (of which the multivariate logistic distribution is a part), the probability expression in Equation (7) can then be written as follows:

$$P\left(c_2, c_3, \ ..., c_{M+1}, 0, ..., 0, 0 \right) = \sum_{S=1}^{S=2^M} (-1)^{L_S} \boldsymbol{F}_{K-1}(\boldsymbol{W}_S, \tilde{V}_{M+2,0}, ..., \tilde{V}_{K-1,0}, \tilde{V}_{K,0}), \tag{10}$$

where S represents a specific combination of length M of the W_{k,c_k-1} and W_{k,c_k} scalars across all the consumed inside goods (k=2,3,…,M+1) such that both W_{k,c_k-1} and W_{k,c_k} are disallowed in the combination for any k (there are 2^M such combinations, and we will represent the resulting vector of elements in combination S as $\boldsymbol{W}_S$), and L_S is a count of the number of lower thresholds W_{k,c_k-1} (k=2,3,…,M+1) appearing in the vector $\boldsymbol{W}_S$.

In the specific case that all the inside goods are consumed (that is, $M = K-1$), the corresponding consumption probability is as follows:

$$P\left(c_2, \ c_3, ..., \ c_{M+1}, c_{M+2}, ..., c_{K-1}, \ c_K \right) = \sum_{S=1}^{S=2^{K-1}} (-1)^{L_S} \boldsymbol{F}_{K-1}(\boldsymbol{W}_S), \tag{11}$$

In the case when none of the inside goods are consumed (that is, $M = 0$), the corresponding consumption probability is:

$$P\left(0,0,...,0,0,...,0,0\right) = \boldsymbol{F}_{K-1}(\tilde{V}_{2,0}, \tilde{V}_{3,0}, ..., \tilde{V}_{M+1,0}, \tilde{V}_{M+2,0}, ...\tilde{V}_{K-1,0}, \tilde{V}_{K,0}) \tag{12}$$

For completeness, we also write an expression for the probability of discrete consumptions, which should be helpful in some of the methods of forecasting to be discussed in the next section. In particular, the KKT conditions imply the following for the discrete consumption:

$$\eta_k = \varepsilon_k - \varepsilon_1 > \tilde{V}_{k,0} \text{ if } x_k^* > 0 \ (k = 2, 3, \ldots, K), \ \tilde{V}_{k,0} = \boldsymbol{\beta}'\boldsymbol{z}_1 - \left(\boldsymbol{\beta}'\boldsymbol{z}_k - \ln p_k\right) \tag{13}$$

$$\eta_k = \varepsilon_k - \varepsilon_1 < \tilde{V}_{k,0} \text{ if } x_k^* = 0 \ (k = 2, 3, \ldots, K).$$

The first condition above states that good k will be consumed to a non-zero amount only if the price normalized random marginal utility of consumption of the first unit ($\boldsymbol{\beta}'\boldsymbol{z}_k - \ln p_k + \varepsilon_k$) is greater than the random (and constant across consumption values x_1^*) marginal utility $\left(\boldsymbol{\beta}'\boldsymbol{z}_1 + \varepsilon_1\right)$ of the outside good. Let d_k be a dummy variable that takes a value 1 if good k (k = 2, 3,…, K) is consumed, and zero otherwise. Then, the multivariate probability that the individual consumes a

non-zero amount of the first M of the K–1 inside goods (that is, the goods 2, 3,…, M+1) and zero amounts of the remaining K–1–M goods (that is, the goods M+2, M+3,…, K) takes the following form:

$$P(d_2=1,d_3=1,\cdots d_{M+1}=1,d_{M+2}=0,\cdots d_{K-1}=0,d_K=0)$$
$$=\int_{\eta_2=\tilde{V}_{2,0}}^{\eta_2=\infty}\int_{\eta_3=\tilde{V}_{3,0}}^{\eta_3=\infty}....\int_{\eta_{M+1}=\tilde{V}_{M+1,0}}^{\eta_{M+1}=\infty}\int_{\eta_{M+2}=-\infty}^{\eta_{M+2}=\tilde{V}_{M+2,0}}....\int_{\eta_{K-1}=-\infty}^{\eta_{K-1}=\tilde{V}_{K-1,0}}\int_{\eta_K=-\infty}^{\eta_K=\tilde{V}_{K,0}}\boldsymbol{f}(\eta_2,\eta_3,...,\eta_K)\,d\eta_K d\eta_{K-1,}...d\eta_2\ , \tag{14}$$

where $\boldsymbol{f}(\eta_2,\eta_3,...,\eta_K)$ represents the multivariate logistic probability density function (pdf) of the random variates $\eta_2,\eta_3,...,\eta_K$. The above expression may be written as:

$$P(d_2=1,\cdots d_{M+1}=1,d_{M+2}=0,\cdots d_{K-1}=0,d_K=0)$$
$$=\boldsymbol{F}_{K-M-1}(\tilde{V}_{M+2,0},...\tilde{V}_{K-1,0},\tilde{V}_{K,0})+\sum_{R\subset\{2,3,...M+1\},|R|\geq 1}(-1)^{|R|}\,\boldsymbol{F}_{K-M-1+|R|}(\tilde{\boldsymbol{V}}_{R,0},\tilde{V}_{M+2,0},...\tilde{V}_{K-1,0},\tilde{V}_{K,0}), \tag{15}$$

where $\boldsymbol{F}_{K-1}(.)$ for any dimension K–1 is the multivariate logistic CDF, R represents a specific combination of the consumed goods (there are a total of $M+C(M,2)+C(M,3)+...C(M,M)=2^M-1$ possible combinations of the consumed goods), $|R|$ is the cardinality of the specific combination R, and $\tilde{\boldsymbol{V}}_{R,0}$ is a vector of utility elements drawn from $\{\tilde{V}_{2,0},\tilde{V}_{3,0},...\tilde{V}_{M+1,0}\}$ that belong to the specific combination R. The multivariate logistic CDF $\boldsymbol{F}_{K-1}(.)$ takes the general form already shown in Equation (9). The CDF of any subset of the $\boldsymbol{\eta}$ vector is readily obtained from that CDF expression. For example, the CDF of only the first two elements is:

$$\boldsymbol{F}(\eta_2<h_2,\eta_3<h_3)=\left(1+e^{\frac{-h_2}{\sigma}}+e^{\frac{-h_3}{\sigma}}\right)^{-1}. \tag{16}$$

Thus, by plugging the appropriate CDF functions in the expression of (14), one can obtain a closed-form expression for the probability of any pattern of discrete consumption of the many alternatives in the MDCEV model.

The model proposed here, which handles grouped consumption data when a good is consumed, may be extended to the case when the baseline preference for each inside good is explicitly separated out into a discrete component and continuous component (this is the fully flexible model proposed in Bhat, 2018). The extension to this more general case is conceptually straightforward, though the resulting model can be profligate in parameters. The mathematical

extension to this general model is provided in Appendix B, though we will stick with the single baseline preference for each inside good case in our empirical application.

2.3. Scale Parameter of the Error Term in the Baseline Marginal Utility

In the traditional MDCEV model, the scale parameter is not identified in the absence of price variation for a general utility profile or the α-profile (see Bhat, 2008). Bhat (2018) shows, however, that the scale parameter is indeed identified even in the absence of price variation if the γ-profile is used. We now discuss the identification of the scale in this traditional MDCEV model as well as the linear outside good MDCEV model that forms the basis for the MDGEV model proposed in this paper (thus, any identification conditions that apply to the linear outside good MDCEV model will immediately apply to the MDGEV model).

2.3.1. Identification of Scale in the Traditional MDCEV Model

In this traditional MDCEV, where satiation is also allowed in the outside good (that is, a non-linear utility form is used even for the outside good), the γ-profile utility function takes the following form:

$$U(\boldsymbol{x}) = \psi_1 \ln x_1 + \sum_{k=2}^{K} \gamma_k \psi_k \ln\left\{\left(\frac{x_k}{\gamma_k}+1\right)\right\} \tag{17}$$

To find the optimal allocation of goods, the Lagrangian is constructed in the usual manner and the equivalent KKT conditions of Equation (4) in this traditional MDCEV model are similar to those in Equation (4) with the change that $V_1 = \boldsymbol{\beta}'\boldsymbol{z}_1 - \ln(x_1^*)$ instead of $V_1 = \boldsymbol{\beta}'\boldsymbol{z}_1$. It is the presence of the $\ln(x_1^*)$ in the expression for V_1 that causes the tight linkage between the continuous and discrete consumptions. It also immediately implies the need for knowledge of the budget, and it also requires observation of the consumption quantities in a strict continuous form (see Bhat, 2018 for other repercussions of the presence of $\ln(x_1^*)$ in the expression for V_1). But, as will now show, it is the presence of the outside good's consumption in V_1 that also allows for the clear estimation of the scale parameter in the traditional MDCEV even without price variation. To see this, in standardized form and without price variation, the KKT conditions for the consumed goods in the traditional MDCEV (Equation (5)) without price variation may be written as:

$$\frac{\eta_k}{\sigma}=\tilde{V}_k^*,\ \ \eta_k=\varepsilon_k-\varepsilon_1,\ \tilde{V}_k^*=\left[\left(\boldsymbol{\beta}^*\right)'\boldsymbol{z}_1-\left(\boldsymbol{\beta}^*\right)'\boldsymbol{z}_k\right]-\left[\left(\frac{1}{\sigma}\right)\ln(x_1^*)-\left(\frac{1}{\sigma}\right)\ln\left(\frac{x_k^*}{\gamma_k}+1\right)\right];\boldsymbol{\beta}^*=\left(\frac{\boldsymbol{\beta}}{\sigma}\right) \quad (18)$$

The scale parameter is distinctly estimable here because it is essentially the coefficient on the natural logarithm term of the continuous consumption quantities in the expression above. Specifically, it is the presence of the $\left(\frac{1}{\sigma}\right)\ln(x_1^*)$ that allows the distinct estimation of the scale parameter σ, because this term is not tangled up with the γ_k parameters in any way (as in the $\left(\frac{1}{\sigma}\right)\ln\left(\frac{x_k^*}{\gamma_k}+1\right)$ terms). We will again show this from a different perspective after discussing the identification case for the linear outside good utility MDCEV below.

2.3.2. Identification of Scale in the Linear Outside Good MDCEV Model

Now consider the case of the utility expression for the consumed goods for the linear outside good MDCEV model of this paper. Let us start with a different more general utility expression (see Bhat, 2008) as follows ($\alpha<1$):

$$U(\boldsymbol{x})=\psi_1^{1-\alpha}x_1+\sum_{k=2}^{K}\frac{\gamma_k}{\alpha}\psi_k{}^{(1-\alpha)}\left\{\left(\frac{x_k}{\gamma_k}+1\right)^{\alpha}-1\right\}. \quad (19)$$

Importantly, note that the above utility expression retains a linear utility profile for the outside good. It also so happens that the above utility expression collapses back exactly to the simpler linear outside good utility function of Equation (1) when $\alpha\to 0$. Because of the linear profile and lack of satiation effects in the outside good, any hope for the estimation of the α satiation parameter as well as the γ_k satiation parameters (k=2,3,…,K) reside entirely in the inside good utility specification (that is, in the second component of the utility function in Equation (19)). Of course, the way that the α parameter generates a satiation effect is through an exponentiation approach, while the way that the γ_k parameters generate satiation is through a translational mechanism. These two mechanisms are distinct, and theoretically there is no reason for both these effects not to be present simultaneously. But, especially when there is a separate α_k parameter for each of the inside goods, Bhat (2008) shows clearly that it is next to impossible to empirically identify both sets of α_k and γ_k parameters, because, for any given ψ_k value, it is

possible to closely approximate a sub-utility function profile for good *k* based on a combination of α_k and γ_k values with a sub-utility function based solely on the α_k or based solely on the γ_k values. In the case of Equation (19), the situation is a little less dire because the α_k parameters are held to be the same across the sub-utility profiles of the different inside goods. However, the same issue as in the more general case with separate α_k parameters will arise in many empirical situations even with a fixed α parameter across all the inside goods. Specifically, it will be possible to mimic literally the same sub-utility profile in Equation (19) for all the inside goods by either normalizing the α parameter or normalizing one of the γ_k parameters (we have confirmed this empirically in the two case studies discussed later). The net result is that, in many contexts, the analyst will need to either normalize the α parameter or normalize one of the γ_k parameters. The analyst can estimate both these models and select the one that provides a better fit (in most cases, this will come out to be the model that normalizes the α parameter).

The implication from the above discussion is that, in most empirical contexts, after allowing for a complete set of γ_k parameters for the inside goods, it will not be empirically possible in the linear outside good utility MDCEV model to distinguish between a specification for the baseline preference that uses $\psi_k = \exp\left(\boldsymbol{\beta}'\boldsymbol{z}_k + \varepsilon_k\right)$ and that uses $\psi_k^* = \left(\psi_k\right)^{1-\alpha} = \left[\exp\left(\boldsymbol{\beta}'\boldsymbol{z}_k + \varepsilon_k\right)\right]^{1-\alpha}$. Putting, for convenience, $\frac{1}{\sigma} = (1-\alpha)$ and taking the logarithm of the baseline preferences that appear in the KKT conditions, the net result is that it is difficult to distinguish between the specifications of $\ln\psi_k = \left(\boldsymbol{\beta}'\boldsymbol{z}_k + \varepsilon_k\right)$ and $\ln\psi_k^* = \frac{1}{\sigma}\left(\boldsymbol{\beta}'\boldsymbol{z}_k + \varepsilon_k\right)$. That is, the scale of the error term in the linear outside good utility model will not be empirically identifiable in most contexts in the absence of price variation. A convenient normalization then is to set the scale σ to 1 (or, equivalently, $\alpha \to 0$) in the linear outside good MDCEV model with a γ-profile for the case with no price variation.

2.3.3. A Revisit of the Identification Discussion in the Traditional MDCEV Model

With the discussion above, we are able to develop another perspective regarding why identification of the scale becomes possible in the traditional MDCEV model. To see this, consider the following general utility profile:

$$U(\boldsymbol{x}) = \frac{1}{\alpha}\psi_1 \left\{\left(x_1+1\right)^{\alpha}-1\right\} + \sum_{k=2}^{K} \frac{\gamma_k}{\alpha}\psi_k{}^{(1-\alpha)} \left\{\left(\frac{x_k}{\gamma_k}+1\right)^{\alpha}-1\right\} \tag{20}$$

This is different from that of Equation (19) in the specification of the outside good, and, using L'Hopital's technique (see Bhat, 2008), it is easy to see that, in the specific case that $\alpha \to 0$, the above utility function collapses exactly to the utility profile of the traditional MDCEV with a γ-profile as shown in Equation (17). In this setting, following Section 3.2 of Bhat (2008) and Section 2.1.1. of Bhat (2018), one can show that both the parameter α and the parameters γ_k are estimable. This is because the α parameter is obtaining a "pinning effect" for the satiation for the outside good (which is lacking in the linear outside utility good specification). But, also as shown in the earlier Bhat studies, the parameter α and the scale of the error terms (embedded in the logarithm of the baseline preferences) are not distinguishable in the specification of Equation (20). Specifically, any scaling may be used for the error terms, with the identity relationship between one set of α^* and σ, and another set of α with an arbitrarily normalized σ^*=1, as follows: $\sigma = \frac{1-\alpha^*}{1-\alpha}$. So, if one uses the normalization $\alpha^* = 0$, rather than the normalization σ^*=1, nothing changes except that we now get a value of $\sigma = \frac{1}{1-\alpha}$. Thus, because the parameter α is estimable in the utility profile of Equation (20) with the scale normalized to one, it immediately implies that the scale in the traditional MDCEV model with a γ-profile of Equation (17) (in which the α parameter is normalized) is estimable.

2.3.4. Intuitive Interpretation and Summary of Identification Considerations

There is an intuitive explanation for the identification issues discussed above. In the traditional MDCEV model with a γ-profile, there is satiation in the outside good too as shown in Equation (17). Thus, the baseline preference for the outside good provides a marker regarding the discrete consumption decision, while the actual outside good consumption provides a second marker for

determining the intensity of consumption of the inside goods (because the KKT conditions for the consumed goods imply that the marginal utility at the optimal consumed values of the inside goods should equal the marginal utility at the consumed value of the outside good). With the linear outside good utility and the γ-profile, the baseline preference for the outside good serves as a single marker for the discrete consumption decision. In this situation, a second marker is needed to determine the continuous consumptions of the consumed inside goods, which is obtained by either setting set the scale to 1 (or, equivalently, $\alpha \to 0$) as we have done above, or by setting the γ_k parameter for one of the inside goods (which is consumed at least by some individuals in the sample) to an arbitrary value such as one. Effectively, both of these normalizations constrain the satiation profile for one of the inside goods, which provides the second marker for continuous consumptions at the point where the marginal utility of this inside good (with a normalized satiation profile) is equal to the baseline preference for the outside good.

A few important summary notes regarding the above identification discussion in relation to the linear outside good MDCEV model (back to the exclusive γ-profile). First, the analyst should attempt to estimate a model with a free scale and a full set of γ_k parameters for the inside goods. In most cases, such an estimation will fail. Then, the analyst can either normalize the scale parameter or normalize the γ_k parameter for one of the inside goods, and pick the one that provides a better data fit (in our experience, it will be the one that normalizes the scale parameter). Second, the condition above related to the inestimability of the scale parameter in the absence of price variation holds even if the satiation parameter is parameterized as a function of individual characteristics (this can be observed using the same strategy as above). Third, to be complete, we must state again that the scale parameter is immediately identifiable in the presence of price variation, even in the linear outside good MDCEV model with the γ-profile utility functional form. Fourth, in the case of more advanced MDC models with a linear outside good profile that allow for a heteroscedastic error specification, the scale of one of the alternatives has to be set to unity with no variation in unit prices across alternatives (similar to the case of the heteroscedastic extreme value or HEV model of Bhat, 1995). With a general error structure and no variation in unit prices, the identification considerations associated with a standard discrete choice model with correlated errors apply (see Train, 2003; Chapter 2). Finally, issues of

identification in the context of the MDCEV model and other MDC variants are nuanced, and are highly dependent on the specific utility profile used within the MDC framework. But, with the field moving clearly toward the use of a γ-profile utility form within MDC contexts, more definitive and clear guidelines are now available based on this paper and the Bhat (2008, 2018) papers.

3. FORECASTING

The forecasting approach in the grouped consumption model with an unobserved budget may be done in a manner similar to, but different from that described in Bhat (2018) (the procedure in Bhat, 2018 applies to the case where the continuous baseline preference is completely different from the discrete baseline preference, and there are two stochastic terms for each good, one in the discrete baseline preference and the other in the continuous baseline preference; on the other hand, in the current model, a single baseline preference exists and a single stochastic effort term applies for each consumer across both the discrete and continuous baseline preferences). The approach is also different from the one proposed in Pinjari and Bhat (2011), which applies to the traditional MDC model and that is generally more complicated than the procedure that can be employed in the current model.

In the specific case of the current model, the forecasts can be made for continuous consumptions or for the grouped consumptions.

3.1. Forecasting Procedure for Continuous Consumptions

The forecasting approach depends on whether the analyst wants to impose some upper bound on the budget or not (in either case, the model still considers the budget as being unobserved; it is just that there is the possibility of some consumers being predicted to consume a very high and unrealistic continuous consumption if an upper bound on the budget is not imposed). The upper bound may be determined based on what is considered reasonable in a specific setting, or may be obtained using methods such as a stochastic frontier approach (see Pellegrini et al., 2020 and Pinjari et al., 2016).

3.1.1. No Upper Bound on the Budget

The KKT conditions of Equation (11) for the inside goods (k=2,3,…,K) translate to the following conditions on the error terms:

$$\varepsilon_1 < \varepsilon_k - \tilde{V}_{k0} \text{ and } x_k^* = \left[\exp(\varepsilon_k - \varepsilon_1 - \tilde{V}_{k0}) - 1\right] \gamma_k \text{ if } x_k^* > 0 \quad (21)$$

$$\varepsilon_1 > \varepsilon_k - \tilde{V}_{k,0} \text{ if } x_k^* = 0$$

The simplest forecasting procedure (Procedure A) for each observation is as follows:

- Step 1: Draw K independent realizations of ε_k (say μ_k), one for each good k ($k = 1,2,...,K$)), from the extreme value distribution with location parameter of 0 and the scale parameter equal to the estimated σ value (label this distribution as EV(0, $\hat{\sigma}$)).
- Step 2: If $\mu_1 < \mu_k - \tilde{V}_{k,1}$, declare the inside good as being selected for consumption ($d_k = 1$); otherwise, declare the inside good as not being selected for consumption ($d_k = 0$).
- Step 3: For the inside goods that are selected ($d_k = 1$), forecast the continuous value of consumption as follows: $x_k^* = \left[\exp(\mu_k - \mu_1 - \tilde{V}_{k0}) - 1\right] \gamma_k$.

A problem with the forecasting procedure above is that the predictions will have high variance (depending on the single realization of error terms taken for each observation). The one time that this may not be much of a problem is if the prediction is being done on a very large synthetic population of interest. A second approach (say, Approach B) is then to repeat steps 1 through 3 above for many sets of realizations. Count the number of times each of the possible $(2^{K-1} - 1)$ combinations of discrete consumption of the inside goods appear as the chosen combination. Also, estimate the probability P_n of each discrete consumption combination n as the number of times it appears as the chosen combination relative to the total number of sets of realizations. Next, for each combination n (n=1,2,…,N, N=$2^{K-1} - 1$), compute the mean value $\bar{x}_{kn}^*$ of the continuous consumption values across the many realizations. Finally, forecast the continuous amount of consumption for each alternative k as $x_k^* = \sum_n P_n \bar{x}_{kn}^*$. This approach will provide more accurate aggregate-level predictions (that is, predictions of consumption quantities across multiple individuals) than the first approach with small forecasting samples. But, for a given individual, given enough number of sets of realizations, it will always forecast a positive value of consumption for each and every alternative.

A third approach (Approach C), somewhere in-between the two approaches above in terms of computation time, is to first use Equation (14) to compute the discrete probability P_n for each combination *n*, then use the usual discrete probability-to-deterministic choice procedure (used in traditional simulation approaches) to determine the most likely market basket of consumption, and forecast the consumption quantities for this single market basket. Specifically, the procedure is as follows

- Step 1: Use Equation (14) to compute the discrete consumption probability for each possible consumption bundle *n*.
- Step 2: Order the combinations from 1 to *N* in an arbitrary order (but retain this from hereon), and, for each combination *n* up to the penultimate combination (*n*=1,2,…,*N*–1), obtain the cumulative probability from combination 1 to combination *n* as $CP_n = \sum_{d=1}^{n} P_d$.
- Step 3: Partition the 0-1 line into *N* segments (each corresponding to a specific combination *n*) using the (*N*–1) CP_n values. Draw a random uniformly distributed realization from {0,1} and superimpose this value over the 0-1 line with the *N* segments. Identify the segment where the realization falls, and declare the combination corresponding to that line segment as the deterministic discrete event of consumption for the individual.
- Step 4: For the specific combination declared as the discrete bundle of consumption from Step 3, forecast the continuous consumption as follows. Draw an independent realization of ε_1 (say μ_1) for the outside good. For each of the consumed goods in the bundle, draw a realization of ε_k (say μ_k) from $\text{EV}(0,\hat{\sigma})$ truncated from below at $\mu_1 + \tilde{V}_{k,1}$ (that is, such that $\mu_k > \mu_1 + \tilde{V}_{k,1}$). Predict the continuous consumption value for the consumed goods as: $x_k^* = \left[\exp(\mu_k - \mu_1 - \tilde{V}_{k0}) - 1\right] \gamma_k$ and set $x_k^* = 0$ for the non-consumed goods. A variant of this step (4) would be to repeat step (4) multiple times with different sets of realizations, and take the mean across the resulting x_k^* predictions.

3.1.2. Upper Bound Imposed on Budget

There may be forecasting situations where the analyst may want to bound the total consumption budget possible, based on what is feasible or what is reasonable. For example, in the case of a

daily time-use case, the feasible budget would be 24 hours. In the case of annual miles driven, based on the estimation sample and other information, an upper bound of 100,000 miles may be imposed by the analyst. In such instances, the forecasting approach needs to be modified, because the discrete and continuous consumption patterns get a little more intertwined in the prediction process. In particular, it should be true that the sum of the continuous consumptions in the inside goods should be less than the externally provided upper bound of budget. But the first and second approaches for the case of no upper bound will not apply here because the draws for the consumed inside goods and the outside good get inter-related through the upper bound of the budget, but these draws also dictate which goods are consumed and which goods are not consumed at the discrete level. The forecasting approach (say, Approach D) in this "upper budget bound" case then is similar to Approach C for the "no upper budget bound case, and is as follows:

- Step 1: Follow Steps 1,2, and 3 from Approach C of the previous section.
- Step 2: For each of the consumed inside goods in the combination from Step 1 (arranged such that the first M consumed goods appear first; k=2,3,…,M+1), draw an independent realization of ε_k (say μ_k) from EV$(0,\hat{\sigma})$. If no inside goods are consumed (M=0), proceed to Step 4.
- Step 3: Compute $H_{k,0} = \mu_k - \tilde{V}_{k,0}$ for k=2,3,…,M+1. Then, identify the minimum (say R^1) of the $H_{k,0}$ values across these consumed inside goods (there is no need to compute R^1 if the combination from step 1 corresponds to no inside good being consumed)
- Step 4: Draw an independent realization μ_k (k=M+2,M+3,…,K) now for each of the K-M-1 non-consumed goods in the combination from step 1 from EV$(0,\hat{\sigma})$, truncated from above at R^1 if M>1 (that is, such that $\mu_k < R^1$ for the non-consumed goods) and untruncated if M=0.
- Step 5: Compute $H_{k,0} = \mu_k - \tilde{V}_{k,0}$ for k=M+1,M+2,…,K. Then, identify the maximum (say R^2) of these $H_{k,0}$ values across these non-consumed inside goods. Ignore this step if all inside goods are consumed.

- Step 6: For combinations of some goods being consumed and others not, determine the maximum of R^2 and $\ln\left(\frac{\sum_{k=2}^{M+1}\left[\exp(\mu_k - \tilde{V}_{k0})\right]}{E + \sum_{k=2}^{M+1}\gamma_k}\right)$. Label this as R^3. Draw a realization μ_1 for the first outside alternative from the lower truncated univariate extreme value distribution (again with the extreme value distribution being $\text{EV}(0, \hat{\sigma})$) such that $\mu_1 > R^3$. For the combination corresponding to all of the inside goods being consumed, draw a realization for the first outside alternative from the singly truncated (from below) univariate extreme value distribution such that $\mu_1 > \ln\left(\frac{\sum_{k=2}^{M+1}\left[\exp(\mu_k - \tilde{V}_{k0})\right]}{E + \sum_{k=2}^{M+1}\gamma_k}\right)$.

A variant of the procedure above would be to repeat step (2) through (6) multiple times with different sets of realizations, and take the mean across the resulting x_k^* predictions.

3.2. Forecasting Procedure for Grouped Consumptions

The simplest forecasting approach (Approach E) for each observation in this case is as follows:

- Step 1: Draw *K* independent realizations (say μ_k), one for each good k ($k = 1, 2, ..., K$), from the extreme value distribution with location parameter of 0 and the scale parameter equal to the estimated σ value (label this distribution as $\text{EV}(0, \hat{\sigma})$).
- Step 2: If $\mu_1 < \mu_k - \tilde{V}_{k,0}$, declare the inside good as being selected for consumption ($d_k = 1$); otherwise, declare the inside good as not being selected for consumption ($d_k = 0$).
- For the inside goods that are consumed (based on Step 2), if $\mu_1 + W_{k,c_k-1} < \mu_k < \mu_1 + W_{k,c_k}$, declare the inside good as being selected for consumption ($d_k = 1$) with a grouped consumption value of c_k; otherwise, declare the inside good as not being selected for consumption ($d_k = 0$).

An alternate procedure (Approach F) is similar to Approach C. First predict the discrete probability P_n for each combination *n*, then translate this to a deterministic prediction of the

market basket of consumption, and forecast the consumption quantities for this single market basket. Specifically, the procedure is as follows:

- Step 1: Same as Steps 1, 2, and 3 of Approach C.
- Step 2: For the specific combination predicted as the discrete bundle of consumption from Step 1, forecast the grouped consumption as follows. Draw an independent realization μ_1 for the outside good. For each of the consumed goods in the bundle, draw a realization μ_k from EV$(0,\hat{\sigma})$ truncated from below at $\mu_1+\tilde{V}_{k,0}$ (that is, such that $\mu_k > \mu_1+\tilde{V}_{k,0}$). If $\mu_1+W_{k,c_k-1} < \mu_k < \mu_1+W_{k,c_k}$, predict a grouped consumption for the *k*th inside good as c_k.

A third procedure (Procedure G) is more direct. This would compute the multivariate probability for each grouped outcome for each good (including zero consumptions) using Equation (6). If a deterministic outcome is to be predicted, one can use the usual discrete probability-to-deterministic choice procedure (used in traditional simulation approaches) to determine the most likely market basket of consumption. The problem with this is that the number of possible combinations can get very high as the number of alternatives increase and/or the number of grouped categories for each alternative increases.

4. EMPIRICAL APPLICATION

4.1. Sample Description

To demonstrate applications of the MDGEV model, we consider two empirical cases. The first is the case of the time-use of individuals. We consider the 2000 San Francisco Bay Area Travel Survey (BATS) data (also used by Bhat, 2005), along with supplementary zonal-level land-use and demographics data for each of the Traffic Analysis Zones (TAZ) in the San Francisco Bay area. The dependent variable corresponds to individual-level time investments in social-recreational activities over a weekend day. Specifically, the total time invested during the weekend day in each of the following four activity purpose categories was computed based on appropriate time aggregation across individual episodes within each category: (1) time spent in in-home social activities (IHS), (2) time spent in in-home recreational (IHR) activities, (3) time spent in out-of-home social (OHS) activities, and (4) time spent in out-of-home recreational (OHR) activities. Details of the activity purpose classification are provided in Bhat (2005), but, generally speaking, social activity episodes included conversation and visiting family/friends,

and recreational activity episodes included such activities as hobbies, exercising, and watching TV. The sample for analysis includes the weekend day time-use information of 1917 individuals, which we partition into an estimation sample of 1500 individuals and a hold-out validation sample of 417 individuals. The analysis of interest is the participation and time invested in four types of discretionary activities over the weekend day: in-home social (IHS), in-home recreation (IHR), out-of-home social (OHS), and out-of-home recreation (OHR). These four activity purposes constitute the "inside" goods in our analysis. The outside good may be thought off here as the time spent in all other non-social and non-recreational activities during the weekend day. Interestingly, this data set did not show too much clustering as most time-use data sets do, and so we used the data set to examine the performance of clustering and to test the ability of the proposed MDGEV model to recover the estimates from an MDCEV estimation on the continuous data. We used increasing sized clustering to examine the effect of cluster size on the ability of the MDGEV model to recover accurate estimates of the variable effects. Specifically, we used clustering sizes of 15 minutes, 30 minutes, and 60 minutes, and estimated MDGEV models. For future reference, we label these models as MDGEV-M1 for the 15-minute cluster size, MDGEV-M2 for the 30-minute cluster size and MDGEV-M3 for the 60-minute cluster size. We also estimated a second MDCEV model (which we will label as the MDCEV-M4 model) that used the 60-minute cluster size observations, and assumed the continuous value of consumption to be the midpoint of the grouped category in which an individual's consumption fell.[9] In a way, we are using the real time-use data as simulated data to examine the effect of cluster size, and the effect of assuming midpoints of grouped categories as the continuous consumption values, on the ability of the model to recover variable effects (as assessed by closeness of estimated coefficients with those obtained from the MDCEV model). We also assess the ability of the models with different levels of coarseness in the groupings to predict the continuous values of time-use in both the estimation sample as well as a hold-out validation sample that we do not use in estimation.

[9] For the final time window category of 570 minutes and above, we assigned the continuous value of 750 minutes based on computing the mean of all the observed values higher than 570 minutes. Important to note here also is that this "midpoint" method, while convenient, is tantamount to assuming a uniform distribution for the η_k terms, which is fundamentally inconsistent with the structure of the MDCEV model (which assumes a logistic distribution for the η_k terms). We include this MDCEV-M4 model here simply for an empirical comparison, although the level of incorrectness due to the inconsistency of the uniform distribution assumption and the use of the MDCEV model will be very context-dependent and will depend on the size of the grouping windows. The narrower the width of the grouping windows, lesser will be the inconsistency.

The second demonstration is based on vehicle ownership and use data from the 2017 National Household Travel Survey in the state of Texas. This is a new data set with grouped consumptions constructed for the specific purpose of this paper. The vehicles owned by each household are categorized into one of five vehicle types: (1) Passenger cars (coupes, sedans, hatchbacks, crossovers, and station wagons), (2) Vans, (3) Sports Utility Vehicles (SUVs), (4) Pickup trucks, and (5) Other (non-pickup trucks and recreational vehicles). For this demonstration exercise, the final estimation sample includes 1375 Texan households with non-zero vehicle ownership, and who owned no more than one vehicle within each of the five vehicle types (if a vehicle type is owned at all). Of course, a household might own multiple vehicle types. A separate hold-out validation sample of 403 Texan households was also created. The MDC variable corresponds to ownership of each vehicle type and the amount of annual miles on each vehicle type.[10] We noticed that the annual mileage of vehicles had a distinct clustering at the multiples of 1000 miles, indicating clearly that household reporting of mileage is in grouped form (indeed, many surveys explicitly recognize this issue and seek annual mileage in grouped form rather than a continuous form). In fact, about 70% of the respondents reported their annual miles in multiples of 1000 (the clustering tends to be at the multiples of 5000 for higher mileage reporting). Thus, in our grouped mileage estimation, we considered the dependent variable to be clustered in mileage windows of 1,000 until a reported mileage of 20,000, beyond which we used a mileage window of 5,000. Table 1 provides information on the distribution of vehicle types in the vehicle-use dataset, assuming midpoint mileage for the intermediate windows and a mileage of 750 for the first mileage window (of 0-1000 miles) and a mileage of 75,000 for the highest mileage window (of 35,000-200,000 miles) (this table summarizes the statistics across the estimation and validation samples, for a total of 1778 households). The table indicates that most of the one-vehicle households own passenger cars (about 56% of one-vehicle households) or SUVs (29% of one-vehicle households). The percentage of one-vehicle households holding pickup trucks and vans is about 10.5% and 4.2%, respectively. However, as one would expect, the percentage of pickup trucks and vans in the mix increases within households with more than one vehicle. Across all households, it is clear that passenger cars are the most likely to be represented in the vehicle mix of households. Specifically, adding across columns for the

[10] The outside good may be thought of here as the miles traveled by non-motorized and other non-private motorized modes.

"passenger car" row of Table 1, it is observed that 1,138 of the 1,778 (64%) of households hold a passenger car. Besides passenger cars, SUVs are also relatively likely to be held by households, with 825 of the 1,778 (46%) households owning an SUV. At the other end, vans and other types of vehicles (non-pickup trucks and recreational vehicles) are the least likely to be present in household vehicle fleets, with only 163 (9.2%) households owning vans and only 95 households (5.3%) owning non-pickup/recreational vehicles. In terms of vehicle-use, the last column of Table 1 indicates that SUVs tend to be the most widely used if held by a household, followed by pickup trucks and passenger cars.

4.2. Model Specifications and Performance Evaluation

In both of the case studies, the emphasis is on demonstrating the application of the proposed model rather than necessarily on substantive interpretations and policy implications. But, within the context of the data available, we explored alternative variable specifications to arrive at the best possible specification (including considering alternative functional forms for continuous independent variables such as income and age, including a linear form, piecewise linear forms in the form of spline functions, and dummy variable specifications for different groupings). The final variable specification was based on statistical significance testing as well as intuitive reasoning based on the results of earlier studies. For both the demonstration case studies, and as discussed earlier in Section 2.3, we normalize the scale of the error terms to one. Also, while not the express focus of our empirical analyses, we do provide brief discussions of the results for completeness purposes. For the time-use case, we present the substantive results only for the MDGEV model with 15-minute clustering size (labeled the MDGEV-M1 model), because there were little differences in the variable effects across the differently clustered MDGEV and the MDCEV models. For the vehicle-use demonstration, the dependent variable is intrinsically clustered, and so only the one MDGEV model is estimated and reported. Note also that, in the specifications, we allow heterogeneity across individuals due to observed variables not only in the baseline preference function (the ψ_k function as in Equation (3)), but also in the satiation parameters (the γ_k parameters). Doing so acknowledges that the intensity of satiation for a particular alternative may vary across individuals, and also allows for additional flexibility in allowing the discrete choice of consuming an alternative to be less closely tied to the continuous choice of the amount of consumption of that alternative (see Bhat, 2008). This is particularly

useful when imposing a linear baseline preference for the outside good. The constraint that $\gamma_k > 0$ for $k=2,\ldots,K$ is maintained by reparametrizing γ_k as $\exp(\delta_k'\omega_k)$, where ω_k is a vector of decision maker-related characteristics and δ_k is a vector to be estimated.

The purpose of our proposed model is to accommodate the case of intrinsically grouped consumption data in multiple discrete situations as well as unobserved budgets. To test the ability of our proposed model to provide a good data fit in such situations, we examine the performance of our model on both the estimation sample as well as a separate holdout validation sample. Of course, there is no clear baseline model to compare the model results with, because earlier MDC models are applicable for the case of continuous consumptions and explicitly provided budgets. But, in our time-use empirical example, we do have the reported continuous consumption values. Thus, for comparison purposes, we also estimate a model based on the reported continuous values while maintaining the final variable specification obtained from our proposed model. These correspond to the linear outside good MDCEV model (which we will label henceforth simply as the MDCEV model) and compare these with the proposed MDGEV formulation. For the time-use case study, the evaluation of data fit is based on the ability to predict the combined multiple discrete plus continuous observed consumption component (MDC component) of consumption as well as, separately, the discrete component (MD) of consumption (whether an alternative is consumed at all or not). For the vehicle type/use case study, the evaluation of data fit is based on the ability to predict the combined multiple discrete plus grouped observed consumption component (MDG component) of consumption and the discrete component (MD) of consumption. The performance metrics include likelihood-based data fit measures as well as non-likelihood based data fit measures, and on both the estimation sample as well as the hold-out validation sample.

4.2.1. Likelihood-Based Data Fit Measures

In the time-use sample, we cannot directly compare the log-likelihood values at convergence of the different MDGEV models estimated with different cluster sizes and the MDCEV-M4 model estimated using midpoint continuous values. So, we compute an effective predictive log-likelihood of all the estimated MDGEV/MDCEV-M4 models (as well as the corresponding MDCEV model estimated on the continuous values) at the common platform of the observed multiple discrete-continuous (MDC) values (that is, using Equation (6)). We also compute the

log-likelihood with only the constants in the baseline preferences and only the constants in the satiation parameters, using the MDCEV model. We then compute values of an effective nested likelihood ratio test relative to the constants only likelihood of the MDCEV model, to test if all the models provide similar values of the resulting test (it is true that the closeness of the predictive likelihood values across the different models will immediately provide an intuitive sense of the performance of the different models, but we compute the effective nested likelihood ratio test value to examine the ability of the different models to show statistically significant improvement over the simple constants only specification). We also compute a predictive Bayesian Information Criterion (BIC) values[= $-\mathcal{L}(\hat{\theta})$ + 0.5(# of model parameters) log(sample size)] with respect to the continuous observations ($\mathcal{L}(\hat{\theta})$ is the log-likelihood at convergence). All of the above metrics correspond to the MDC component of fit. We then use the estimated values from the MDGEV/MDCEV-M4 models and the MDCEV model to predict the purely discrete component (MD component) of fit using Equation (13) and compute corresponding predictive log-likelihood function and information criterion values. We then compute the log-likelihood at constants only for the pure discrete component (using the actual discrete shares of the many multiple discrete combinations), and compute an informal nested likelihood ratio test value for the discrete component (technically speaking, this is only an informal test because the likelihood is maximized for the continuous consumptions, not the discrete consumptions). At this discrete level, we also compute an informal "Adjusted likelihood ratio index" ($\bar{\rho}^2$) for each of the MDGEV, MDCEV-M4, and MDCEV models as:

$$\bar{\rho}^2 = 1 - \frac{\mathcal{L}(\hat{\theta}) - M}{\mathcal{L}(C)}, \tag{22}$$

where $\mathcal{L}(\hat{\theta})$ is the predictive log-likelihood function at convergence for the purely discrete component, and $\mathcal{L}(C)$ is the log-likelihood function at constants, also only for the purely discrete component. *M* is the number of parameters (not including the constants appearing in the baseline preference). For the hold-out validation sample, none of the statistical tests discussed above hold, so we simply compute a predictive likelihood value and the Bayesian Information Criterion using the model estimates for the observed MDC and MD choices.

For the vehicle type and use case study, we have the grouped consumption values. So, we redo the same analysis as in the time-use study, except that all the computations as above are

undertaken for a single MDGEV model (with the observed grouped consumption values) and the fit measures are computed on the estimation sample at the level of the observed grouped values of consumption.

4.2.2. Non-Likelihood Based Data Fit Measures

To further supplement the disaggregate likelihood-based performance at the multivariate and disaggregate levels, we evaluate the performance of the MDGEV models intuitively and informally at a disaggregate and aggregate level. Since these non-likelihood based data fit measures are more easily undertaken at discrete/grouped consumption levels, at the disaggregate level, we estimate the probability of the observed MDG outcome for each individual, and compute an average probability of correct prediction for the MDG outcome in both the estimation and hold-out samples. For the time-use case study, to keep the presentation simple, we undertake this analysis only for the MDGEV model estimated at the finest level of grouping (we also do not present the results for the MDCEV-M4 model, again to keep the presentation manageable). This MDGEV at the finest grouping level corresponds to the model grouped at cluster size of 15 minutes (MDGEV-M1). A similar analysis is undertaken to obtain the average probability of correct prediction at the MD outcome level. At the aggregate level, we design an informal heuristic diagnostic check of model fit by computing the predicted MDG and MD components (in terms of aggregate share of individuals) for specific multivariate discrete outcomes. In this analysis, to keep things presentable and understandable (because the number of groupings per alternative is already very high, and the number of multivariate combinations across different alternatives explodes), we focus on a combined MDG and MD prediction through a simple trinary prediction, for each alternative, of whether an individual participates in that alternative, whether the participation is between 0+ and 120 minutes in the time-use case (between 0+ and 7500 miles in the vehicle use case), or whether the participation is over 120 minutes in the time-use case (over 7500 miles in the vehicle use case). These probabilistic predictions are easily obtained based on the property that the multivariate logistic distribution has the univariate logistic distribution as its marginal. We then compute the aggregate predicted values within each of the three categories for each alternative, and compare the predicted versus actual fractions within each category for each alternative using the mean absolute percentage

error (MAPE) statistic. The above procedure is implemented for both the estimation sample and the holdout sample.

4.3. Substantive Model Results

For completeness, we now discuss the substantive results from the MDGEV model (the substantive results from the MDCEV model were the same as that from the MDGEV model, and so the MDCEV model results are not presented here).

4.3.1. Time-Use Model

Table 2 provides the results for the proposed MDGEV model (MDGEV-M1) in the time-use context. In this section, we discuss the effects of the variables on the time-use activity participation by variable category. The labels used for the four activity types are IHS (In-Home Social), OHS (Out-of-Home Social), In-Home-Recreational (IHR), and Out-of-Home Recreational (OHR). The effects relate to the impact of variables on the logarithm of the baseline preference (that is, they correspond to the $\boldsymbol{\beta}$ vector elements in Equation (2)), except when discussing the satiation effects toward the end of this section.

Household sociodemographic: The children variables indicate that individuals in households with very young children (0-4 years of age) are more likely to participate in out-of-home social activities over the weekend relative to their peers, presumably a result of wanting a break from in-home child-rearing responsibilities during the extended weekend home stay. Participation in out-of-home social activity rather than out-of-home recreation perhaps offers a mechanism to remain with the child (which many parents may actually enjoy to a good extent, and that also dispenses with the need for child-care arrangements), while also potentially benefitting from the help in taking care of the child that an individual may receive from extended family members during social visits. Interestingly, the participation propensity for out-of-home social activities takes an abrupt turn in the presence of children in the age group of 5-15 years, with individuals more likely to participate in in-home activities and out of-home recreation rather than out-of-home social pursuits. This turn suggests a distinctly different lifecycle stage as children grow up from being an infant/toddler to being more independent. According to the human development literature, there are three possible reasons for this (see Batra, 2013 and Chiarlitti and Kolen,

2018). First, beyond the infant/toddler stage, children become easier to be around with because they become more communicative with language and can respond back. This reward in the form of distinct responses increases the propensity for the family to be together at home or to pursue outdoor recreation activities as a single family unit. Second, around five years of age (coincidentally this is also around the time most children begin kindergarten), children are transitioning from the developmental stage of "play age" (3-5 years of age), when their main network still revolves around the family unit, to "school age" (5-12 years), when their network radius expands to include school and community. During "play age", children are still in a very experimental phase; but during "school age", children are learning independence and self-coping skills in new environments. This likely leads to parents being in a less "monitoring" role, and becoming more comfortable to allow their children to spend time outside of their supervision in the form of "playdates" at classmates' homes. Thus, parents themselves spend less time in "out-of-home" social activities as children age, and spend more social time with the child within the home instead. A third reason is the increasingly structured activities undertaken by children over five years of age in the U.S. and elsewhere over the weekends, including those undertaken in-home (such as taking piano lessons or tutoring lessons) and participation in youth sports leagues.

The number of adults in a household has a positive impact on in-home recreation (relative to other activity purposes). Seo et al. (2013) also observe that larger households have more possibilities for mutually rewarding in-home leisure activities, significantly lowering their propensity to travel out-of-home. The effects of motorized vehicle availability and bicycles are as expected, and generally increase out-of-home activity participation. Of course, the one caveat here is that these variables may be endogenous, in that individuals who are more outdoor-oriented may be the ones who decide to own more number of vehicles and bicycles. These endogeneities can be handled by modelling time-use jointly with motorized vehicle ownership, residential location, and bicycle ownership, as undertaken by Pinjari et al. (2011). This suggests extension of joint models to include a new MDG variable, which we leave for future research.

Finally, within the group of household socio-demographics, a higher household income leads to more out-of-home leisure activities, a result that is not surprising given that social and recreational activities (especially the latter) have a financial cost (of transportation and goods/services consumption) associated with them, and higher income households are better positioned to absorb these costs (see Highfill and Franks, 2019 and Parady et al., 2019).

Household location attributes: Several variables pertaining to locational attributes were tested, out of which only land-use mix diversity turned out to be statistically significant. Land-use mix diversity variable is computed as a fraction between 0 and 1 for each traffic analysis zone of the San Francisco Bay area (see Bhat and Gossen, 2004; Bhat, 2005). Zones with a value closer to one have a richer land-use mix than zones with a value closer to zero. Three categories of land-uses are considered in the computation of the mix diversity variable: acres in residential use (r), acres in commercial/industrial use (c), and acres in other land-uses (o). The actual form of the land-use mix diversity variable is:

$$\text{Land-use mix diversity } = 1 - \left\{ \frac{\left(\left| \frac{r}{D} - \frac{1}{3} \right| + \left| \frac{c}{D} - \frac{1}{3} \right| + \left| \frac{o}{D} - \frac{1}{3} \right| \right)}{(4/3)} \right\}, \tag{23}$$

where $D = r + c + o$. The functional form assigns the value of zero to zones in which land-use is focused in only one category, and assigns a value of 1 to zones in which land-use is equally split among the three land-use categories. The results in Table 2 indicate that individuals residing in areas with high land-use mix diversity tend to have a lower preference for OHR activity and higher preference for IHR activity during the weekend. This is admittedly a little difficult to explain, but may also be an artifact of the spatial scale used in the computation (that is, finer spatial resolutions rather than traffic analysis zones may be at play in the effect of land-use mix on activity-travel behaviour; see Guo and Bhat, 2004).

Individual characteristics: Among individual characteristics, women appear to be less likely to pursue in-home recreation over the weekends, while age appears to have a dampening effect on out-of-home recreation pursuits. The age effect is to be expected, as older individuals are likely to have physiological constraints that hamper mobility and lead to more in-home recreational activity pursuits such as reading and watching TV. In fact, Paillard-Borg et al., 2009 find that reading and other activities at home that involve some level of mental stimulation dominate the time-use of older adults. Further, individuals above the age of 65 years are more likely to be retired, and thus have a more limited social network. As a result, weekend recreational activities engendered through social networks (such as going bowling or going hiking) also gets limited. In addition, the human development literature identifies different network correlates of social and

emotional loneliness between young and older adults (Green et al., 2001). While younger adults appear to revel in the size of their social networks and look to pursue outdoor recreational activities (with members of their expansive social network) as a means to reduce loneliness, older adults appear to prefer close (and small-sized) social networks and relationships within the home (such as with an intimate partner or a romantic relationship or a deep trusted friend) to alleviate feelings of loneliness. Moving on to other individual variables, employed individuals are found to have a lower baseline preference for IHS activity during the weekends compared to unemployed individuals, presumably a reflection of the desire to catch up on some private recreational in-home activities and/or outdoor activities rather than entertaining guests within homes on non-work days. The ethnicity variable is also found to impact the extent of activity participation during the weekend – Hispanics are found to have a higher baseline preference for OHS activity, which is consistent with a vast literature in family science that suggests a closer-knit extended family and community unit of socialization among non-Anglo speaking cultures (see, for example, Pernice-Duca, 2010 and Viruell-Fuentes et al., 2013).

Day of the week and seasonal effects: The higher preference for in-home recreation on Sundays relative to Saturdays is consistent with the notion that individuals treat Sundays as an in-home rest and "chill" day as they get recharged for the coming week. The winter season is associated with a higher propensity (relative to other seasons) for out-of-home social activity, while both the fall and winter seasons are associated with a lower preference for out-of-home recreation activity. The former result is a clear reflection of the festive season of getting together with family and friends, while the latter finding is to be expected as the weather in Spring and Summer is more favorable for outdoor recreational activities compared to winter and fall in the San Francisco Bay area.

Baseline preference constants: The constants have no substantive interpretation because of the presence of the continuous land-use mix variable. But, loosely speaking, the constants reflect an overall lower preference for in-home social activity and higher preference for out-of-home recreation during the weekends, consistent with the general descriptive statistic that only 6.3% of individuals participate in in-home social activities, while close to 35% of individuals participate in out-of-home recreation.

Satiation effects through γ_k parameters: As indicated in Section 4.2, to allow heterogeneity in the parameters across individuals, while also guaranteeing the positivity of the parameters, these were parameterized as $\gamma_k = \exp(\delta_k' \omega_k)$. The estimates in Table 2 for the satiation effects correspond to the elements of the δ_k vector. A positive value for a δ_k element implies that an increase in the corresponding element of the ω_k vector increases γ_k, which has the result of reducing satiation effects and increasing the continuous consumption quantity of alternative *k* (conditional on consumption of alternative *k*). On the other hand, a negative value for a δ_k element implies that an increase in the corresponding element of the ω_k vector decreases γ_k, which has the result of increasing satiation effects and decreasing the continuous consumption quantity of alternative *k* (conditional on consumption of alternative *k*). In our final specification, two exogenous variables turned out to be marginally statistically significant, both associated with the out-of-home social (OHS) activity purpose. In particular, individuals in large households (relative to their peers) invest shorter times in out-of-home social pursuits if they participate in such pursuits. In addition, out-of-home social pursuits on Sundays are shorter than on Saturdays. Finally, a comparison of the constants across the activity purposes indicates the higher satiation rates (lower time durations) in out-of-home recreation and out-of-home social pursuits (especially after adding up the constant and the coefficients on "household size and "Sunday" for the out-of-home activity purpose), relative to the in-home activity purposes.

4.3.2. Vehicle-Use Model

Table 3 presents the results for the vehicle use case study. Again, the effects relate to the impact of variables on the logarithm of the baseline preference, except when discussing the results specific to satiation effects toward the end of this section. The five vehicle type alternatives are (1) Passenger car, (2) Van, (3) SUV, (4) Pickup truck, and (5) Other.

Household sociodemographic: Among the set of household sociodemographic variables, the effect of annual household income in Table 3 indicates that low income households (less than $35,000 annual income) are more likely to own vans, while high income households (with more than $125,000 annual income) are more likely to own passenger cars, SUVs, and pick-up trucks.

The association of SUVs with high income households is particularly discernible in the results. This is not surprising, because most of the luxury vehicles reside within the SUV class, and SUVs are known to be gas-guzzlers that require quite a bit of fuel cost outlays (see Bhat et al., 2009).

The presence of many children in the household leads to a strong preference for vans relative to other types of vehicles, which is to be expected because vans are more spacious, safe, and comfortable for travel with small children. Also, this may be a collective vehicle buying strategy of a group of households with children so that carpooling arrangements to transport children become possible in an efficient and mutually beneficial manner. In addition to the effect of children on the preference for vans, the results also indicate that households with more individuals prefer vans to other vehicle types. On the other hand, the results in Table 3 indicate that households with many workers are disinclined to own vans, and prefer to have passenger cars. With a dispersed set of work locations, it stands to reason that households would make the conscious decision to own more passenger cars that provide better fuel efficiency for the commute trips of the many workers (see Clark et al., 2016 for a similar result). Interestingly, the race of the household is also found to impact vehicle-type holding and usage, even after controlling for income effects. Specifically, White households are clearly much more likely to own pickup trucks and other vehicle types (other truck types/recreational vehicles). This is consistent with a recent study by a digital marketing firm, which observes that about 75% of the purchases of the top five pickup trucks (especially the #1 selling Ford F-150 vehicle in the pickup class) are by White households.[11]

Household location attributes: Households in locations with high population density (more than 4000 persons per square mile) have a higher preference for passenger cars than those in less dense areas. This result may reflect the relative ease of maneuverability afforded by smaller vehicles in highly dense travel areas, especially in the context of parking and keeping within relatively narrow lanes when driving. The other household location-related result in Table 3 regarding the higher inclination of households residing in less dense employment locations to own pickup trucks may simply be a consequence of individuals in such households more likely

[11] See https://hedgescompany.com/blog/2018/10/pickup-truck-owner-demographics/, accessed May 2, 2020.

to be self-employed in farming and other related pursuits; in such contexts, being able to haul large-sized items and operate in relatively rugged terrain become desirable attributes in a vehicle.

Baseline preference constants: The baseline preference constants do not have any clear and substantive interpretations because of the presence of count variables (such as number of adults, children, and workers in the household). But it is illustrative to note that, by far, the highest negative constants are associated with vans and other (non-pickup trucks/recreational) vehicle types, conforming to their very low representation in household vehicle fleets, as discussed earlier in the context of Table 1.

Satiation effects through γ_k parameters: The results for satiation (lower panel of Table 3) reveal that, conditional on ownership, lower income households put less mileage on passenger cars and pickup trucks (relative to SUVs, given the very low ownership of vans and non-pickup/recreational vehicles in the sample). That is, should a lower income household own both a passenger car and an SUV, or both a pickup truck and an SUV, it will put more mileage on the SUV in both cases. While this may seem counter-intuitive, the model is actually reflecting the reality that, even though very low income households have only a small probability of owning an SUV (as discussed earlier in the estimates of the baseline preference), an SUV tends to be used much more than other vehicle types if it is actually owned (and this is the case across all income categories). Thus, given the very low baseline preference for SUVs among low income households, and the fact that the baseline preference not only dictates the discrete consumption choice, but also serves as the basis from which satiation effects start operating, the model lowers the satiation parameter for low income households to ensure that SUV use tends to be still high when owned. Other results from Table 3 indicate the higher use of pickup trucks, conditional on ownership, among households with many workers, and the lower use of passenger cars, again conditional on ownership, among households residing in urban areas. The latter result is again the model simply trying to reconcile the high ownership of passenger cars in highly dense areas with the lower general use (by way of mileage) on passenger cars relative to pick-ups and SUVs. The constants related to the satiation parameters (the last row of Table 3) may be viewed as the satiation effects for households with no workers that reside in non-urban areas, and earn an income of U.S. $35,000 or more. For such households, a comparison of the magnitudes of the

satiation constants clearly implies the inclination to use SUVs the most and to use non-pickup trucks/recreational vehicles the least, all conditional on ownership. These results comport with the mileages in the final column of Table 1.

4.4. Data Fit Measures

As mentioned in Section 4.2.1 and Section 4.2.2, data fit measures are presented in two forms – likelihood-based data fit measures and non-likelihood based data fit measures, separately for the time-use and vehicle-use cases in the following sub-sections. The data fit measures are provided for both the estimation sample and the hold-out sample.

4.4.1. Time-Use Case

4.4.1.1. Likelihood based fit measures

The likelihood based data fit measures for the time-use case study are provided in Table 4. For the time-use case study, our emphasis is on investigating the performance of the grouped MDGEV models (MDGEV-M1, MDGEV-M2, and MDGEV-M3) and the "midpoint" MDCEV-M4 relative to that of the MDCEV model. While not presented here, we will point out that all the grouped models and the "midpoint" model recovered the variable coefficients accurately, with the overall APE (across all coefficients and with respect to the estimates from the MDCEV model) being 1.10%, 1.40%, 8.4%, and 1.25%, respectively for the MDGEV-M1, MDGEV-M2, MDGEV-M3, and MDCEV-M4 models (the actual model results for each of the four models are available in an online supplement to this paper at https://www.caee.utexas.edu/prof/bhat/ABSTRACTS/MDGEV/OnlineSupplement.pdf).

The predictive log-likelihood at convergence for all the models at the MDC level are almost exactly the same for both the estimation and hold-out samples, especially for the MDCEV, MDGEV-M1, and MDGEV-M2 models (see the first numeric row of Table 4). This supports the ability of the MDGEV models (at different levels of clustering) to predict the discrete-continuous consumption values well. Indeed, the robustness of our proposed method to different clustering sizes is remarkable (at least in this case study). The likelihood-based statistics and the Bayesian Information Criterion statistic from the different grouped models also are virtually identical (except for the MDGEV-M3 model, though even this exception fades for the hold-out sample), supporting the notion that all the models show statistically significant

improvement over the simple constants only specification. The likelihood data fit results for the MDCEV-M4 model indicate a definitively poorer performance (relative to the grouped models) for the MDC component in the estimation sample. Interestingly, though, in the current empirical context, the MDCEV-M4 model does as well as the grouped models (and even just a little better) in the hold-out sample (which can happen because of a chance occurrence).

The results for the pure discrete MD component (see the lower panel of Table 4) are similar to those from the MDC component, with literally no difference in the fit statistics across the MDCEV model, the many grouped models, and the MDCEV-M4 model in the estimation sample. The same holds true in the hold-out sample too, except for the clear poorer performance of the MDCEV-M4 model. Overall, however, the fit measures for the "midpoint" method can vary quite substantially based on the value assigned for the highest time window (that does not have a "midpoint" because of the open-ended nature of the upper time point) as well as the width of the time windows, as already discussed earlier. Thus, it is much more preferable to adopt our grouped and consistent model formulation rather than the convenient (but inconsistent) approach of assigning midpoint values.

In summary, the likelihood-based fit measures indicate strongly that the MDGEV models for the time-use case at different cluster sizes are able to effectively recover the actual parameters of the MDCEV model, as well as provide data fit measures that are literally unchanged from those of the MDCEV model. This suggests that our proposed MDGEV model is effective even at relatively large grouped windows of consumption, though there is some inevitable (though surprisingly limited) deterioration at very large grouped windows.

4.4.1.2. Non-likelihood based fit measures

The non-likelihood based fit measures for the time-use case are provided for the MDGEV level with the finest level of grouping (i.e. the MDGEV-M1 model). At the disaggregate level, we computed the average probability of correct prediction at both the MDG and MD levels. These values turned out to be 0.085 (for the MDG component) and 0.162 (for the MD component) in the estimation sample, with corresponding values of 0.064 and 0.142, respectively, for the hold out sample. While these may seem low, it must be observed that this is to be expected, given the number of different multivariate combinations that are possible. In particular, at the MDG level, there are of the order of 31 grouped intervals for each of the four activity purpose alternatives in

addition to the non-participation alternative, generating a total of 32^4 (=1,048,576) possible alternatives. An equal share model would, therefore, provide an average probability of correct prediction at the MDG level of $1/1{,}048{,}576=0.954*10^{-6}$, which is substantially lower than our model predictions on both the estimation and hold-out samples. At the MD level, the number of alternatives is more manageable at 2^4 (=16) possible discrete choice combinations. In our case, given the shares in each of the estimation and hold-out samples, the average probabilities of correct prediction for the sample shares model are 0.131 and 0.122, respectively, for the estimation and hold-out samples. Again, our MDGEV model outperforms the sample shares (constants only) model even on this metric of average probability of correct prediction.

The aggregate fit measures are presented in Table 5. As indicated earlier in Section 4.2.2, this measure is based on a heuristic check for the combined multiple discrete-grouped consumption in a trinary prediction context of an individual not participating in an activity (labeled as "0 minutes" in Table 5), or whether the participation is for 0^+ to 120 minutes, or whether the participation is for greater than or equal to 120 minutes. We report the aggregate predicted values for each of the three grouped categories and compare them with the observed number of individuals in each of these categories for ach activity type. For both the estimation and the hold-out samples, the prediction for whether or not an individual participates in an activity is the most accurate (with the weighted MAPE being about 3% and 4% respectively for the estimation and hold-out sample), followed by activity participation in the 0^+ to 120 minutes category and greater than 120 minutes category. The overall weighted MAPE values of 5.45% and 7.60% indicate convincingly accurate trinary predictions, reinforcing the efficacy of the grouped-consumption model.

4.4.2. Vehicle-Use Case

4.4.2.1. Likelihood based fit measures

The likelihood based data fit measures for the vehicle-use case are provided in Table 6, in a format similar to that of the time-use case. However, because the vehicle-use case study is intrinsically a grouped data situation, the issue of comparing different models does not arise. However, it is important to note that, at both the MDG and MD levels, the proposed MDGEV specification rejects the one with only constants, demonstrating the value of our variable

specification (even though the primary emphasis of this research is to demonstrate the application of the methodology, and not necessarily on the substantive behavioural results).

4.4.2.2. Non-likelihood fit measures

The average probability of correct predictions in the vehicle use case for the discrete-grouped (MDG) consumption component and for the purely discrete consumption (MD) component are found to be 0.0040 and 0.1073, respectively for the estimation sample (0.0031 and 0.0969 respectively for the hold-out sample). These values, especially the average probability of correct discrete-grouped consumption predictions, are even lower than the time-use case; this is because the vehicle-use case has five alternatives and 21 grouped intervals (including the case of non-ownership) for each alternative, leading to 21^5 (about 4,084,101) possible MDG alternatives. The average probability of correct prediction for a random prediction among these over 4 million MDG alternatives would yield a value of $0.245*10^{-6}$, which is far below our model value. At the MD level, there are five elemental alternatives for a total of 2^5 (=32) possible discrete choice combinations. Given the shares in each of the estimation and hold-out samples, the average probabilities of correct prediction are 0.0889 and 0.0771, respectively, for the estimation and hold-out samples, which are again lower than those obtained in our MDGEV model.

Table 7 provides the aggregate non-likelihood fit measures for the vehicle-use case. The aggregate fit measure is based on the simple trinary prediction (for each vehicle type) in the grouped categories of 0 miles (household does not hold a vehicle of the specific type), 0^+ to 7500 miles, and greater than 7500 miles. As before, we report the aggregate predicted values for each of the three grouped categories and compare them with the observed number of households in each of these categories across the vehicle-types. For both, the estimation and the hold-out samples, the discrete prediction of holding a specific vehicle type or not is the best (with weighted MAPEs of 3.92% for the estimation sample and 7.98% for the hold-out sample). The overall weighted MAPE values of 6.40% and 14.25%, respectively, for the two samples indicate reasonably accurate trinary predictions.

5. CONCLUSIONS

The traditional multiple discrete-continuous (MDC) model (which allows a satiation effect in the outside good) tightly links the discrete and continuous consumption quantities and also requires the knowledge of the budget. Bhat's (2018) flexible MDCEV model delinks the discrete-preference of a good from its continuous preference by allowing a linear utility profile for the outside good; however, this model is profligate in parameters. In this paper, we further explore the linear utility specification of the outside good while keeping a single consumption preference for the inside goods (which is a balanced trade-off between the traditional and the flexible MDCEV). In doing so, we discuss an important identification issue that is specifically relevant to such a model. The formulation also immediately allows us to incorporate the case of multiple discrete-grouped consumption using the same stochastic structure as before, and without the need for budget observations. Such grouped consumption observations are often encountered in many consumer survey data, including in time-use surveys, vehicle use surveys, consumption surveys, and scanner panel data of packaged product purchases, to identify just a few.

The paper proposes a closed-form multiple discrete-grouped extreme value (MDGEV) model by assuming the distribution of the error terms in the baseline utilities of the alternatives to be IID type-I extreme value. Forecasting methods for the model are proposed and discussed. To demonstrate the application of the model, we consider two case studies: (1) a time use case study involving individuals' leisure activity participation in four weekend activity purposes: in-home social, out-of-home social, in-home recreational, and out-of-home recreational, and (2) a vehicle-use case study involving household choice of vehicle type and use from among five vehicle body types – passenger car, van, SUV, pickup truck and other. For the time-use case study, we noticed relatively little clustering from a San Francisco Bay Area survey, and so used the data as a simulation engine to examine the effects of different levels of artificial clustering (15-minute, 30-minute, and 60-minute) on the ability of the MDGEV model to recover the "true" parameter estimates from an MDCEV model estimation on the underlying data. Based on a multitude of data fit measures, we found that all the MDGEV models are able to recover the "true" parameters quite well and perform well in data fit evaluations in both the estimation sample as well as a hold-out sample. For the vehicle-use case study, we demonstrated the application of an MDGEV model on the grouped mileage data.

To conclude, the proposed MDGEV should prove to be beneficial in a number of applications across disciplines, because many survey-related and other endogenous variables of interest are often reported or solicited in clusters or grouped categories. Of course, there is substantial scope for enhancing the simple closed-form model proposed in this paper, including relaxing the IID assumption across the error terms of alternatives, allowing for random coefficients (especially when there are alternative-specific variables available), and extending the MDGEV model to a flexible MDGEV model with distinct baseline preferences for the discrete and grouped consumptions (similar to the move from the MDCEV model to the flexible MDCEV model). However, any of these directions of enhancement will inevitably involve non-closed form probability structures and/or become profligate in model parameters. If the popularity of the simple closed-form MDCEV model in research applications is any indication (relative to the number of research applications of advanced variants of the MDCEV model), we would expect (or, at least, we certainly would hope) that our proposed simple closed-form MDGEV model of this paper would open up a new world of application possibilities in the context of multiple discrete-grouped consumption data.

ACKNOWLEDGMENTS

This research was partially supported by the U.S. Department of Transportation through the Data-Supported Transportation Operations and Planning (D-STOP) Tier 1 University Transportation Center (Grant No. DTRT13GUTC58). The authors are grateful to Lisa Macias for her assistance in formatting this document, and appreciate the comments of an anonymous reviewer on an earlier version of the paper.

REFERENCES

Batra, S., 2013. The psychosocial development of children: Implications for education and society — Erik Erikson in context. *Contemporary Education Dialogue*, 10(2), 249-278. DOI: 10.1177/0973184913485014

Bhat, C.R., 1995. A heteroscedastic extreme value model of intercity mode choice. *Transportation Research Part B*, 29(6), 471-483.

Bhat, C.R., 1996. A hazard-based duration model of shopping activity with nonparametric baseline specification and nonparametric control for unobserved heterogeneity. *Transportation Research Part B*, 30(3), 189-207.

Bhat, C.R., 2005. A multiple discrete-continuous extreme value model: Formulation and application to discretionary time-use decisions. *Transportation Research Part B*, 39(8), 679-707.

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*, 42(3), 274-303.

Bhat, C.R., 2018. A new flexible multiple discrete-continuous extreme value (MDCEV) choice model. *Transportation Research Part B*, 110, 261-279.

Bhat, C.R., and Gossen, R., 2004. A mixed multinomial logit model analysis of weekend recreational episode type choice. *Transportation Research Part B*, 38(9), 767-787.

Bhat, C.R., and Sen, S., 2006. Household vehicle type holdings and usage: An application of the multiple discrete-continuous extreme value (MDCEV) model. *Transportation Research Part B*, 40(1), 35-53.

Bhat, C.R., Astroza, S., Bhat, A.C., and Nagel, K., 2016. Incorporating a multiple discrete-continuous outcome in the generalized heterogeneous data model: Application to residential self-selection effects analysis in an activity time-use behavior model. *Transportation Research Part B*, 91, 52-76.

Bhat, C.R., Sen, S., and Eluru, N., 2009. The impact of demographics, built environment attributes, vehicle characteristics, and gasoline prices on household vehicle holdings and use. *Transportation Research Part B*, 43, 1-18.

Born, K., Yasmin, S., You, D., Eluru, N., Bhat, C.R., and Pendyala, R.M., 2014. Joint model of weekend discretionary activity participation and episode duration. *Transportation Research Record: Journal of the Transportation Research Board*, 2413, 34-44.

Castro, M., Eluru, N., Bhat, C.R., and Pendyala, R.M., 2011. Joint model of participation in nonwork activities and time-of-day choice set formation for workers. *Transportation Research Record: Journal of the Transportation Research Board*, 2254, 140-150.

Chiarlitti, N.A., and Kolen, A.M., 2018. Are children and their parents more active when children engage in more structured activities? *International Journal of Exercise Science*, 11(5), 106-115.

Clark, B., Lyons, G., and Chatterjee, K., 2016. Understanding the process that gives rise to household car ownership level changes. *Journal of Transport Geography*, 55, 110-120.

Deaton, A., and Muellbauer, J., 1980. *Economics and Consumer Behavior*. Cambridge University Press, Cambridge.

Garikapati, V.M., Sidharthan, R., Pendyala, R.M., and Bhat, C.R., 2014. Characterizing household vehicle fleet composition and count by type in integrated modeling framework. *Transportation Research Record: Journal of the Transportation Research Board*, 2429, 129-137.

Green, L.R., Richardson, D.S., Lago, T., and Schatten-Jones, E.C., 2001. Network correlates of social and emotional loneliness in young and older adults. *Society for Personality and Social Psychology*, 27(3), 281-288. https://doi.org/10.1177/0146167201273002

Guo, J.Y., and Bhat, C.R., 2004. Modifiable areal units: Problem or perception in modeling of residential location choice? *Transportation Research Record: Journal of the Transportation Research Board*, 1898, 138-147.

von Haefen, R.H., and Phaneuf, D.J., 2003. Estimating preferences for outdoor recreation: A comparison of continuous and count data demand system frameworks. *Journal of Environmental Economics and Management*, 45, 612-630.

Hendel, I., 1999. Estimating multiple-discrete choice models: An application to computerization returns. *Review of Economic Studies*, 66, 423-446.

Highfill, T., and Franks, C., 2019. Measuring the U.S. outdoor recreation economy, 2012–2016. *Journal of Outdoor Recreation and Tourism*, 27, 100233.

Jäggi, B., Weis, C., and Axhausen, K.W., 2013. Stated response and multiple discrete-continuous choice models: Analyses of residuals. *Journal of Choice Modelling*, 6, 44-59.

Jian, S.S., Rashidi, T.H., and Dixit, V., 2017. An analysis of carsharing vehicle choice and utilization patterns using multiple discrete-continuous extreme value (MDCEV) models. *Transportation Research Part A*, 103, 362-376.

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

Kuriyama, K., and Hanemann, W.M., 2006. The integer programming approach to a generalized corner-solution model: An application to recreation demand. *Working paper*, Waseda University, Tokyo.

Lee, S., and Allenby, G.M., 2014. Modeling indivisible demand. *Marketing Science*, 33(3), 364-381. https://doi.org/10.1287/mksc.2013.0829

Lu, H., Hess, S., Daly, A., and Rohr, C., 2017. Measuring the impact of alcohol multi-buy promotions on consumers' purchase behavior. *Journal of Choice Modelling*, 24, 75-95.

Ma, J., Ye, X., and Pinjari, A.R., 2019. Practical method to simulate multiple discrete-continuous generalized extreme value model: Application to examine substitution patterns of household transportation expenditures. *Transportation Research Record: Journal of the Transportation Research Board*, 2673(8), 145-156. https://doi.org/10.1177/0361198119842819

Mäler, K.G., 1974. Environmental economics: A theoretical inquiry. *The Johns Hopkins University Press for Resources for the Future*, Baltimore, MD.

Paillard-Borg, S., Wang, H., Winblad, B., and Fratiglioni, L., 2009. Pattern of participation in leisure activities among older people in relation to their health conditions and contextual factors: A survey in a Swedish urban area. *Ageing and Society*, 29(5), 803-821. https://doi.org/10.1017/S0144686X08008337

Parady, G., Katayama, G., and Yamazaki, H., 2019. Analysis of social networks, social interactions, and out-of-home leisure activity generation: Evidence from Japan. *Transportation*, 46(3), 537-562. https://doi.org/10.1007/s11116-018-9873-8

Pellegrini, A., Sarman, I., and Maggi, R., 2020. Understanding tourists' expenditure patterns: A stochastic frontier approach within the framework of multiple discrete-continuous choices. *Transportation*, forthcoming. https://doi.org/10.1007/s11116-020-10083-2

Pernice-Duca, F.M., 2010. An examination of family and social support networks as a function of ethnicity and gender: A descriptive study of youths from three ethnic reference groups. *Journal of Youth Studies*, 13(3), 391-402. https://doi.org/10.1080/13676260903447536

Pinjari, A.R., 2011. Generalized extreme value (GEV)-based error structures for multiple discrete-continuous choice models. *Transportation Research Part B*, 45(3), 474-489.

Pinjari, A.R., and Bhat, C.R., 2011. Computationally efficient forecasting procedures for Kuhn-Tucker consumer demand model systems: application to residential energy consumption analysis. Technical paper, Department of Civil and Environmental Engineering, University of South Florida.

Pinjari, A.R., Augustin, B., Imani, V.S., Eluru, N., and Pendyala, R.M., 2016. Stochastic frontier estimation of budgets for Kuhn–Tucker demand systems: Application to activity time-use analysis. *Transportation Research Part A*, 88, 117-133.

Pinjari, A.R., Pendyala, R.M., Bhat, C.R., and Waddell, P.A., 2011. Modeling the choice continuum: An integrated model of residential location, auto ownership, bicycle ownership, and commute tour mode choice decisions. *Transportation*, 38(6), 933-958.

Seo, S., Ohmori, N., and Harata, N., 2013. Effects of household structure and accessibility on travel. *Transportation*, 40, 847-865. https://doi.org/10.1007/s11116-013-9468-3.

Shin, J., Hwang, W.S., and Choi, H., 2019. Can hydrogen fuel vehicles be a sustainable alternative on vehicle market? Comparison of electric and hydrogen fuel cell vehicles. *Technological Forecasting and Social Change*, 143, 239-248.

Train, K., 2003. *Discrete Choice Methods with Simulation.* Cambridge University Press, New York.

Varghese, V., and Jana, A., 2019. Multitasking during travel in Mumbai, India: Effect of satiation in heterogeneous urban settings. *Journal of Urban Planning and Development*, 145(2), 04019002. https://doi.org/10.1061/(ASCE)UP.1943-5444.0000504

Viruell-Fuentes, E.A., Morenoff, J.D., Williams, D.R., and House, J.S., 2013. Contextualizing nativity status, Latino social ties, and ethnic enclaves: An examination of the 'immigrant social ties hypothesis'. *Ethnicity & Health*, 18(6), 586-609. https://doi.org/10.1080/13557858.2013.814763

Wales, T.J., and Woodland, A.D., 1983. Estimation of consumer demand systems with binding non-negativity constraints. *Journal of Econometrics*, 21(3), 263-85.

You, D., Garikapati, V.M., Pendyala, R.M., Bhat, C.R., Dubey, S., Jeon, K., and Livshits, V., 2014. Development of vehicle fleet composition model system for implementation in activity-based travel model. *Transportation Research Record: Journal of the Transportation Research Board*, 2430, 145-154.

Appendix A:
Intermediate derivation to show that the probability expression for grouped consumption collapses to a closed form expression

To show that our probability expression for grouped consumption collapses to a closed-form expression (essentially, the multivariate logistic CDF), we start off with the integrand in Equation (7) of the text and integrate it from $-\infty$ to the upper bounds (this is an *M*-dimensional integration). The integration steps are shown below.

Let $G_{K-1}(x_2^* < a_{2,c_2}, x_3^* < a_{3,c_3}, ..., x_{M+1}^* < a_{M+1,c_{M+1}}, x_{M+2}^* = 0, ..., x_{K-1}^* = 0, x_K^* = 0)$

$$= \int_{x_2^*=-\infty}^{a_{2,c2}} \int_{x_3^*=-\infty}^{a_{3,c3}} ... \int_{x_{M+1}^*=-\infty}^{a_{M+1,cM+1}} |J| \left(\frac{M!}{\sigma^M} \times \frac{\prod_{k=2}^{M+1} e^{\frac{-(\tilde{V}_k)}{\sigma}}}{\left(1 + \sum_{k=2}^{M+1} e^{\frac{-(\tilde{V}_k)}{\sigma}} + \sum_{k=M+2}^{K} e^{\frac{-(\tilde{V}_{k0})}{\sigma}}\right)^{M+1}} \right) dx_{M+1}^* ... dx_3^* dx_2^* \quad \text{(A.1)}$$

Now, $\tilde{V}_k = \tilde{V}_{k0} + \ln\left(\frac{x_k^*}{\gamma_k} + 1\right)$, and $|J| = \left[\prod_{i=2}^{M+1} f_i\right] = \left(\prod_{i=2}^{M+1} \frac{1}{x_i^* + \gamma_i}\right)$ therefore, the above integration expression can be re-written as,

$$\frac{M!}{\sigma^M} \int_{x_2^*=-\infty}^{a_{2,c2}} \int_{x_3^*=-\infty}^{a_{3,c3}} ... \int_{x_{M+1}^*=-\infty}^{a_{M+1,cM+1}} \left(\times \left(\prod_{i=2}^{M+1} \frac{1}{x_i^* + \gamma_i}\right) \times \frac{\prod_{k=2}^{M+1} e^{\frac{-(\tilde{V}_{k0})}{\sigma}} \left(\frac{x_k^*}{\gamma_k} + 1\right)^{(-1/\sigma)}}{\left(1 + \sum_{k=2}^{M+1} e^{\frac{-(\tilde{V}_{k0})}{\sigma}} \left(\frac{x_k^*}{\gamma_k} + 1\right)^{(-1/\sigma)} + \sum_{k=M+2}^{K} e^{\frac{-(\tilde{V}_{k0})}{\sigma}}\right)^{M+1}} \right) dx_{M+1}^* ... dx_3^* dx_2^* \quad \text{(A.2)}$$

We evaluate this by starting from the innermost integral and solving one at a time. Let the first integration variable be x_{M+1}^*, so we focus only on terms in the numerator that contains the variable x_{M+1}^* (this is easy to deal with since the numerator only contains terms in a multiplicative form).

Hence, the first integration (the innermost one) can be written as,

$$\mathrm{I}_{x_{M+1}} = \int_{x^*_{M+1}=-\infty}^{a_{M+1,c_{M+1}}} \left[\left(\left(\frac{1}{x^*_{M+1}+\gamma_{M+1}} \right) \times \frac{e^{\frac{-(\tilde{V}_{M+1,0})}{\sigma}} \left(\frac{x^*_{M+1}}{\gamma_{M+1}}+1 \right)^{(-1/\sigma)}}{\left(1+\sum_{k=2}^{M+1} e^{\frac{-(\tilde{V}_{k0})}{\sigma}} \left(\frac{x^*_k}{\gamma_k}+1 \right)^{(-1/\sigma)} + \sum_{k=M+2}^{K} e^{\frac{-(\tilde{V}_{k0})}{\sigma}} \right)^{M+1}} \right) dx^*_{M+1} \right] \quad \text{(A.3)}$$

To evaluate this integral, let $t = 1+\sum_{k=2}^{M+1} e^{\frac{-(\tilde{V}_{k0})}{\sigma}} \left(\frac{x^*_k}{\gamma_k}+1 \right)^{(-1/\sigma)} + \sum_{k=M+2}^{K} e^{\frac{-(\tilde{V}_{k0})}{\sigma}}$

Then $dt = e^{\frac{-(\tilde{V}_{M+1,0})}{\sigma}} \left(-\frac{1}{\sigma} \left(\frac{x^*_{M+1}}{\gamma_{M+1}}+1 \right)^{(-1/\sigma)-1} \times \frac{1}{\gamma_{M+1}} \right) dx^*_{M+1}$

or, $dt = e^{\frac{-(\tilde{V}_{M+1,0})}{\sigma}} \left(-\frac{1}{\sigma} \left(\frac{x^*_{M+1}}{\gamma_{M+1}}+1 \right)^{(-1/\sigma)} \times \left(\frac{1}{x^*_{M+1}+\gamma_{M+1}} \right) \right) dx^*_{M+1}$

Therefore, the integral $\mathrm{I}_{x_{M+1}}$ can be re-written as (ignoring the integration limits for the moment)

$$\mathrm{I}_{x_{M+1}} = \int -\sigma \frac{dt}{t^{M+1}}$$

This is a straightforward integration to evaluate. Now substituting the values in terms of x^*_{M+1} with the appropriate limits, we have, after evaluating the integral,

$$\mathrm{I}_{x_{M+1}} = \frac{\sigma}{M} \frac{1}{\left(1+\sum_{k=2}^{M+1} e^{\frac{-(\tilde{V}_{k0})}{\sigma}} \left(\frac{x^*_k}{\gamma_k}+1 \right)^{(-1/\sigma)} + \sum_{k=M+2}^{K} e^{\frac{-(\tilde{V}_{k0})}{\sigma}} \right)^{M}} \Bigg|_{x^*_{M+1}=-\infty}^{a_{M+1,c_{M+1}}} \quad \text{(A.4)}$$

which evaluates to,

$$\mathrm{I}_{x_{M+1}} = \frac{\sigma}{M} \frac{1}{\left(1+\sum_{k=2}^{M} e^{\frac{-(\tilde{V}_k)}{\sigma}} + e^{\frac{-(\tilde{V}_{M+1,0})}{\sigma} - \frac{1}{\sigma}\ln\left(\frac{a_{M+1,c_{M+1}}}{\gamma_{M+1}}+1 \right)} + \sum_{k=M+2}^{K} e^{\frac{-(\tilde{V}_{k0})}{\sigma}} \right)^{M}} \quad \text{(A.5)}$$

Now the integration expression in Equation (A.1) can be re-written as follows (this is now an *M*–1 dimensional integration)

$$\frac{M!}{\sigma^M}\int_{x_2^*=-\infty}^{a_{2,c2}}\int_{x_3^*=-\infty}^{a_{3,c3}}\cdots\int_{x_M^*=-\infty}^{a_{M,cM}}\frac{\sigma}{M}\left(\times\left(\prod_{i=2}^{M}\frac{1}{x_i^*+\gamma_i}\right)\times\frac{\prod_{k=2}^{M}e^{\frac{-(\tilde{V}_{k0})}{\sigma}}\left(\frac{x_k^*}{\gamma_k}+1\right)^{(-1/\sigma)}}{\left(1+\sum_{k=2}^{M}e^{\frac{-(\tilde{V}_k)}{\sigma}}+e^{\frac{-(\tilde{V}_{M+1,0})}{\sigma}-\frac{1}{\sigma}\ln\left(\frac{a_{M+1,cM+1}}{\gamma_{M+1}}+1\right)}+\sum_{k=M+2}^{K}e^{\frac{-(\tilde{V}_{k0})}{\sigma}}\right)^{M}}\right)dx_M^*\ldots dx_3^*dx_2^* \tag{A.6}$$

Now we repeat the same steps for the next innermost integral, i.e. with respect to the integration variable x_M^*. Like before, we write this as,

$$\mathrm{I}_{x_M}=\int_{x_M^*=-\infty}^{a_{M,cM}}\left[\left(\left(\frac{1}{x_M^*+\gamma_M}\right)\times\frac{e^{\frac{-(\tilde{V}_{M,0})}{\sigma}}\left(\frac{x_M^*}{\gamma_M}+1\right)^{(-1/\sigma)}}{\left(1+\sum_{k=2}^{M}e^{\frac{-(\tilde{V}_k)}{\sigma}}+e^{\frac{-(\tilde{V}_{M+1,0})}{\sigma}-\frac{1}{\sigma}\ln\left(\frac{a_{M+1,cM+1}}{\gamma_{M+1}}+1\right)}+\sum_{k=M+2}^{K}e^{\frac{-(\tilde{V}_{k0})}{\sigma}}\right)^{M}}\right)dx_M^*\right] \tag{A.7}$$

Proceeding in a similar manner like in the case of the earlier variable, it is easy to see that this integration will eventually take the following form (after evaluating the appropriate limits).

$$\mathrm{I}_{x_M}=\frac{\sigma}{(M-1)}\frac{1}{\left(1+\sum_{k=2}^{M-1}e^{\frac{-(\tilde{V}_k)}{\sigma}}+\sum_{j=M}^{M+1}e^{\frac{-(\tilde{V}_{j0})}{\sigma}-\frac{1}{\sigma}\ln\left(\frac{a_{j,cj}}{\gamma_j}+1\right)}+\sum_{k=M+2}^{K}e^{\frac{-(\tilde{V}_{k0})}{\sigma}}\right)^{M-1}} \tag{A.8}$$

Now, our main integration expression in Equation (A.1) will look like the following,

$$\frac{M!}{\sigma^M}\int_{x_2^*=-\infty}^{a_{2,c2}}\int_{x_3^*=-\infty}^{a_{3,c3}}\cdots\int_{x_{M-1}^*=-\infty}^{a_{M-1,cM-1}}\frac{\sigma^2}{M(M-1)}\left(\times\left(\prod_{i=2}^{M-1}\frac{1}{x_i^*+\gamma_i}\right)\times\frac{\prod_{k=2}^{M-1}e^{\frac{-(\tilde{V}_{k0})}{\sigma}}\left(\frac{x_k^*}{\gamma_k}+1\right)^{(-1/\sigma)}}{\left(1+\sum_{k=2}^{M-1}e^{\frac{-(\tilde{V}_k)}{\sigma}}+\sum_{j=M}^{M+1}e^{\frac{-(\tilde{V}_{j0})}{\sigma}-\frac{1}{\sigma}\ln\left(\frac{a_{j,cj}}{\gamma_j}+1\right)}+\sum_{k=M+2}^{K}e^{\frac{-(\tilde{V}_{k0})}{\sigma}}\right)^{M-1}}\right)dx_{M-1}^*\ldots dx_3^*dx_2^* \tag{A.9}$$

The entire integration is completed when the above process is repeated until we evaluate the outermost integral with respect to the x_2^* variable. In fact, just before the outermost (last) integral is evaluated, the integration expression (in Equation A.1) will take the following form,

$$\frac{M!}{\sigma^M}\int_{x_2^*=-\infty}^{a_{2,c2}} \frac{\sigma^{M-1}}{M(M-1)(M-2)...2}\left(\times\left(\frac{1}{x_2^*+\gamma_2}\right)\times\frac{e^{\frac{-(\tilde{V}_{2,0})}{\sigma}}\left(\frac{x_2^*}{\gamma_2}+1\right)^{(-1/\sigma)}}{\left(1+e^{\frac{-(\tilde{V}_2)}{\sigma}}+\sum_{j=3}^{M+1} e^{\frac{-(\tilde{V}_{j0})}{\sigma}-\frac{1}{\sigma}\ln\left(\frac{a_{j,cj}}{\gamma_j}+1\right)}+\sum_{k=M+2}^{K} e^{\frac{-(\tilde{V}_{k0})}{\sigma}}\right)^{M-1}}\right)dx_2^*$$

(A.10)

By evaluating this integral with respect to the x_2^* variable in a similar fashion as earlier variables, we obtain a closed-form expression for the integral in Equation (A.1) as below,

$$G_{K-1}(x_2^*<a_{2,c_2}, x_3^*<a_{3,c_3},...,x_{M+1}^*<a_{M+1,c_{M+1}}, x_{M+2}^*=0,...,x_{K-1}^*=0, x_K^*=0)$$

$$=\int_{x_2^*=-\infty}^{a_{2,c2}}\int_{x_3^*=-\infty}^{a_{3,c3}}...\int_{x_{M+1}^*=-\infty}^{a_{M+1,cM+1}}|J|\left(\frac{M!}{\sigma^M}\times\frac{\prod_{k=2}^{M+1} e^{\frac{-(\tilde{V}_k)}{\sigma}}}{\left(1+\sum_{k=2}^{M+1} e^{\frac{-(\tilde{V}_k)}{\sigma}}+\sum_{k=M+2}^{K} e^{\frac{-(\tilde{V}_{k0})}{\sigma}}\right)^{M+1}}\right)dx_{M+1}^*...dx_3^*dx_2^*$$

$$=\frac{1}{\left(1+\sum_{j=2}^{M+1} e^{\frac{-(\tilde{V}_{j0})}{\sigma}-\frac{1}{\sigma}\ln\left(\frac{a_{j,cj}}{\gamma_j}+1\right)}+\sum_{k=M+2}^{K} e^{\frac{-(\tilde{V}_{k0})}{\sigma}}\right)} \qquad \text{(A.11)}$$

This closed-form expression shows that the probability expression for multiple discrete-grouped consumption, as shown in Equation (10) of the text, is also a closed-form expression.

Appendix B:
Extension of the proposed MDGEV formulation to a flexible form wherein the baseline utility is explicitly separated along the discrete and continuous consumption dimensions

Following Bhat's (2018) formulation, let the utility function be,

$$U(\boldsymbol{x}) = \psi_1 x_1 + \sum_{k=2}^{K} \gamma_k \left(\left[\psi_{kd}\right]^{1(x_k=0)} \times \left[\psi_{kc}\right]^{1(x_k>0)} \right) \ln\left\{ \left(\frac{x_k}{\gamma_k} + 1 \right) \right\}, \tag{B.1}$$

where the original ψ_k is partitioned into two multiplicative components, ψ_{kd} and ψ_{kc}. The first component ψ_{kd} corresponds to the baseline preference that determines whether or not good *k* will be consumed (the *D*-preference component) and the second component ψ_{kc} corresponds to the baseline preference if good *k* is consumed (the *C*-preference component).

The KKT conditions then take the following form:

$$\psi_{kd} - \lambda p_k > 0 \text{ and } (\psi_{kc}) \left(\frac{x_k^*}{\gamma_k} + 1 \right)^{-1} = \lambda p_k \text{ for } k = 2,\ldots, K \text{ with consumption } x_k^* (x_k^* > 0)$$

$$\psi_{kd} - \lambda p_k < 0 \text{ if } x_k^* = 0,\ k = 2,..., K \tag{B.2}$$

$$\psi_1 = \lambda .$$

Substituting for λ from the last equation into the earlier equations for the inside goods, and taking logarithms, we can rewrite the KKT conditions as:

$$\ln(\psi_{kd}) - \ln(\psi_1) - \ln p_k > 0 \text{ and } \ln(\psi_{kc}) - \ln\left(\frac{x_k^*}{\gamma_k} + 1 \right) - \ln(\psi_1) - \ln p_k = 0$$

$$\text{for } k = 2,\ldots, K \text{ with consumption } x_k^* (x_k^* > 0) \tag{B.3}$$

$$\ln(\psi_{kd}) - \ln(\psi_1) - \ln p_k < 0 \text{ if } x_k^* = 0,\ k = 2,..., K .$$

To ensure the positivity of the *D*-preference and the *C*-preference terms, we specify these two components for each inside good as follows:

$$\psi_{kd} = \exp(\boldsymbol{\beta}' \boldsymbol{z}_k + \varepsilon_k) \text{ and } \psi_{kc} = \exp(\boldsymbol{\theta}' \mathrm{q}_k + \xi_k) , \tag{B.4}$$

where $\boldsymbol{z}_k$ and ε_k are as defined earlier in the text, but now are specific to the *D*-preference component of good *k*, and $\boldsymbol{q}_k$ and ξ_k are similarly defined for the *C*-preference component. The vectors $\boldsymbol{z}_k$ and $\boldsymbol{q}_k$ can include some common attributes, but can also have different attributes. Using notations already defined in the text, the KKT conditions can be reframed as follows:

$\eta_k > \tilde{V}_{k,1}$ and $\varsigma_k = \vec{V}_{k,1}$ if $x_k^* > 0$ ($k = 2, 3,\ldots, K$), $\eta_k = \varepsilon_k - \varepsilon_1$ and $\varsigma_k = \xi_k - \varepsilon_1$,

$\eta_k < \tilde{V}_{k,1}$ if $x_k^* = 0$ ($k = 2, 3,\ldots, K$), where

$$\tilde{V}_{k,1} = \boldsymbol{\beta}'\boldsymbol{z}_1 - \left(\boldsymbol{\beta}'\boldsymbol{z}_k - \ln p_k\right), \text{ and} \tag{B.5}$$

$$\vec{V}_{k,1} = \boldsymbol{\beta}'\boldsymbol{z}_1 - \left(\boldsymbol{\theta}'\boldsymbol{q}_k - \ln p_k\right) + \ln\left(\frac{x_k^*}{\gamma_k} + 1\right).$$

The error terms η_k (k = 2, 3,…, K) and the error terms ς_k (k = 2, 3,…, K) are jointly multivariate logistically distributed (with a fixed correlation of 0.5 across all pairings of these error terms), if we assume that the error terms ε_k (k = 1, 2,…, K) and the error terms ξ_k (k = 2, 3,…, K) are all identically and independently Gumbel distributed with a scale parameter σ. Then, following all the notations already defined in the text and in this appendix, the probability of the consumption pattern for the case of $M \geq 1$ and $M < K - 1$ may be written as follows:

$$P\left(c_2, c_3, \ldots, c_{M+1}, 0, \ldots, 0, 0\right) = \sum_{S'=1}^{S'=2^M} (-1)^{L_{S'}} \boldsymbol{F}_{K-1}(\vec{\boldsymbol{W}}_{S'}, \tilde{V}_{M+2,1}, \ldots \tilde{V}_{K-1,1}, \tilde{V}_{K,1}) +$$

$$\sum_{S \subset \{2,3,\ldots M+1\}, |S| \geq 1} (-1)^{|S|} \sum_{S'=1}^{S'=2^M} (-1)^{L_{S'}} \boldsymbol{F}_{K-1+|S|}(\vec{\boldsymbol{W}}_{S'}, \tilde{V}_{M+2,1}, \ldots \tilde{V}_{K-1,1}, \tilde{V}_{K,1}, \tilde{\boldsymbol{V}}_{S,1}) \tag{B.6}$$

Where $\vec{W}_{k,c_k} = \boldsymbol{\beta}'\boldsymbol{z}_1 - \left(\boldsymbol{\theta}'\boldsymbol{q}_k - \ln\left(\frac{a_{k,c_k}}{\gamma_k} + 1\right) - \ln p_k\right)$ and S' represents a specific combination of length M of the $\vec{W}_{k,c_k-1}$ and $\vec{W}_{k,c_k}$ scalars across all the consumed inside goods (k=2,3,…,M+1) such that both $\vec{W}_{k,c_k-1}$ and $\vec{W}_{k,c_k}$ are disallowed in the combination for any k (there are 2^M such combinations, and we will represent the resulting vector of elements in combination S' as $\vec{\boldsymbol{W}}_{S'}$), and $L_{S'}$ is a count of the number of lower thresholds $\vec{W}_{k,c_k-1}$ (k=2,3,…,M+1) appearing in the vector $\vec{\boldsymbol{W}}_{S'}$.

In the specific case that all the inside goods are consumed (that is, $M = K - 1$), the corresponding consumption probability is as follows:

$$P\left(\ c_2,\ c_3,\ ...,\ c_{M+1}, c_{M+2},\ ..., c_{K-1}, c_K\ \right) = \sum_{S'=1}^{S'=2^{K-1}} (-1)^{L_{S'}} \boldsymbol{F}_{K-1}(\vec{\boldsymbol{W}}_{S'}) +$$

$$\sum_{S\subset\{2,3,...K\},|S|\geq 1} (-1)^{|S|} \sum_{S'=1}^{S'=2^{K-1}} (-1)^{L_{S'}} \boldsymbol{F}_{K-1+|S|}(\vec{\boldsymbol{W}}_{S'}, \tilde{\boldsymbol{V}}_{S,1}) \qquad \text{(B.7)}$$

In the case when none of the inside goods are consumed (that is, $M = 0$), the corresponding consumption probability is:

$$P\left(\ 0,0,...,0,0,...,0,0\right) = \boldsymbol{F}_{K-1}(\tilde{V}_{2,1}, \tilde{V}_{3,1}, ..., \tilde{V}_{M+1,1}, \tilde{V}_{M+2,1}, ...\tilde{V}_{K-1,1}, \tilde{V}_{K,1}) \qquad \text{(B.8)}$$

LIST OF TABLES

Table 1: Data description for the vehicle-use case study (sample size = 1778)

Vehicle-type distribution					
Vehicle-type	**Household (HH) vehicle ownership levels**				**Average annual mileage**
	1-vehicle HH	2-vehicles HH	3-vehicles HH	4 or more vehicles HH	
Passenger Car	490 (55.9%)	528 (34.7%)	106 (27.8%)	14 (24.1%)	8620
Van	37 (4.2%)	97 (6.4%)	22 (5.8%)	7 (12.1%)	7520
SUV	254 (29.0%)	463 (30.5%)	96 (25.2%)	12 (20.7%)	9895
Pickup truck	92 (10.5%)	404 (26.6%)	106 (27.8%)	13 (22.4%)	8805
Other	4 (0.4%)	28 (1.8%)	51 (13.4%)	12 (20.7%)	3740
Total	877 (100%)	1520 (100%)	381 (100%)	58 (100%)	---

Table 2: MDGEV result for the time-use case at 15-minutes clustering (M1)

Variables	Coefficient estimates (t-stats)			
	In-Home Social (IHS)	**Out-of-Home Social (OHS)**	**In-Home Recreational (IHR)**	**Out-of-Home Recreational (OHR)**
Household sociodemographic				
Number of children aged 0-4 years	-	0.309 (2.98)	-	-
Number of children aged 5-15 years	-	-0.161 (-2.03)	-	-
Number of adults	-	-	0.312 (4.21)	-
Number of household vehicles	-	-	-0.191 (-3.20)	-
Number of bicycles in the household	-	-	-	0.092 (3.54)
Household income (Base: >$60,000/yr)				
Household income less than $35,000/yr	-	-	0.676 (4.74)	-
Household income $35,000/yr-$60,000/yr	-	-	0.263 (2.37)	-
Household location attribute				
Land-use mix	-	-	0.685 (2.59)	-0.576 (-2.08)
Individual characteristics				
Female	-	-	-0.330 (-3.65)	-
Age of individual (Base: Less than 50 years)				
Age 50-65	-	-	-	-0.247 (-2.12)
Age greater than 65			0.663 (3.62)	-0.373 (-2.02)
Employed	-0.512 (-2.38)	-	-	-
Hispanic	-	0.609 (2.79)	-	-
Day and seasonal effects				
Weekend day is Sunday (Base: Saturday)	-	-	0.369 (4.09)	-
Winter (Base: Summer and Spring)	-	0.348 (1.97)	-	-0.388 (-2.05)
Fall (Base: Summer and Spring)	-	-	-	-0.278 (-2.51)
Baseline preference constants	-2.422 (-14.81)	-1.223 (-14.23)	-1.304 (-6.39)	-0.560 (-3.81)
Satiation effects				
Household size	-	-0.156 (-1.78)	-	-
Weekend day is Sunday (Base: Saturday)	-	-0.339 (-1.66)	-	-
Satiation constant	4.963 (20.13)	5.138 (18.53)	5.056 (43.48)	4.690 (39.91)

Table 3: MDGEV result for the vehicle-use case

Variables	Coefficient estimates (t-stats)				
	Passenger car	**Van**	**SUV**	**Pickup truck**	**Other**
Household sociodemographic					
Household Income (Base: Greater than $125,000)					
Income less than $35,000 annually	-0.287 (-2.73)	-	-1.116 (-9.17)	-0.379 (-3.03)	-
Income between $35,000 - $75,000 annually	-0.287 (-2.73)	-	-0.563 (-5.82)	-	-
Income between $75,000 - $125,000 annually	-0.287 (-2.73)	-	-	-	-
Number of children in the household	-	0.613 (8.72)	-	-	-
Number of adults in the household	-0.167 (-2.40)	0.820 (5.19)	-	-	-
Number of workers in the household	0.247 (4.44)	-0.399 (-3.20)	-	-	-
Race is White (Base: Non-white)	-	-	-	0.655 (4.19)	1.264 (2.42)
Household location attributes					
Population density more than 4000 persons/sq. mile (Base: less than 4000 persons/sq. mile)	0.366 (4.40)	-	-	-	-
Employment density more than 500 workers/sq. mile (Base: less than 500 workers/sq. mile)	-	-	-	-0.735 (-7.08)	-
Baseline preference constants	0.061 (0.395)	-3.809 (-11.84)	0.034 (0.35)	-0.830 (-4.51)	-4.151 (-8.15)
Satiation effects					
Income less than $35,000 annually (Base: more than $35,000)	-0.519 (-3.04)	-	-	-0.435 (-1.77)	-
Number of Workers		-	-	0.378 (2.79)	-
Household is in an urban area (Base: Non-urban)	-0.342 (-1.61)	-	-	-	-
Satiation constant	8.640 (35.27)	8.686 (44.77)	8.863 (79.19)	8.241 (43.09)	7.245 (32.16)

Table 4: Likelihood based data fit measures for the time-use case study

	Estimation sample (N=1500)					Hold-out sample (N=417)				
	MDCEV (M0)	**MDGEV (M1)**	**MDGEV (M2)**	**MDGEV (M3)**	**MDCEV (M4)**	**MDCEV (M0)**	**MDGEV (M1)**	**MDGEV (M2)**	**MDGEV (M3)**	**MDCEV (M4)**
For the multiple discrete-continuous consumption (MDC) component										
Predictive log-likelihood at convergence	-13560.54	-13560.85	-13560.67	-13576.70	-13613.80	-3903.87	-3904.17	-3904.13	-3905.11	-3903.95
Log-likelihood at constants	-13641.03	-13641.03	-13641.03	-13641.03	-13641.03	-3913.13	-3913.13	-3913.13	-3913.13	-3913.13
Number of model parameters	29	29	29	29	29	29	29	29	29	29
Number of non-constants parameters	21	21	21	21	21	21	21	21	21	21
Bayesian Information Criterion	13666.58	13666.89	13666.71	13682.74	13719.84	3991.35	3991.65	3991.61	3992.59	3991.43
Nested likelihood ratio test ($-2 * [\mathcal{L}(\hat{\theta}) - \mathcal{L}(C)]$)	160.98*	160.35*	160.73*	128.67*	54.45*					
	*All values are greater than Chi-squared statistics with 21 degrees of freedom at any reasonable level of significance, indicating superior fit relative to the constants-only model.									
For the purely discrete (MD) component										
Predictive log-likelihood at convergence	-3366.87	-3366.85	-3366.82	-3369.24	-3366.95	-971.95	-972.10	-972.18	-972.81	-975.82
Log-likelihood at constants	-3444.17	-3444.17	-3444.17	-3444.17	-3444.17	-975.99	-975.99	-975.99	-975.99	-975.99
Number of model parameters	23	23	23	23	23	23	23	23	23	23
Number of non-constants parameters	19	19	19	19	19	19	19	19	19	19
Bayesian Information Criterion	3450.97	3450.95	3450.92	3453.34	3451.05	1041.33	1041.48	1041.56	1042.19	1045.20
Adjusted likelihood ratio index	0.0169	0.0169	0.0169	0.0162	0.0169					
Nested likelihood ratio test ($-2 * [\mathcal{L}(\hat{\theta}) - \mathcal{L}(C)]$)	154.6#	154.64#	154.70#	149.86#	154.55#					
	#All values are greater than Chi-squared statistics with 19 degrees of freedom at any reasonable level of significance, indicating superior fit relative to the constants-only model.									

Table 5: Aggregate non-likelihood fit measures for the time-use case for 15-minute clustering MDGEV model

	ESTIMATION SAMPLE (N = 1500)						HOLD-OUT SAMPLE (N = 417)					
Aggregate heuristic check for multiple discrete-grouped consumption based on trinary prediction												
Activity participation	**Number of individuals participating in the respective activity for the following grouped interval**											
	0 minute		0+ to 120 minutes		≥ 120 minutes		0 minute		0+ to 120 minutes		≥ 120 minutes	
	Obs.	Pred.	Obs.	Pred.	Obs.	Pred.	Obs.	Pred.	Obs.	Pred.	Obs.	Pred.
IHS	1406	1408	43	40	51	52	393	392	14	11	10	14
OHS	1114	1142	176	170	210	188	307	318	44	47	66	52
IHR	937	993	133	165	430	342	242	272	39	47	136	99
OHR	1002	1047	196	194	302	259	283	291	57	54	77	72
Weighted mean absolute percentage error for each group (%)		2.94		7.62		15.54		4.05		10.82		20.95
Overall weighted mean absolute percentage error (%)						5.45						7.60

Table 6: Likelihood based data fit measures for the vehicle-use case study

	Estimation sample (N=1375)	Hold-out sample (N=403)
For grouped consumption (MDG) component		
Log-likelihood at convergence	-10597.95	-3157.64
Log-likelihood at constants	-10776.91	-3184.17
Number of model parameters	27	27
Number of non-constants parameters	17	17
Bayesian Information Criterion (BIC)	10695.50	3238.63
Nested likelihood ratio test ($-2 * [\mathcal{L}(\hat{\theta}) - \mathcal{L}(C)]$)	357.92 (greater than Chi-squared statistics at 17 degrees of freedom for any reasonable level of significance, indicating superior fit relative to the constants-only model.)	
For purely discrete (MD) component		
Predictive log-likelihood at convergence	-3709.61	-1133.71
Log-likelihood at constants	-3866.73	-1159.03
Number of model parameters	18	18
Number of non-constants parameters	13	13
Bayesian Information Criterion (BIC)	3774.65	1187.70
Adjusted likelihood ratio index	0.0373	
Nested likelihood ratio test ($-2 * [\mathcal{L}(\hat{\theta}) - \mathcal{L}(C)]$)	314.24 (greater than Chi-squared statistics at 13 degrees of freedom for any reasonable level of significance, indicating superior fit relative to the constants-only model.)	

Table 7: Aggregate fit measures for the vehicle-use MDGEV model

	ESTIMATION SAMPLE (N = 1375)						**HOLD-OUT SAMPLE (N = 403)**					
Aggregate heuristic check for multiple discrete-grouped (MDG) consumption based on trinary prediction												
Vehicle type	**Number of households in the respective mileage group for the vehicle-types usage**											
	0 mile		**0+ to 7500 miles**		**> 7500 miles**		**0 mile**		**0+ to 7500 miles**		**> 7500 miles**	
	Obs.	Pred.	Obs.	Pred.	Obs.	Pred.	Obs.	Pred.	Obs.	Pred.	Obs.	Pred.
Passenger Car	593	618	377	358	405	399	138	181	102	115	163	107
Vans	1245	1259	56	60	74	56	370	371	14	17	19	16
SUV	744	838	202	226	429	310	211	243	49	67	143	93
Pickup-truck	917	972	207	206	251	198	251	282	61	61	91	60
Other	1311	1312	57	53	7	10	382	384	19	16	2	3
Weighted mean absolute percentage error for group (%)		3.92		5.81		17.1		7.98		15.2		33.96
Overall weighted mean absolute percentage error (%)						6.40						14.25