Application of an Econometric Multiple Discrete Continuous Fusion Approach to Link Residential Sector Energy Demand and Travel Infrastructure and Usage

Md Istiak Jahan*
Postdoctoral Research Scholar
Department of Civil, Environmental & Construction Engineering
University of Central Florida, Orlando, Florida 32816, USA
Email: md.istiak.jahan@ucf.edu
ORCiD number: 0000-0002-4056-7816

Naveen Eluru
Professor
Department of Civil, Environmental & Construction Engineering
University of Central Florida, Orlando, Florida 32816, USA
Email: naveen.eluru@ucf.edu
ORCiD number: 0000-0003-1221-4113

* Corresponding author

Abstract
The considerable body of earlier research on household level residential sector energy demand does not consider the impact of household residents' travel infrastructure on energy consumption patterns. The absence of travel infrastructure elements in energy demand models can be attributed to the lack of data providing this information in energy surveys. In this study, a novel econometric fusion approach is utilized to combine the traditional energy dataset -Residential Energy Consumption Survey (RECS) data – with a transportation survey data - National Household Travel Survey (NHTS) data. The probabilistic fusion approach is employed to study energy consumption by end use type with a Multiple Discrete Continuous Extreme Value model framework. The framework will quantify the impact of travel infrastructure and usage related attributes on household end-use energy demand and remedy the over-estimation of the impact of socioeconomic attributes. The model results reveal the impact of several household socioeconomic attributes (i.e., household size, location and income) and travel infrastructure and usage related attributes (i.e. number of vehicles of different fuel and body types, household annual mileage and frequency of long-distance trips) on household end-use energy demand. The model estimation results are augmented with an elasticity analysis and policy analysis to highlight the implementation of the proposed framework. The elasticity results reveal that ignoring the influence of travel infrastructure and usage variables can contribute to errors for elasticity values for other independent variables such as household size (up to 1800%), number of adults (up to 50%) and residential location (up to 15%).

Keywords: Household energy, travel infrastructure and usage, data fusion, MDCEV model**,** energy end-uses

1 Background

Residential energy demand contributes significantly to the overall energy demand of a country. According to the U.S. Energy Information Administration (EIA), the residential sector accounted for around 20% of the total U.S. energy consumption in 2023 (EIA, 2024b). Currently, residential energy demand patterns are in a state of flux. The rapid adoption of advanced home technologies (such as energy efficient air conditioners, high-definition TVs, gaming consoles, and smart home devices), prevalence of emerging transportation alternatives (such as electric vehicle charging facilities), increasing share of residential solar charging, and work culture evolution post-COVID with higher rate of hybrid work will continue to impact household energy consumption patterns into the future. In addition to affecting national energy demand, the residential sector has a large impact on global warming. In 2021, residential energy consumption released about 365 million metric tons of carbon dioxide (USAFacts, 2024). As countries around the world tackle greenhouse gas emissions in their fight against global warming, identifying the factors affecting residential energy demand in the future will be an important contribution.

Residential energy demand patterns are well researched in energy literature. The demand patterns are typically analyzed along three key dimensions: (a) overall energy demand (see (Besagni & Borgarello, 2018; Wang et al., 2021)), (b) energy consumption by energy source, such as electricity, natural gas, and Liquefied Petroleum (LP) gas (see (Iraganaboina & Eluru, 2021; Lu et al., 2022)), and (c) energy demand for specific end-use type such as space heating, water heating, and cooling (see (Kuang et al., 2023a; Malla, 2022)). From a demand estimation perspective, assessing energy consumption across different end-use categories provides a thorough insight into the factors affecting household energy consumption patterns. The methodological approaches considered for residential energy modeling include regression-based approaches (Lee & Song, 2022; Xie & Noor, 2022), machine learning approaches (such as random-forest regression, k-nearest neighbors, gradient boosting method, extreme gradient boosting method, support vector machine, artificial neural network and deep extreme machine learning approach) (Burnett & Kiesling, 2022; Kuang et al., 2023b), bottom-up simulation approaches (Malla, 2022; Shimoda et al., 2021), engineering methodologies (such as Intergovernmental Panel on Climate Change (IPCC) method for carbon emission and Australian Zero Emissions House design tool (Fan et al., 2015; Ren et al., 2013) and advanced econometric frameworks (such as Multiple Discrete Continuous Extreme Value (MDCEV) model and mixed multinomial logit-MDCEV model) (Yu et al., 2013; Yu & Zhang, 2015). Across most of these approaches, energy demand by end-use is modeled separately by end-use. Thus, these approaches ignore that the different end-use energy demands arise from the same household. The MDCEV model provides an elegant framework in studying choice scenarios where multiple alternatives are selected. Examples of such choice scenarios include purchasing different types of vehicles in a household (Bhat & Sen, 2006), activity participation in a day (Bhat, 2005), and energy consumption by energy source (Iraganaboina & Eluru, 2021). The MDCEV approach, employed in this study, recognizes overall energy allocation across end-uses (such as space cooling, lighting, cooking and water heating) as a single behavioral process allowing for substitution and complementarity effects across end-use alternatives.

These studies reveal several important determinants of energy demand including – (a) household sociodemographic attributes, (b) dwelling attributes, (c) household appliance-use related attributes and (d) climate related attributes. Sociodemographic attributes such as household size, income, and location (urban vs. rural) significantly influence energy consumption patterns across different end-uses (Malla, 2022). Larger households often consume more energy due to increased usage of appliances for heating, cooling, and other purposes (Ellsworth-Krebs, 2020; Zou & Luo, 2019). Higher-income households also consume more energy than their counterparts, especially for discretionary uses such as advanced heating/cooling systems, larger living spaces, and additional appliances. They are also more likely to own EVs, influencing residential charging demand (Dixit & Singh, 2022; Shin et al., 2019). Furthermore, variations in energy consumption for space heating and cooling are also noticeable between urban and rural areas due to differences in dwelling size, insulation systems and access to modern technologies[1] (Chun-sheng et al., 2012; Heinonen & Junnila, 2014).

Interestingly, the considerable body of research on household level residential sector energy demand does not consider the impact of household residents' travel infrastructure and usage on energy consumption patterns. The out-of-home activities of all or a subset of household residents impacts the energy consumption pattern of various end-uses (such as space heating, water heating, space cooling, lighting, and cooking) and potential post-travel activities (such as cooling, washing and cleaning). The variables that can offer insights on travel infrastructure and usage can include household vehicle ownership, employment status, long-distance travel behavior and household annual mileage. Households with high vehicle ownership levels potentially represent increased private vehicle usage behavior such as spending significant time outside and this can potentially contribute to lower energy consumption at home. Further, the number of long-distance trips and annual mileage indicate how often household members stay away from home. These metrics directly affect the frequency of appliance usage, space heating/cooling and washing and drying. Therefore, the increased number of long-distance trips or longer annual mileage might influence the energy consumption for different household appliances. Similarly, vehicle fleet fuel type mix influences the energy consumption of various household end uses. For instance, households with electric vehicles (EVs) may potentially install solar infrastructure and energy management systems to manage the increased electricity demand affecting other end uses, such as space heating and cooling. The importance of considering the link between travel infrastructure and usage and energy consumption is likely to increase further with the growing adoption of EVs. With increasing adoption of EVs and the corresponding impact of EV charging on residential energy demand is a dimension of importance.

The absence of travel infrastructure and usage elements in energy demand models can be attributed to the lack of data providing this information in energy surveys. Traditional survey data or smart meter data is rarely augmented with detailed data related to household travel infrastructure and usage data. Transportation survey data on the other hand compiles this information. However, given these data are compiled on entirely different respondents there has not been any consideration for using the datasets together. In this study, a novel framework is employed that allows us to combine two datasets without any common identifier (see (Bhowmik, et al., 2024a; Jahan et al., 2024)). The 2020 Residential Energy Consumption Survey (RECS) data and the 2022 National Household Travel Survey (NHTS) data are employed for this fusion exercise. The dependent variable of interest is energy consumption by end-use dimension (such as space heating, water heating, refrigeration, cooling and ventilation, EV charging, and lighting). Given the nature of the dependent variable, the MDCEV model is employed in this analysis.

The current study contributes to the existing literature along two dimensions. *First*, this study contributes to end-use energy demand modeling systems by employing a novel fusion approach merging two datasets without a common identifier. *Second*, the study conducts a first of its kind assessment of the impact of travel infrastructure related variables on various household energy end-uses. The study recognizes that traditional approach to energy models that ignore travel infrastructure variables might offer erroneous energy estimates for future scenarios. The proposed model system accurately quantifies the impact of all independent variables by accommodating for travel infrastructure variables. The independent variables considered from energy data include household socioeconomic attributes, dwelling attributes and appliance-use related attributes. The new variables considered from NHTS data include household socioeconomic attributes and travel infrastructure and usage related attributes. The proposed approach allows us to easily examine if the data fusion contributes to improving our understanding of energy consumption by end-use in a straightforward manner using traditional model fit metrics (such as log-

[1]There is a growing body of literature examining how to arrive at Net Zero Energy Buildings. These studies focus on the development of energy share framework (Minelli et al., 2024), evaluation of different technology installations (D'Agostino et al., 2024), contributions of renewable energy generation (hydropower, wind energy, solar, heat pumps, and bioenergy) (Ahmed et al., 2022) and evaluation of design parameters affecting the energy performance of net zero energy buildings (Kaitouni et al., 2024).

likelihood and Information Criterion). The model results clearly identify which of the newly incorporated variables from NHTS provide improved data fit while also correctly quantifying the effect of independent variables from RECS dataset. The model estimation results are augmented with an elasticity analysis to highlight the inconsistencies of energy models that ignore travel infrastructure and usage variables while also showing the potential impact of travel pattern variables. It is important to note here that earlier work using the fusion approach considered continuous variables (Bhowmik, et al., 2024b) and binary outcome variables (Bhowmik, et al., 2024a). In the current paper, this approach is applied to accommodate multiple discrete continuous variables thus generalizing the fusion approach for different kinds of dependent variables.

The current study aims to develop an energy model framework that can estimate household level energy demand with information on household socioeconomic attributes and travel infrastructure and usage related attributes. The proposed model framework can be employed to estimate regional energy demand based on detailed census population data. The estimates can offer insights on energy demand in response to demographic and climate changes over time. Understanding these trends can offer policy makers with insights on what important treatments (such as improving insulation) can offer energy use reductions in different regions. The viability of the proposed model for scenario analysis has been illustrated in the policy analysis section of the paper.

The insights can also be used to design integrated energy management strategies that account for both residential and transportation energy needs. For instance, the number of EV sales is increasing in recent years and according to the US Department of Energy, 80% of EVs are charged at home (John, 2022). Therefore, in the future years, an increasing pressure on the national electricity grid is inevitable. This study provides with valuable insights into the factors that impact household EV charging infrastructure. By identifying the locations of higher concentration of households that consume energy for EV charging, the proposed framework can help in redesigning the national grid and assist with determining optimal locations for new EV charging infrastructure. Further, as household socioeconomic conditions vary across different regions and cities, the proposed framework can also serve as a valuable tool to estimate total energy demand across diverse geographic areas.

The remainder of the paper is organized as follows: The next section presents the experimental design of the fusion approach followed by a discussion of the methodological details. The data description section summarizes the datasets. The following section specifies the selection of the best fitted model while the estimation results of the best fitted model is summarized in the model estimation section. The applicability of the proposed framework is illustrated in the elasticity analysis section. The final section concludes the paper.

2 Data Fusion Algorithm

In this paper, the energy use data is employed from 2020 Residential Energy Consumption Survey (RECS) dataset and the travel infrastructure and usage data is employed from the 2022 National Household Travel Survey (NHTS) dataset. The two datasets do not share a unique identifier to easily merge them. Hence, a probabilistic fusion approach is employed for our analysis of energy consumption by end use.

The data fusion is performed based on the common attributes across the two datasets. The fusion algorithm hypothesizes that households with matching attributes across the two datasets are likely to share similar attributes (such as travel infrastructure and usage attributes form NHTS). Seven common variables – household division[2], household income, household location, household ownership, household region, household size and housing type – are present in the RECS and NHTS datasets. While it might be ideal to consider fusion with only a completely matched set of records, it is not always possible to get enough records for fusion with strict matching policy. As the number of matching variables increases, the number of possible matches across the two datasets reduces very rapidly. For example, when all seven variables are

[2] Household region partitions America into 4 regions West, South, Northeast and Midwest. Household Division is a more disaggregate classification of the region including East North Central, East South Central, Middle Atlantic, Mountain, New England, Pacific, South Atlantic, West North Central, and West South-Central region.

matched, the average number of matches is only 14. In addition, 13% RECS records do not have any matches from the NHTS dataset. Hence, in this context, among the seven common variables, matching at most six variables is considered to ensure adequate matching records exist. To be sure, the performance of each possible set of matching variables are carefully compared.

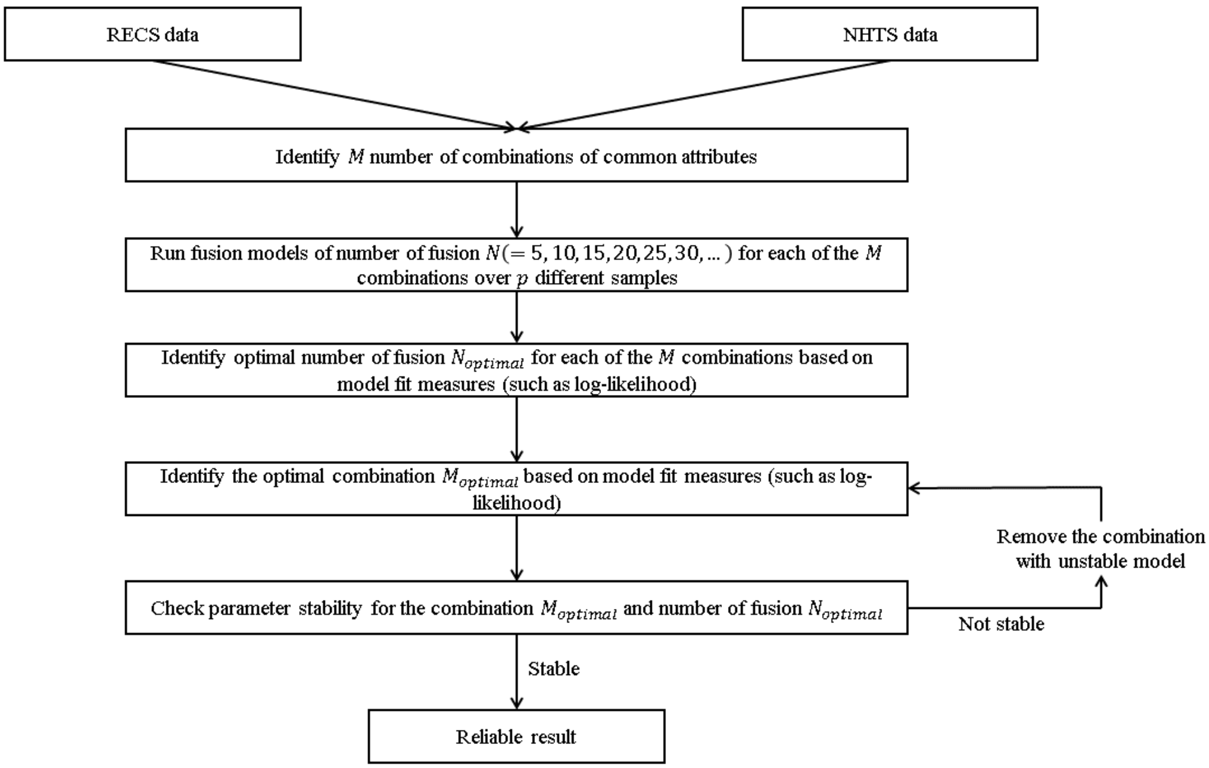

Figure 1: Flow chart of fusion algorithm

The reader would note that for any set of matching variables, multiple matches are likely to exist. Among these matches, there is no way to identify an "ideal match". Hence, the fusion process with multiple candidates from the pool of matched records is considered. The model fit of the dependent variable for different number of fusion records is systematically tested to identify the "optimal" number of fusion records. While increasing the size of fusion records can possibly improve the model, the increase in computation time needs to be recognized. When multiple NHTS records are fused for each RECS record, the energy consumption record will repeat with each fused record. Hence, a weight variable is considered that ensures each RECS record accounts for only one record. For example, if the number of matched records is 10, it is ensured that across the 10 records the newly added weight adds to 1. In this research, two approaches are employed for determining the weight. The first approach takes the form of a deterministic weight – $1 / (\text{Number of NHTS records})$ (so 1/10 in our example). In the second approach - probabilistic approach - the model is let to allocate the weight for each of the fused records based on the variables not used in the matching process. The weight function is scored based on the similarity/dissimilarity of the common attributes that were not used for fusion. The weight score is expected to be higher for records with higher similarity. In this process, the records that offer the largest improvement in prediction will have higher weights. Finally, given the inherent random nature of the fusion process, the fusion process is repeated for a fixed number of fused records multiple times to ensure that the results are reliable. The model parameters across these samples are compared using a modified Wald t-test to ensure parameter stability. After establishing parameter stability, the fused dataset and the model is finalized. The complete data fusion process is summarized in Figure 1 (see (Bhowmik, et al., 2024a) for similar research). In addition, an example of data fusion process is presented in Figure 2.

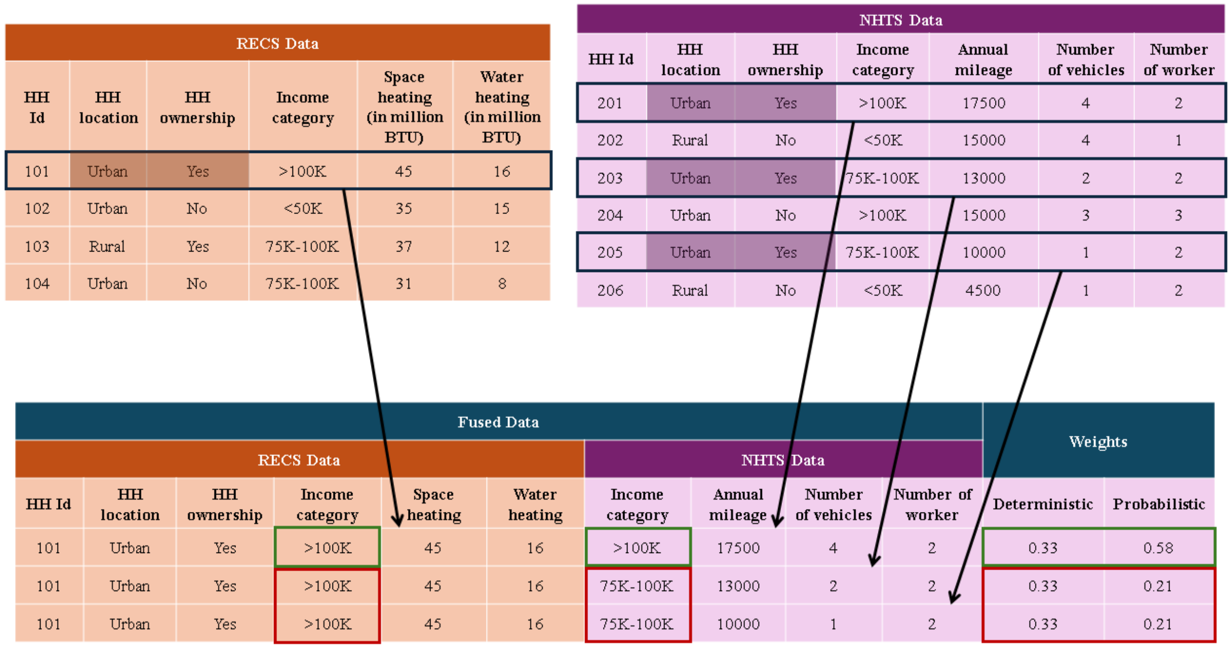

Figure 2: Illustration of data fusion between RECS and NHTS datasets

From Figure 2, it can be observed that the RECS data and the NHTS data do not share a common identifier to merge them. While considering household location and ownership as matching variables, three cases can be found in the NHTS dataset for the first case of the RECS data. When the three NHTS records are fused to the RECS record, the energy consumption record in the RECS data repeats three times. Hence, a weight variable is considered to ensure that the RECS record accounts for only one record. In the case of deterministic weight, all the cases in the fused dataset are assigned an equal weight (1/3). However, in the case of probabilistic weights, the variable income category is considered for allocating the weights because the income category is present in both datasets but is not used in the matching process. It can be noticed that, only for the first fused record the income category matches in both datasets. Thus, the weight score for this record is computed to be higher (for illustration). In our proposed framework, the appropriate weight allocation approach will be selected based on the impact of the weight variable (income category in our example) on model performance.

3 Modeling Methodology

With the preliminaries provided on the data fusion process, the methodological framework of the MDCEV model that is adopted in our study to estimate household energy demand for various end-uses with fused data is outlined here.

3.1 MDCEV Fusion Model

The fusion model structure includes a decision component and a weight component. In the decision model component, the MDCEV model formulation is employed for estimating the household energy demand for different end-uses (see (Bhat, 2005, 2008) for more details).

Consider, there are $n(= 1,2,3, \dots, N)$ households in the RECS dataset and s possible matches from the NHTS dataset. Let $t_{n,k}$ be the total energy consumed by the n^{th} household for the end-use type k ($k =$

$1,2, \dots, K)$. With these notations, the utility equation over the $K(=9)$ end-use types will take the following form (see (Bhat, 2005, 2008) for detailed explanation about MDCEV model):

$$U_{ns} = \sum_{k=1}^{K} \frac{\gamma_k}{\alpha_k} \left[exp\left(\beta' x_{n,k} + \varphi' y_{ns,k} + \varepsilon_{ns,k}\right)\right] . \left\{ \left(\frac{t_{n,k}}{\gamma_k} + 1 \right)^{\alpha_k} - 1 \right\} \tag{1}$$

where, U_{ns} represents the utility of the household n for the s^{th} fused record. $x_{n,k}$ is a vector of attributes from the RECS dataset that characterize end-use alternative k for the n^{th} household. $y_{ns,k}$ is a vector of attributes from the NHTS data that influence the energy demand. β and φ represent vectors of the corresponding parameters to be estimated. $\varepsilon_{ns,k}$ captures idiosyncratic characteristics that impact the baseline utility for the corresponding alternative; and γ_k and α_k are satiation parameters associated with end-use type k. Now in the presence of outside end-use types, the utility equation will take the following form:

$$U_{ns} = \sum_{k=1}^{Q} \frac{1}{\alpha_k} exp\left(\varepsilon_{ns,k}\right) t_{n,k}{}^{\alpha_k} + \sum_{k=Q+1}^{K} \frac{\gamma_k}{\alpha_k} \left[exp\left(\beta' x_{n,k} + \varphi' y_{ns,k} + \varepsilon_{ns,k}\right)\right] . \left\{ \left(\frac{t_{n,k}}{\gamma_k} + 1 \right)^{\alpha_k} - 1 \right\} \tag{2}$$

where, Q is the number of outside end-use types. The reader would note that, there is no translation parameters γ_k for the outside end-use types, as energy is always consumed for them (Bhat, 2008). Now, due to the common role of α and γ, estimating both in empirical application is challenging. Therefore by imposing selective restrictions on α and γ separately, several functional forms of the utility equation (see (Bhat, 2008; Bhat & Eluru, 2010) for the different functional forms) are examined in this study and the γ −profile of Equation 3 is found to provide the best data fit.

$$U_{ns} = \sum_{k=1}^{Q} exp\left(\varepsilon_{ns,k}\right) ln\left(t_{n,k}\right) + \sum_{k=Q+1}^{K} \gamma_k \left[exp\left(\beta' x_{n,k} + \varphi' y_{ns,k} + \varepsilon_{ns,k}\right)\right] . ln\left(\frac{t_{n,k}}{\gamma_k} + 1 \right) \tag{3}$$

In the MDCEV model framework the utility U_{ns} is derived by allocating the total energy among different end-uses. The optimization problem can be solved by forming Lagrangian function for the usage constraint and subsequently applying Kuhn-Tucker first order conditions (Iraganaboina & Eluru, 2021; Wales & Woodland, 1983). Therefore, the probability that a household allocates total energy consumption across the first M of the K end-use types can be expressed as follows:

$$P_{ns}(t_1{}^*, t_2{}^*, \dots, t_M{}^*, 0,0, \dots, 0) = \frac{1}{\sigma^{M-1}} \left[\prod_{m=1}^{M} f_{n,m} \right] \left[\sum_{m=1}^{M} \frac{p_{n,m}}{f_{n,m}} \right] \left[\frac{\prod_{m=1}^{M} e^{V_{ns,m}}}{\left(\sum_{k=1}^{K} e^{V_{ns,k}}\right)^M} \right] (M-1)! \tag{4}$$

where,

$$f_{n,m} = (1 - \alpha_m)/(t_{n,m}{}^* + \gamma_m) \tag{5}$$

$$V_{ns,k} = \beta' x_{n,k} + \varphi' y_{ns,k} + \varepsilon_{ns,k} - ln\left({t_{n,k}{}^*}/{\gamma_k} + 1 \right) - l\,n\left(p_{n,k}\right) \tag{6}$$

$p_{n,k}$ is the total expenditure for the k^{th} end-use alternative by the n^{th}household.

3.2 Weight Component

The weight component takes the form of a latent multinomial logit structure allocating the probability for each NHTS household being paired with the RECS household. The matched weightage propensity is determined based on a latent probability value estimated using a multinomial logit model as follows:

$$w_{ns} = \frac{exp(\zeta z_{ns})}{\sum_{s=1}^{S} exp(\zeta z_{ns})} \tag{7}$$

where z_{ns} is a vector of attributes for RECS household n and fused record s that influences the propensity of matching the NHTS dataset with the RECS dataset. z_{ns} represents the variables that are present in both datasets but not used for fusion. ζ is the corresponding coefficient vector to be estimated.

3.3 Model Estimation

Based on above notation, the overall weighted probability for each household in the RECS data can be computed as follows:

$$L_n = \sum_{s=1}^{S} w_{ns} * P_{ns} \tag{8}$$

where L_n is the weighted probability of the household n in the RECS dataset. The parameters to be estimated in the proposed MDCEV fusion model of Equation 8 include the β vector, the φ vector, the γ vector and the ζ vector. The maximum likelihood inference approach is used to estimate the parameters. The log-likelihood function for the model estimation is defined as:

$$LL = \sum_{n=1}^{N} \log(L_n) \tag{9}$$

The analysis is conducted using Gauss Matrix Programming software (see (Bhat, 2008) for a brief explanation of the model estimation approach).

4 Data Description

The 2020 Residential Energy Consumption Survey (RECS, 2020) data and the 2022 National Household Travel Survey (NHTS, 2022) data were utilized in our analysis. The fusion of datasets from two different years might ignore the impact of changing socioeconomic characteristics, household energy demand and the impact of major events like Covid-19 pandemic. However, as presented in Figure 3, the socioeconomic indicators, such as household size, household income, home ownership, household location and housing types remain relatively stable across the two years. Further, the main objective of this study is to develop an energy demand framework that can estimate the impact of travel infrastructure and usage related attributes on household end-use energy demand and remedy the over-estimation of the impact of socioeconomic attributes. Considering the documented contribution of the fusion approach (as measured objectively in terms of model improvement), the slight temporal mismatch between the two datasets does not affect the value of the fused dataset for model development. The RECS dataset is comprised of household energy consumption records of 18,496 households from the entire USA. The data provides information about household socioeconomic attributes and total energy consumption by different end-uses. Among the 18,496 household records, 5,000 households are randomly selected for model estimation to avoid model overfitting and manage computational resources. The remaining 13,496 household records are employed for assessing parameter stability. The 2022 NHTS data is used for detailed travel infrastructure and usage related information. The dataset comprises records of 7,893 households from the entire USA. Among these household records, 7,417 households are considered for model estimation after removing the data with missing values.

4.1 Dependent Variable

Energy demand of household end-uses is considered as the variable of interest in this study. The dependent variable is obtained from the 2020 RECS data. In our study, the household energy end-uses are classified into 9 categories such as – (1) space heating (including space heater), (2) water heating (including water heater, boiler pump and hot tub heater), (3) refrigeration (including refrigerator and freezer), (4) cooling and ventilation (including air conditioner, evaporative cooler, ceiling fan, humidifier, and dehumidifier), (5) cooking (including stove, cooktop, and oven), (6) washing and drying (including washing machine, dryer, and dishwasher), (7) EV charging, (8) lighting (including indoor and outdoor lights) and (9) miscellaneous activities (including television, water pump, desktop, computer, laptop, cell phone, smart phone, printer, scanner, and other undefined categories). The distribution of the dependent variable categories is shown in Table 1. Based on the average energy consumption, the table represents that space heating is the highest energy consuming end-use category, which is followed by water heating and EV charging (only for the households that own EV). Two end-use categories – lighting and miscellaneous activities – are found to be consumed by all the households, as shown in Table 1. Therefore, these two categories are used as the outside categories in our MDCEV model (see (Bhat, 2005) for the explanation of outside category).

Table 1: Distribution of different household energy end use categories

Variable	Participation Rate (%)	Minimum	Maximum	Mean	Standard deviation
Space heating	95.4	0.0 (0.0)	407906.1 (119550.4)	40051.8 (11738.5)	38116.8 (11171.4)
Water heating	99.4	0.0 (0.0)	162555.6 (47642.3)	15591.5 (4569.6)	11999.2 (3516.7)
Refrigeration	99.6	0.0 (0.0)	18348.5 (5377.6)	3657.5 (1071.9)	2246.4 (658.40
Cooling and ventilation	95.4	0.0 (0.0)	131030.3 (38402.7)	9501.0 (2784.5)	9658.1 (2830.6)
Cooking	99.8	0.0 (0.0)	19832.3 (5812.5)	1852.8 (543.0)	1425.8 (417.9)
Washing and drying	92.6	0.0 (0.0)	21092.0 (6181.7)	2694.0 (789.5)	2275.3 (666.8)
Lighting	100.0	13.3 (3.9)	46486.0 (13624.2)	2335.1 (684.3)	2888.3 (846.5)
EV charging	1.2	0.0 (0.0)	77599.6 (22743.1)	10423.6 (3055.0)	12334.4 (3615.0)
Miscellaneous activities	100.0	57.1 (16.7)	779554.2 (228474.2)	9917.6 (2906.7)	17236.7 (5051.8)

Note:
1. Participation rate represents the percentage of households with non-zero energy usage. The mean value of energy use is computed only for households with non-zero usage.
2. Energy consumptions are presented in 1000 BTU (kWh)

4.2 Independent Variables

A set of independent variables including household socioeconomic attributes (such as household size, location, and income) and travel infrastructure and usage related attributes (such as number of vehicles of different fuel type and body type, number of drivers, annual mileage and frequency of the usage of intercity buses for long distance trips) are employed in our study. The socioeconomic attributes are obtained from the RECS dataset, while the travel infrastructure and usage related attributes are obtained from the NHTS dataset. The fusion between the RECS and the NHTS datasets is conducted using the seven common variables present in both datasets. The distributions of these variables in both the datasets are presented in Figure 3. In the figure, the variables across the datasets represent very similar distributions. It can be observed that, in the RECS dataset, 67.2% households are from urban areas, and the share is 70.4% in the NHTS data. Further, in both RECS and NHTS datasets, majority of the households are composed of two members, and the share is 38.1% and 43.6% respectively. Households with three members represent the lowest share in both datasets. Regarding household ownership, both datasets represent that the majority of households own their own homes, and their share is 73.2% in the RECS data and 78.4% in the NHTS data. The household income distributions exhibit slight differences between the two datasets. In the RECS data, households with incomes between $25,000 and $49,999 make up the largest share (20.2%), whereas, in the NHTS data, the largest share is constituted by the households with incomes between $100,000 and $149,999 (20.1%). However, their shares in the two datasets are relatively similar. For instance, households with incomes between $25,000 and $49,999 constitute 20.2% in the RECS data and 17.6% in the NHTS data. Regarding the location of the household, the figure represents that, in both datasets, majority of the households are located in the south region, while the smallest share of households is located in the northeast region. Further, in both datasets, most of the households are from the South Atlantic division and their shares are 16.8% and 20.4% respectively. Finally, regarding housing type, the figure represents that most households in both datasets live in single family detached homes. They constitute 66.4% of the RECS data and 71.9% of the NHTS data. The readers should note that if the variables are present in both RECS and NHTS datasets, data available in RECS data are employed for model estimation.

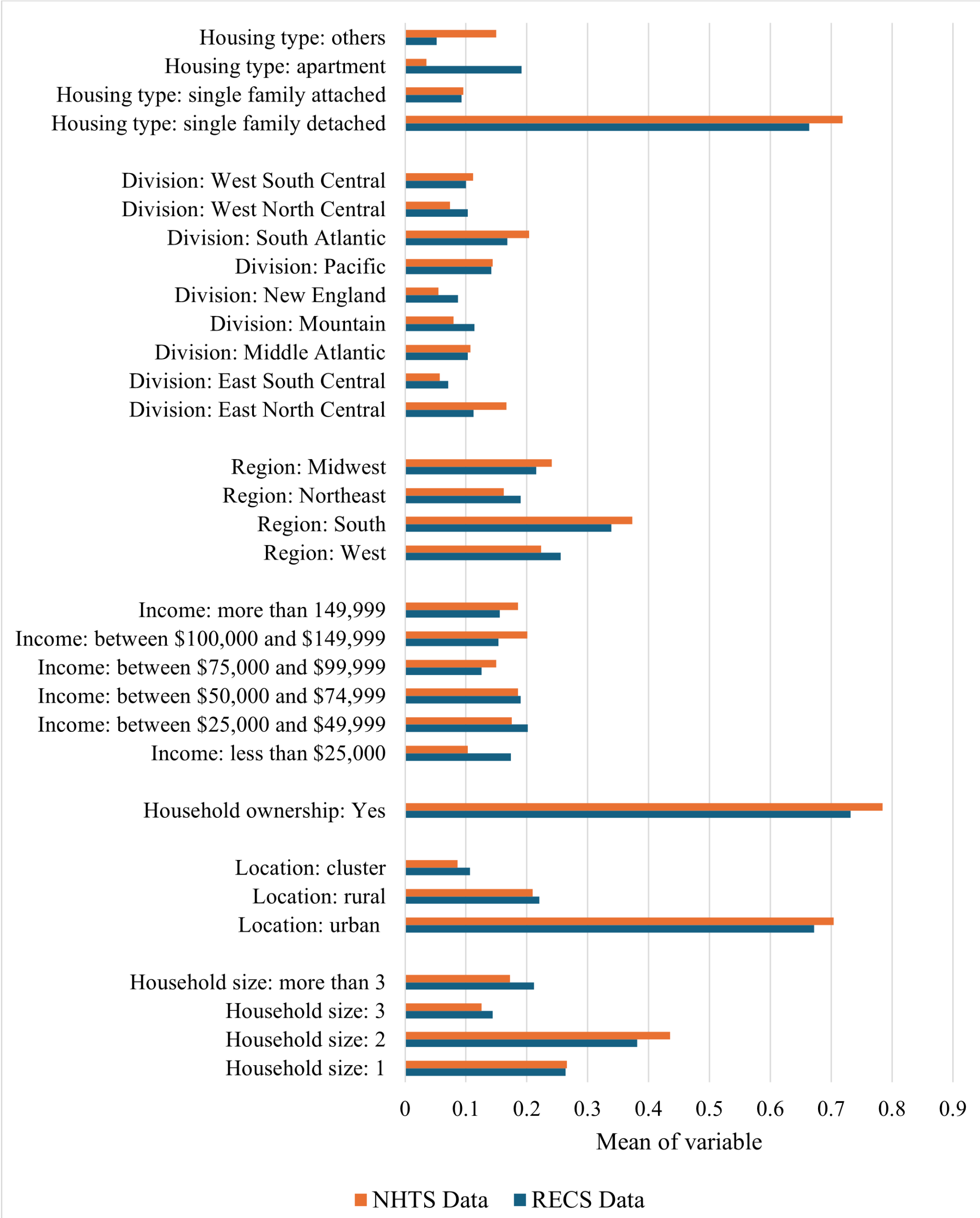

Figure 3: Distribution of common attributes present in both RECS and NHTS datasets

Further, the distributions of various independent variables are presented in Table 2. From the distribution of household socioeconomic attributes in the RECS data it can be observed that, 67% households are located in urban areas, 22% households are located in rural areas, and the rest of the households are located in urban cluster areas. The income distribution shows that 17% households have income less than $25,000, 20% households have income between $25,000 and $49,999, 19% households have income between $50,000 and $74,999, 13% households have income between $75,000 and $99,999, 15% households have income between $100,000 and $149,999 and the rest of the households have income more than $149,999. It is also noticeable that 66% households live in detached houses, 9% households live in attached houses as single families, 19% households live in apartments and the rest 5% live in other housing arrangements. Further, the household ownership distribution shows that 73% of the households live in their own houses. Across all the households, only 2% households are noticed to own EVs in their house (reflective of EV share in 2020). In addition to these binary variables, several count variables (such as household size, number of children and number of adults) are employed from the RECS data in our study.

NHTS data provides detailed information about household travel infrastructure and usage related attributes, such as number of drivers in the household, total number of vehicles in the households, number of vehicles of different fuel types and body types, total annual mileage, number of annual long-distance trips and number of intercity bus or train trips in a year for long-distance tours. Additionally, several socioeconomic attributes that are not present in the RECS data, are also employed from the NHTS data to obtain potential improvement in model fitting. These variables include the number of workers in the household, the number of people of different gender and household ethnicity. The distributions of these variables are presented in Table 2. All the variables presented in Table 2 are tested in model estimation. From this large set of variables, those that offer significant impact on the household energy end-uses at 90% confidence level (t-value ≥ 1.65) are retained in the final specification.

Table 2: Independent variables employed for model estimation

Variable	Minimum	Maximum	Mean	Standard Deviation
RECS Data (N = 5,000)				
Household size	1.0	7.0	2.4	1.4
Number of children	0.0	4.0	0.5	0.9
Number of adults	1.0	7.0	1.9	0.9
Location: urban	0.0	1.0	0.7	0.5
Location: rural	0.0	1.0	0.2	0.4
Location: cluster	0.0	1.0	0.1	0.3
Income: less than $25,000	0.0	1.0	0.2	0.4
Income: between $25,000 and $49,999	0.0	1.0	0.2	0.4
Income: between $50,000 and $74,999	0.0	1.0	0.2	0.4
Income: between $75,000 and $99,999	0.0	1.0	0.1	0.3
Income: between $100,000 and $149,999	0.0	1.0	0.2	0.4
Income: more than 149,999	0.0	1.0	0.2	0.4
EV availability	0.0	1.0	<0.1	0.1
Housing type: single family detached	0.0	1.0	0.7	0.5
Housing type: single family attached	0.0	1.0	0.1	0.3
Housing type: apartment	0.0	1.0	0.2	0.4
Housing type: Others	0.0	1.0	0.1	0.2
Region: West	0.0	1.0	0.3	0.4
Region: South	0.0	1.0	0.3	0.5
Region: Northeast	0.0	1.0	0.2	0.4
Region: Midwest	0.0	1.0	0.2	0.4
Division: East North Central	0.0	1.0	0.1	0.3
Division: East South Central	0.0	1.0	0.1	0.3
Division: Middle Atlantic	0.0	1.0	0.1	0.3
Division: Mountain	0.0	1.0	0.1	0.3
Division: New England	0.0	1.0	0.1	0.3
Division: Pacific	0.0	1.0	0.1	0.4
Division: South Atlantic	0.0	1.0	0.2	0.4
Division: West North Central	0.0	1.0	0.1	0.3
Division: West South Central	0.0	1.0	0.1	0.3
Household ownership: Yes	0.0	1.0	0.7	0.4
NHTS Data (N = 7,417)				

Number of drivers in the household	1.0	7.0	1.8	0.7
Number of vehicles in the household	1.0	17.0	2.0	1.1
Number of sedan cars	0.0	6.0	0.9	0.8
Number of vans	0.0	7.0	0.1	0.3
Number of SUVs	0.0	5.0	0.6	0.7
Number of pickup trucks	0.0	5.0	0.3	0.6
Number of motorcycles	0.0	6.0	0.1	0.3
Number of other type of vehicles	0.0	17.0	<0.1	0.3
Number of gasoline vehicles	0.0	14.0	1.8	1.0
Number of electric vehicles	0.0	2.0	<0.1	0.2
Number of hybrid cars	0.0	4.0	0.1	0.3
Number of other fuel vehicles	0.0	17.0	0.1	0.3
Household total annual mileage	2100.0	1300000.0	85678.9	219656.8
Number of long-distance trips in a month	0.0	414.0	3.7	13.5
Frequency of intercity (Amtrak) train trips in a year	0.0	24.0	0.1	0.8
Frequency of usage of intercity buses in a year	0.0	61.0	0.1	1.1
Number of workers in the household	0.0	6.0	1.0	0.9
Number of males	0.0	8.0	1.1	0.8
Number of females	0.0	7.0	1.1	0.8
Caucasian American households	0.0	1.0	.8	0.4
African American households	0.0	1.0	.1	0.3
Households of other races	0.0	1.0	.1	0.2
Mixed race households	0.0	1.0	.1	0.3

5 Model Selection with Fused Data

In this section, the optimal fusion model is selected following the algorithm presented in Figure 1. *In the first step*, various variable combinations using the 7 common variables present in both RECS and NHTS datasets are identified. These variables include household Division (D), household Income (I), household Location (L), household Ownership (O), household Region (R), household Size (S) and housing Type (T). Based on these common variables, in our study, 9 variable combinations are examined for data fusion: (1) D-I-L-O-R-S, (2) D-I-L-O-S-T, (3) D-I-L-O-R-T, (4) D-I-L-R-S-T, (5) D-I-R-S-T, (6) D-L-O-R-S, (7) D-L-O-R-S-T, (8) D-L-O-S-T and (9) I-L-O-R-S-T. *In the second step*, for each combination 5-, 10-, 15-, 20 and 25 NHTS records were randomly fused to each RECS record, and the MDCEV model was developed with the fused data[3]. *In the third step*, the improvement of the model log-likelihood (LL) (relative to the model with only RECS dataset) at each combination was compared. The LL improvement of the fusion model is computed as follows: $LL\ improvement = LL_{fusion\ model} - LL_{RECS\ only\ model}$. Based on the LL improvement, the optimal number of fusion records for each variable combination was selected. Figure 4 represents the LL improvements for different number of matching variables for different variable combinations. The figure indicates the optimal number of fusion records for all variable combinations.

[3] The reader would note that if a particular RECS record does not have the target number of records, all the available records from NHTS are considered for the fusion.

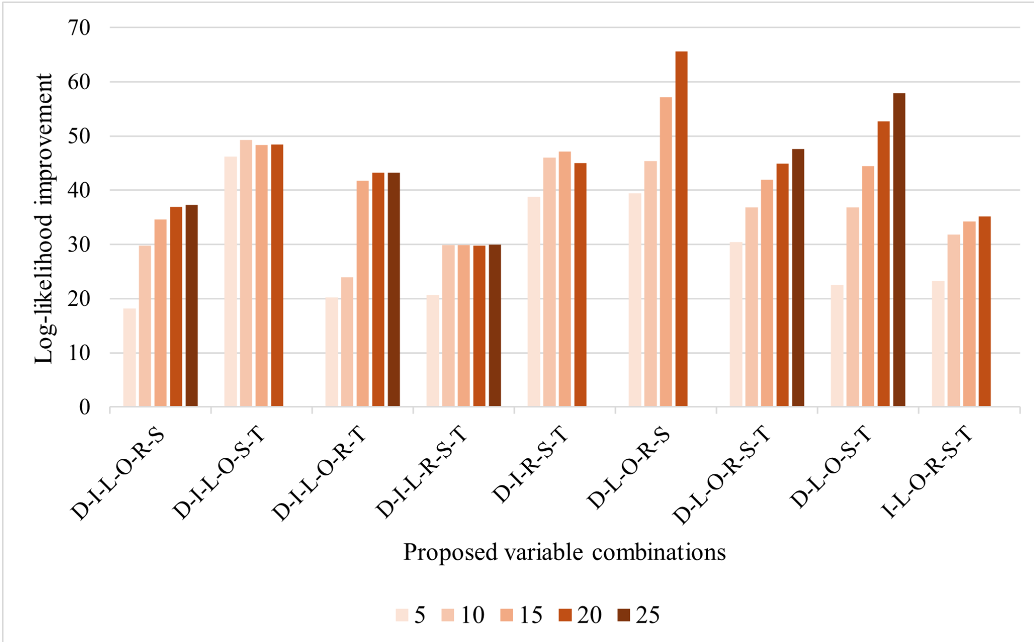

Figure 4: Selection of optimal number of fusion records for all combinations

In the fourth step, the optimal models for all variable combinations were compared, and the final optimal fusion model was selected. Figure 5 shows the LL improvement of all the optimal models for all variable combinations.

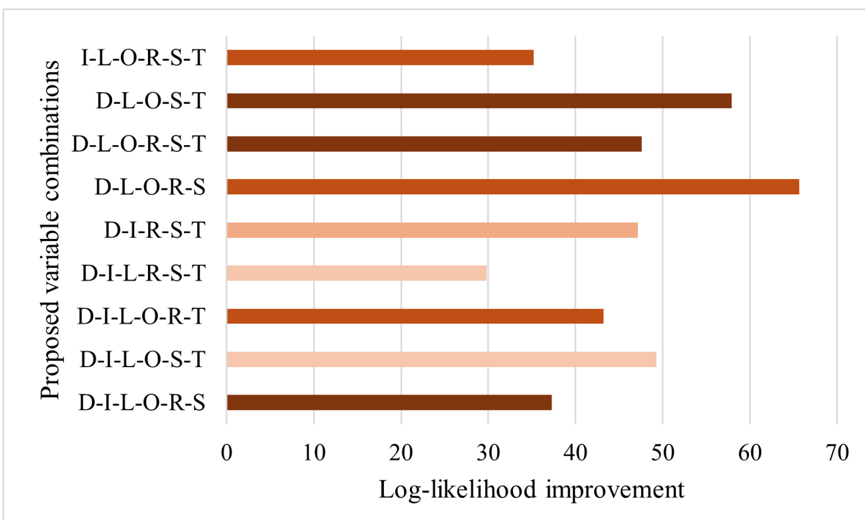

Figure 5: Selection of optimal model combination

The optimal model for the combination D-L-O-R-S is found to offer the best model with the highest LL improvement. From the Figure 4 it is noticeable that, for the combination D-L-O-R-S the model provides the highest LL improvement with 20 fusion records. It is noticeable that, with an increase in the number of fusion sizes the LL improvement also increases. However, increasing the number of fusion records requires increased computational time. It is also noticeable that, at fusion size = 20 for the combination D-L-O-R-S, the model provides the highest LL improvement across all the models for different variable combinations[4]. Thus, the proposed model framework was estimated fusing 20 NHTS records to each RECS record based on the D-L-O-R-S combination. *In the final step*, the parameter stability of the optimal model is examined. Since the 20 NHTS records were selected randomly for data fusion, it is important to ensure the random sampling does not impact the stability of estimates. For this reason, five random samples are created, each consisting of 5,000 random cases drawn from the 13,496 RECS records that were not utilized for model estimation. The random samples were created following the same process used for the estimation sample. For all these samples the optimal fusion model specifications were estimated. The average of the parameters across the samples is considered the population estimate. Subsequently, a revised Wald t-test statistic is computed for each sample parameter relative to population mean parameter as follows:

$$t - statistic = \left| \frac{\beta - \bar{\beta}}{\sqrt{\sigma^2 + \bar{\sigma}^2}} \right| \tag{10}$$

where, β is the sample parameter vector, $\bar{\beta}$ is the mean of the parameter vectors across the samples, σ is the Standard Deviation (SD) vector and $\bar{\sigma}$ is the population SD vector computed as the mean of the SD vectors across the samples.

The distribution of t-statistics of all the variables estimated are presented in Figure 6. To be sure, it is not expected that the parameter estimates are identical. However, the focus is on examining if the parameters across these multiple samples exhibit statistically significant variability. The figure shows that, the parameter test statistics for all variables are less than critical t statistic value for 90% confidence interval (1.65). This implies that sample randomness does not affect the parameter stability across the samples.

[4] For the combination D-L-O-R-S, the BIC value of the RECS only model is 649,219 and at fusion size = 20, the BIC value of the fusion model is 649,198. The BIC values also indicate the improved predictability of the fusion model.

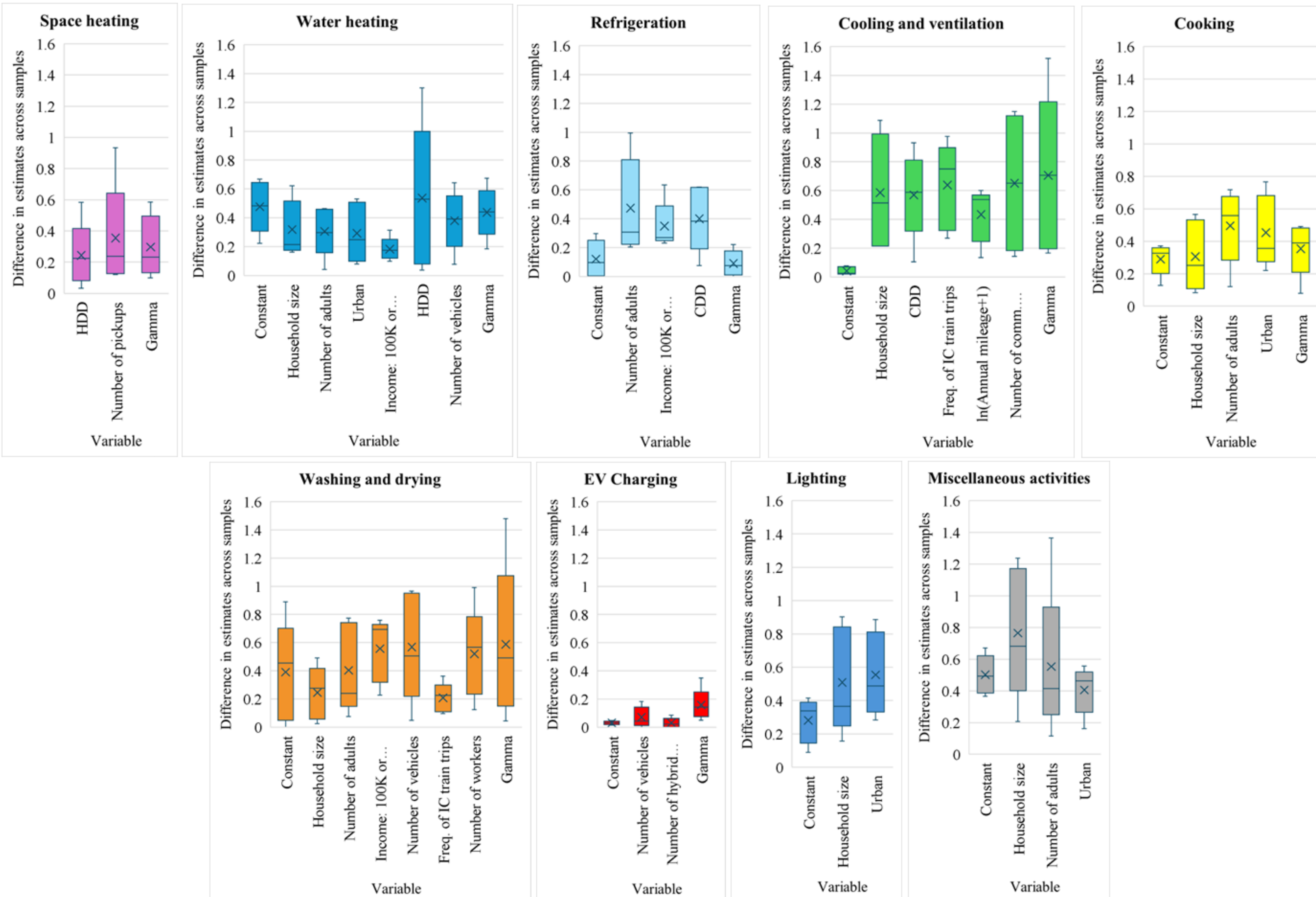

Figure 6: Test statistics for parameter estimates across samples

Table 3: Estimation results of the optimal MDCEV fusion model

Parameter	Space heating	Water heating	Refrigeration	Cooling and ventilation	Cooking	Washing and drying	EV charging	Lighting	Miscellaneous activities
	Estimates (t-value)								
Constant	---	1.75 (12.40)	2.96 (12.86)	0.11 (0.81)	3.40 (9.68)	0.14 (2.22)	-1.01 (-2.29)	5.82 (103.82)	7.38 (135.06)
Variables from the RECS data: *Socioeconomic attributes*									
Household size	---	0.15 (9.73)	---	0.10 (6.45)	0.06 (4.36)	0.16 (8.89)	---	0.08 (5.95)	0.07 (5.47)
Number of adults	---	0.05 (1.99)	0.08 (4.73)	---	0.06 (2.27)	0.05 (1.88)	---	---	0.12 (5.16)
Urban area	---	0.19 (7.57)	---	---	0.10 (3.84)	---	---	0.17 (4.96)	0.08 (3.22)
Income: $100,000 or more	---	-0.17 (-6.99)	-0.11 (-5.75)	---	---	0.10 (3.73)	---	---	---
Variables from the RECS data: *Meteorological attributes*									
HDD (in 1000)	0.25 (35.72)	0.06 (10.64)	---	---	---	---	---	---	---
CDD (in 1000)	---	---	0.04 (5.00)	0.47 (31.14)	---	---	---	---	---
Variables from the NHTS data: *Travel infrastructure and usage related attributes*									
Number of vehicles	---	-0.02 (-1.80)	---	---	---	0.15 (13.11)	0.29 (1.46)	---	---
Number of pickup trucks	-0.11 (-3.43)	---	---	---	---	---	---	---	---
Number of hybrid vehicles	---	---	---	---	---	---	0.51 (2.36)	---	---
Number of vehicles used for business purpose	---	---	---	-0.69 (-2.65)	---	---	---	---	---
Frequency of intercity train trips	---	---	---	-0.23 (-1.26)	---	-0.04 (-1.51)	---	---	---
ln (annual mileage +1)	---	---	---	0.03 (2.87)	---	---	---	---	---
Variables from the NHTS data: *Socioeconomic attributes*									
Number of workers	---	---	---	---	---	-0.06 (-3.43)	---	---	---
Satiation parameter									
Gamma	3352.55 (19.83)	386.68 (6.85)	54.08 (4.54)	642.67 (11.14)	12.93 (2.87)	287.80 (31.00)	2359.32 (4.72)	---	---

6 Model Estimation Results

In this section the estimation results of the optimal fusion model are described. The specification of the model is presented in Table 3, and the model estimates are discussed by different variable categories. The reader would note that a positive (negative) coefficient for a parameter in the MDCEV model is likely to increase (decrease) the utility for that household from participating in the end-use alternative energy consumption for the alternative relative to the base alternative.

6.1 Variables from the RECS Data

6.1.1 Socioeconomic attributes

In this study, several household socioeconomic attributes are employed from the RECS dataset. Among them, several attributes are found to offer significant impacts on household energy end-uses. It can be observed that, with the increase in household size, energy demand for washing and drying are likely to increase by the highest amount. Further, with the increase in the number of adults, the highest increment of energy demand is likely to be seen in washing and drying. The reader would note that the impact of household size and number of adults needs to be considered together. The actual impact of the number of adults is represented by summing up the impacts of household size and number of adults. Thus, the impacts of number of adults on the energy demand for washing and drying, and miscellaneous activities are (0.16+0.05) and (0.07+0.12). The parameter estimates indicate that the increase in energy consumption for miscellaneous activities will be the highest for an additional adult member compared to an additional child. The parameter associated with urban area indicates that households in urban areas are likely to consume the highest amount of energy for water heating. The result implies the dependency of urban residents on the appliances that rely on hot water (such as showers, washing machines, and dish washers). The impact of household income indicates that household with income of $100,000 or more are also likely to consume the highest amount of energy for washing and drying.

6.1.2 Meteorological attributes

To evaluate the impact of different weather conditions, two meteorological attributes – Heating Degree Days (HDD) and Cooling Degree Days (CDD) – are considered in the proposed fusion framework. With an increase in HDD household's energy consumption for space heating is likely to increase by the highest amount. Further, household's energy consumption for cooling and ventilation is likely to increase with an increase in CDD. The results indicate that households from the regions of higher HDD are likely to consume the highest amount of energy for space heating. Households from the regions of higher CDD are likely to consume the highest energy for cooling and ventilation.

6.2 Variables from the NHTS Data

6.2.1 Travel infrastructure and usage related attributes

Several travel infrastructure and usage related attributes, employed from the NHTS dataset, are found to be significant in our model estimation. An increase in the number of vehicles is associated with increased energy consumption for EV charging. The impact of vehicle type shows that with an increase in the number of pick-up trucks, households are likely to consume the least amount of energy for space heating. The result might suggest a lifestyle that involves less time spent indoors and possibly lowers the demand for space heating. The impact of fuel type shows that with the increase in number of hybrid vehicles, households' energy consumption for EV charging is likely to increase.

Furthermore, it can be seen from the table that as the vehicle mileage increases, households are more likely to consume energy for cooling and ventilation. The result indicates the demand for a more comfortable indoor environment after extended period of commuting. Further, the number of intercity train trips represents the likelihood that household members are away from home for a longer period or not. From the table it is noticeable that, with the increasing number of intercity train trips household's energy consumption for cooling and ventilation decreases significantly. Similarly, the numbers of vehicles used for business purposes indicate household members' likelihood of spending more time outside of the home. It can be noticed that with an increase in such vehicles, households are likely to decline their usage of cooling and ventilation appliances.

6.2.2 Socioeconomic attributes

Only one socioeconomic characteristic from the NHTS data – number of workers – are found significant in our model. It can be observed that, with the increase in the number of workers, household energy consumption for washing and drying are noticed to decrease. The results potentially highlight the reduced usage of these end-uses as more household members stay away from home for work activities.

6.2.3 Satiation Parameter

Satiation parameters are estimated only for the non-outside categories in our study. Space heating is found to have the highest satiation implying a rapid drop of its marginal utility for energy usages. This indicates a smaller usage level of energy for space heating. On the other hand, cooking has the lowest satiation value among all alternatives.

6.3 Weight Component

There are seven common variables present in both the RECS and the NHTS datasets. For the fusion model, five of these variables – household division, household location, household ownership, household region and household size – were utilized. The remaining two variables – household income and housing types – are present in both datasets but were not used for fusion. So, two binary indicator variables were created such that the binary variables will take a value of one if the variables match in both datasets and zero otherwise. However, these binary indicator variables were not found to provide any significant impact in our modeling framework. The results indicate that all the NHTS records fused to each RECS record provide similar impact in the model.

7 Elasticity Effect Analysis

The coefficients of the independent variables in Table 3 do not directly provide the exact effect of variables on household's end-use energy demand. The impact of the variables might change across different end-uses. To evaluate this variability, the elasticity of the variables is computed. More specifically, the percentage changes in the energy consumption of various end-use categories are estimated in response to any change in the explanatory variables. For continuous independent variables (such as annual mileage) in the model, the change was conducted in an increment of 10% (see (Bhowmik et al., 2022) for similar analysis). For count variables (such as household size, number of adults and number of vehicles), the change was conducted in an increment of one unit. For indicator variables (such as urban area and Caucasian American household), the change was obtained by changing the value of the variable to one for the subsample of observations for which the variable takes a value of zero and vice-versa (see (Kabli et al., 2020) for similar analysis). The elasticity effects of the variables are shown in Figure 7 and 8. Figure 7 represents the elasticity effects of the RECS variables in both RECS only model (before fusion) and fusion model (after fusion). The exercise was conducted to illustrate how in the absence of travel infrastructure and usage related variables, other variables from RECS are either over-estimated or under-estimated. Figure 8 represents the elasticity effects of the newly added NHTS variables. For the sake of space, the four highest changing end-uses for each of these NHTS variables are presented.

From the figures, several observations can be made. *First*, it can be noticed from Figure 7 that the elasticity of RECS variables with and without NHTS variables changes significantly for some alternatives, for example, the impact of household size on cooling and ventilation. A significant increment of the elasticity can be noticed in the fusion model, while an obvious reduction in elasticity magnitude is expected once NHTS variables are introduced. These results clearly show how mis-specification bias inflates the magnitudes of variables from RECS dataset. *Second*, with an increase in household size, household energy consumption for water heating, washing and drying and lighting experience a significant increase. For one additional household member, energy consumption for water heating increases by 10%, washing and drying increases by 12% and lighting increases by 3%. At the same time, energy consumption for space heating and refrigeration are found to reduce. Further, for an additional adult member in the household, energy consumption for refrigeration increases by 5%, cooking increases by 3% and miscellaneous activities increases by 9%. *Third*, Figure 8 represents the impact of various travel infrastructure and usage related variables on energy end-uses. It can be observed that, with an additional vehicle in the household, the energy demand for EV charging increases significantly. *Fourth*, with an increase in the number of hybrid vehicles in the household, the energy consumption for EV charging also increases. *Fifth*, with an increase in the number of vehicles used for commercial purposes, household energy consumption for EV charging also increases implying an increased utilization of EV(s) for commercial purposes. *Finally*, household energy consumption for cooling and ventilation is noticed to increase with an increment in household annual mileage but decrease with an increased number of long-distance trips.

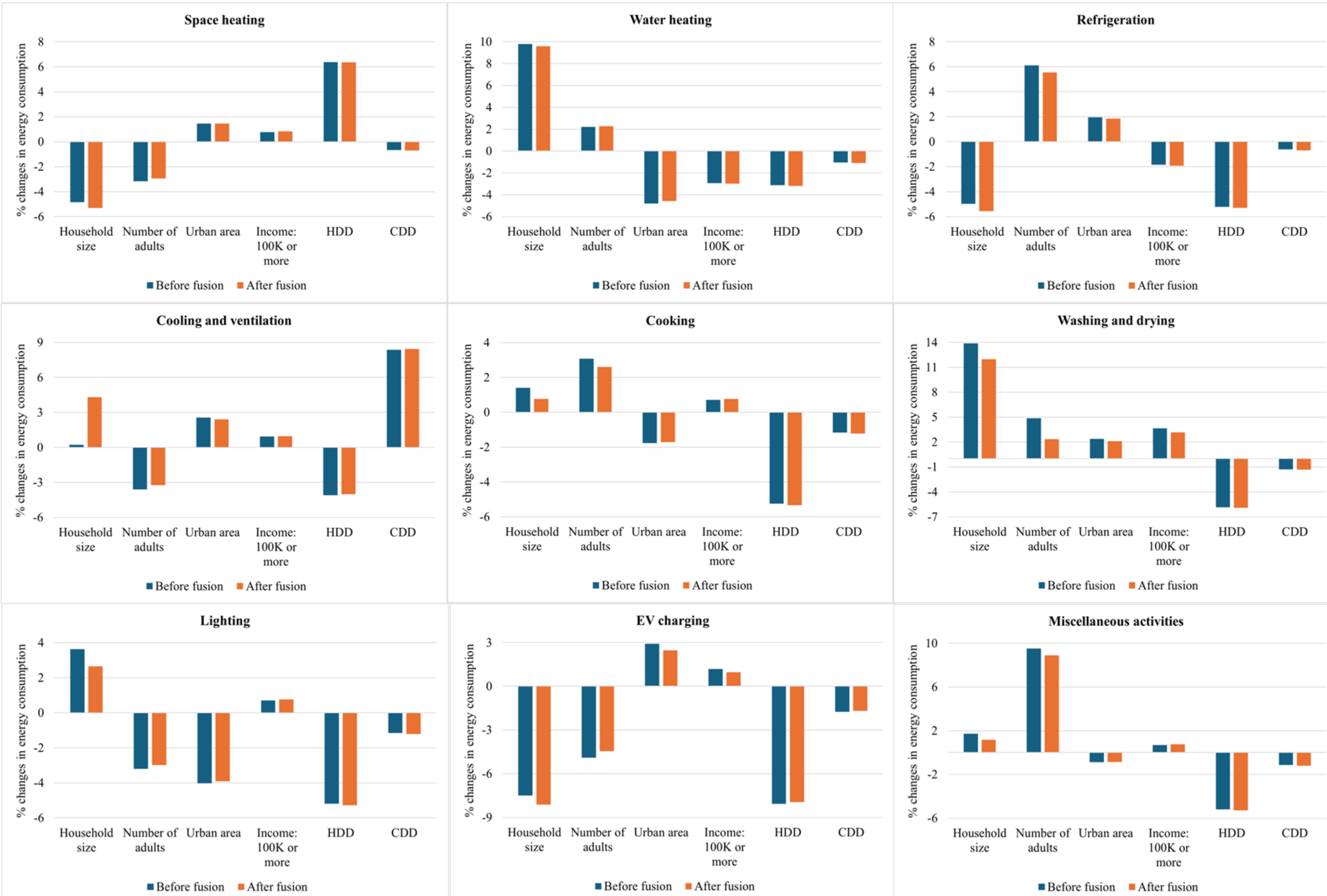

Figure 7: Impact of data fusion on the elasticity effects of the independent variables from RECS data

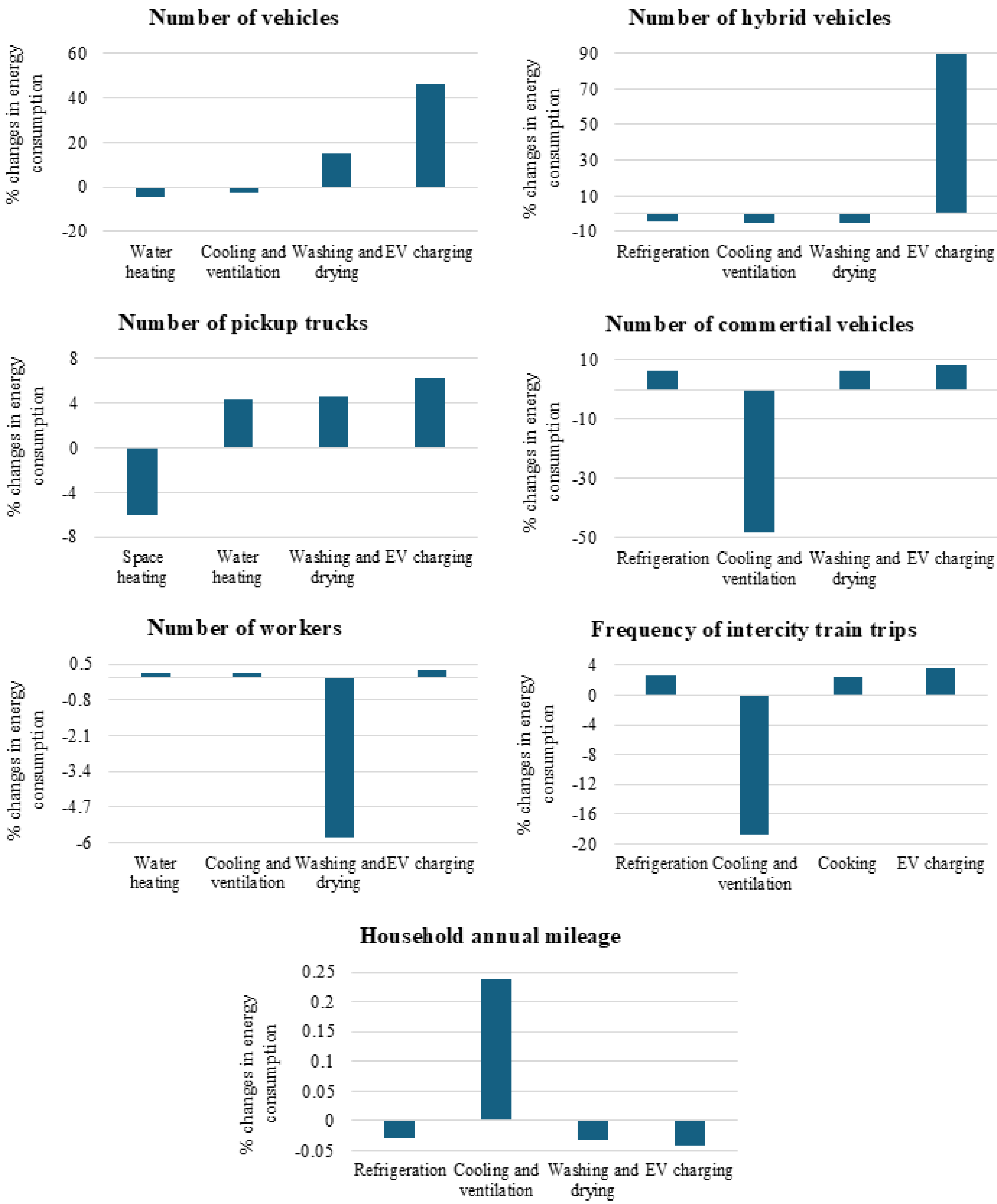

Figure 8: Elasticity effects of variables from NHTS data obtained after data fusion

8 Policy analysis

To illustrate the implementation of our proposed methodology, in this section, the impacts of different weather conditions and EV ownership on household end use energy demand as well as their variations between urban and rural areas are evaluated. To perform these analyses, a synthetic household is considered. The characteristics of the synthetic household are as follows:

- Household size: 5
- Number of adults: 2
- Number of workers: 2
- Household income: more than $100,000
- Number of vehicles: 2
- Number of pickup trucks: 1
- Number of hybrid vehicles: 1
- Number of intercity train trips in a year: 10
- Annual mileage: 30,000
- Number of vehicles used for commercial purpose: 0

To evaluate the impact of different weather conditions, different climatic regions presented in Table 2 are considered and their corresponding HDD and CDD values are employed as inputs in the estimated fusion model from Table 3. As reported by the U.S. Energy Information Administration (EIA), the annual HDD (CDD) of the four regions in the year 2023 are as follows: (a) South region: 2,231(2,353), (b) Northeast region: 5,250 (605), (c) Midwest region: 5,681(875) and (d) West region: 4,347(1158) (EIA, 2024a). Further, to evaluate the impact of EV ownership, the energy consumption for various end-uses is compared under two scenarios: (a) household does not own an EV and (b) household owns an EV. The reader might note that, in the first case, the proposed fusion model in Table 3 is employed to distribute the total household energy across eight end-use categories, excluding the EV charging. In the second case, the total household energy is distributed across all the nine categories. The average household total energy demand in the estimation dataset is considered for this analysis. Finally, to assess the impact of household location, energy demand of the synthetic household is predicted for both urban and rural areas. The impact of different weather conditions and EV ownership in urban and rural areas are presented in Figures 9 and 10.

Several observations can be made from these figures. First, the share of space heating for the synthetic household is the highest in the Midwest region and lowest in the south region. Households in both urban and rural areas follow similar patterns. Second, regarding the energy demand for refrigeration and cooling and ventilation, households in both urban and rural areas of south region consume the highest energy. Third, the impact of EV ownership reveals that the synthetic household in the south region consume more energy for EV charging than other regions. Further, for households in all climatic regions, energy demand for EV charging causes significant changes in the energy consumption for space heating and water heating. EV charging is not found to cause significant changes in other end-use categories. Fourth, compared to other end-use categories, energy demand for space heating and water heating exhibits significant variations between urban and rural areas. Finally, households in rural areas are found to consume slightly more energy for EV charging than households in urban areas.

The results provide valuable insights for a diverse range of stakeholders, including energy policymakers, grid operators, urban and rural planning authorities, renewable energy developers, EV manufacturers, real estate developers, and academic researchers. Policymakers and grid operators can utilize these findings to guide energy efficiency programs, assess regional variations in energy demand in response to emerging climate trends, plan infrastructure for EV charging, forecast energy loads, and strategize future capacity expansions. The study also highlights differences in energy consumption between urban and rural households, enabling planners to design more effective energy distribution strategies. The proposed framework can be a potential tool for evaluating the variations in energy demand across different end uses which can aid the renewable energy developers and product manufacturers to build an energy efficient system. Further, understanding regional energy demands for EV charging helps manufacturers in developing their EV marketing strategies and product development. Finally, the findings provide valuable information to the academic researchers for further studies on residential energy consumption, EV integration, and climate-specific energy behavior.

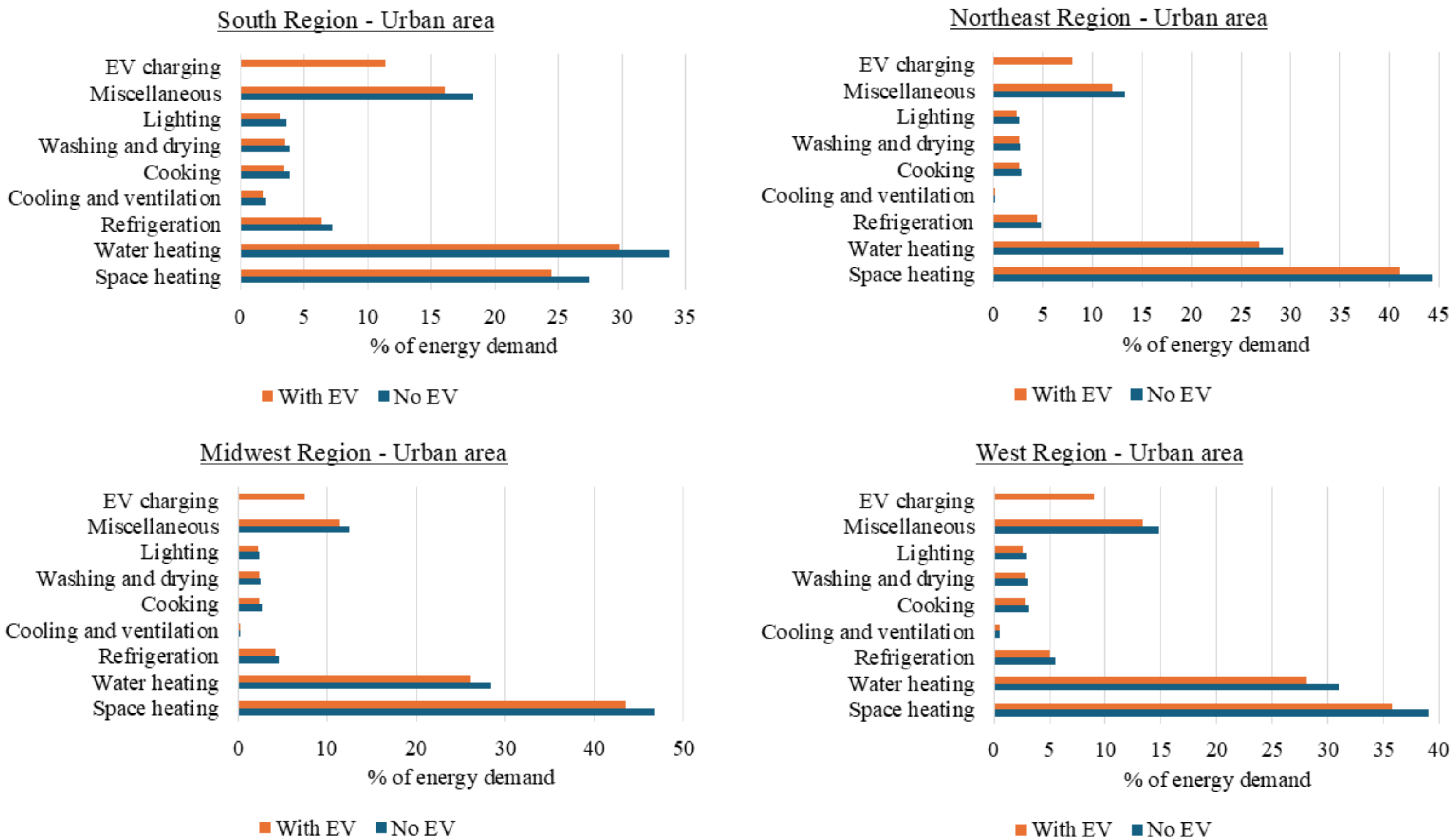

Figure 9: Impact of weather conditions and EV ownership in urban area

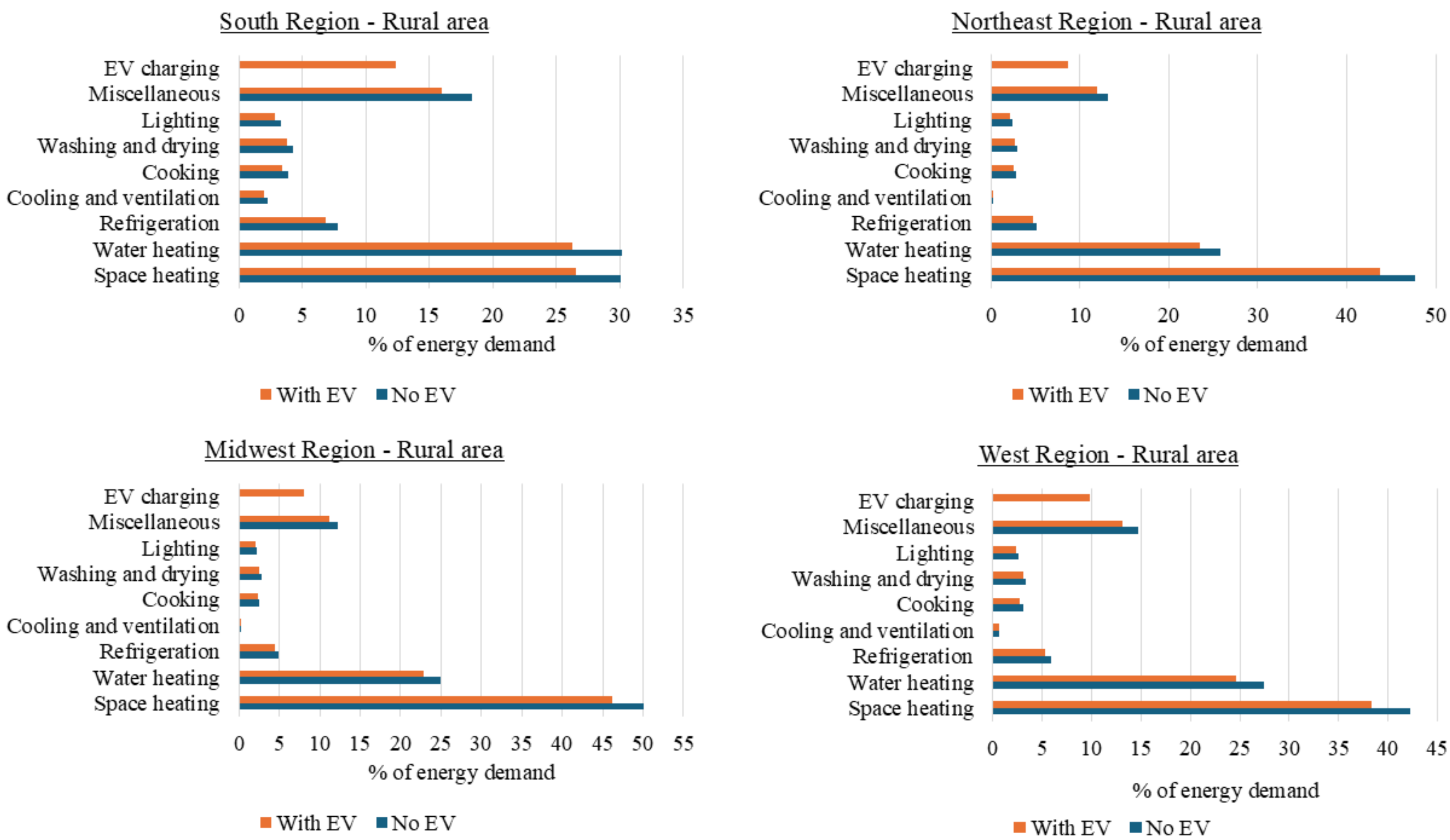

Figure 10: Impact of weather conditions and EV ownership in rural areas

9 Conclusion

Earlier studies examining residential sector energy demand revealed several important determinants of energy demand including – (a) household sociodemographic attributes, (b) dwelling attributes, (c) household appliance-use related attributes and (d) climate related attributes. The considerable body of research on household level residential energy demand does not consider the impact of household residents' travel infrastructure and usage on energy consumption patterns. The absence of travel infrastructure and usage elements in energy demand models can be attributed to the lack of data providing this information in energy surveys. Transportation survey data on the other hand compiles this information. However, given these data are compiled on different respondents there has not been any consideration for using the datasets together. In this study, we proposed a novel framework that allows us to combine two datasets without any common identifier. The 2020 Residential Energy Consumption Survey (RECS) data and the 2022 National Household Travel Survey (NHTS) data are employed for this fusion exercise. The data fusion is performed based on the common attributes across the two datasets. Seven common variables – household Division (D), household Income (I), household Location (L), household Ownership (O), household Region (R), household Size (S) and housing Type (T) – are present in both RECS and NHTS dataset. Based on these common variables, in our study, 9 variable combinations are examined for data fusion: (1) D-I-L-O-R-S, (2) D-I-L-O-S-T, (3) D-I-L-O-R-T, (4) D-I-L-R-S-T, (5) D-I-R-S-T, (6) D-L-O-R-S, (7) D-L-O-R-S-T, (8) D-L-O-S-T and (9) I-L-O-R-S-T. The dependent variable of interest is energy consumption by end-use dimension (such as space heating, water heating, refrigeration, cooling and ventilation, EV charging, and lighting). Given the multiple discrete nature of the dependent variable, the Multiple Discrete Continuous Extreme Value (MDCEV) model is employed in this analysis. Based on the improvement of the model log-likelihood (relative to the model with only RECS dataset), the optimal variable combination and the optimal fusion size for the corresponding combination are selected. In this study, the optimal fusion model is estimated by fusing 20 NHTS records to each RECS record based on the combination D-L-O-R-S.

The model estimation results provide valuable insights into the important factors that affect residential sector energy end use demand. The traditional variables that contribute to increased energy use include household size, urban location, HDD and CDD. The travel infrastructure variables affecting energy usage include number of vehicles, number of hybrid vehicles and inter-city trips. The model results are augmented with an elasticity analysis to highlight the inconsistencies of energy models that ignore travel pattern variables. The elasticity analysis highlights the impact of various travel infrastructure and usage related variables on energy end-uses. It was observed that household energy consumption for cooling and ventilation increases with an increment in household annual mileage but decreases with an increased number of long-distance trips. The results also revealed that an additional vehicle in the household leads to higher energy demand for EV charging. With growing EV adoption, the link between travel infrastructure and usage and energy consumption is useful in understanding the evolving residential sector energy demand.

The insights from our study can also be used to design integrated energy management strategies that account for residential and transportation energy needs. For instance, EV sales are increasing in recent years and according to the US Department of Energy, 80% of EVs are charged at home (John, 2022). This study provides valuable insights into the factors that impact household EV charging infrastructure. By identifying the locations of higher concentration of households that consume energy for EV charging, the proposed framework can help in estimating energy demand across different spatial resolutions, redesigning the national grid and determining optimal locations for new EV charging infrastructure. Thus, agencies interested in predicting future energy demand can incorporate the proposed framework within their toolkits.

To be sure, the research is not without limitations. *First*, in our study, the dependent variable – in the RECS data – is annual energy consumption by different energy end uses. Hence, the analysis is limited to travel infrastructure and usage variables at the annual resolution. Future research efforts might explore the value of considering high resolution travel behavior elements such as number of daily trips, travel mode and trip purpose information in energy demand modeling. *Second*, in our research, the fusion exercise was conducted using the 2020 RECS data and 2022 NHTS data. Given the differences in the time period, it is possible that there are several differences in household energy and travel infrastructure and usage behavior over the time period. In future research, it might be useful to conduct data fusion using datasets from the same year. *Third*, the focus of the current study is developing improved models for annual energy consumption. If the interest is on understanding the impact of peak energy demand due to growing EV adoption and/or potential solutions for grid management, it would be more useful to develop a short-term energy demand model using more detailed energy data from smart meters (see for example (Beckel et al., 2014; Viegas et al., 2016; Yildiz et al., 2017)). The current model framework predictions can serve as an input for short term energy demand models. *Finally*, due to data unavailability, EV charging characteristics such as charging frequency and annual average daily charging hours were not considered in model estimation. It would be useful for future energy consumption surveys to include additional questions on household travel infrastructure variables. Further the survey effort can focus on daily household energy consumption patterns (such as frequency and duration of EV charging in a day, amount of electricity generated by solar panels and daily energy consumption for different household end uses) and travel behavior (such as number of trips on weekdays or weekends, trip purpose, number of people on trip and trip mode). These variables will allow future researchers to focus on building detailed short term energy demand models.

CRediT authorship contribution statement
Md Istiak Jahan: Data curation, Formal analysis, Methodology, Software, Visualization, Writing – original draft. **Naveen Eluru:** Conceptualization, Methodology, Supervision, Validation, Visualization, Writing – review & editing.

Declaration of interests
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data statement
The research employed 2020 Residential Energy Consumption Survey dataset and 2022 National Household Travel Survey dataset. Both datasets are publicly available.

References
Ahmed, A., Ge, T., Peng, J., Yan, W. C., Tee, B. T., & You, S. (2022). Assessment of the renewable energy generation towards net-zero energy buildings: A review. *Energy and Buildings*, *256*, 111755. https://doi.org/10.1016/j.enbuild.2021.111755

Beckel, C., Sadamori, L., Staake, T., & Santini, S. (2014). Revealing household characteristics from smart meter data. *Energy*, *78*, 397–410. https://doi.org/10.1016/j.energy.2014.10.025

Besagni, G., & Borgarello, M. (2018). The determinants of residential energy expenditure in Italy. *Energy*, *165*, 369–386. https://doi.org/10.1016/j.energy.2018.09.108

Bhat, C. R. (2005). A multiple discrete-continuous extreme value model: Formulation and application to discretionary time-use decisions. *Transportation Research Part B: Methodological*, *39*(8), 679–707. https://doi.org/10.1016/j.trb.2004.08.003

Bhat, C. R. (2008). The multiple discrete-continuous extreme value (MDCEV) model: Role of utility function parameters, identification considerations, and model extensions. *Transportation Research Part B: Methodological*, *42*(3), 274–303. https://doi.org/10.1016/j.trb.2007.06.002

Bhat, C. R., & Eluru, N. (2010). The Multiple Discrete-Continuous Extreme Value (MDCEV) Model: Formulation and Applications. In *Choice Modelling: The State-of-the-art and The State-of-practice* (pp. 71–99). Emerald Group Publishing Limited. https://doi.org/10.1108/9781849507738-004

Bhat, C. R., & Sen, S. (2006). Household vehicle type holdings and usage: An application of the multiple discrete-continuous extreme value (MDCEV) model. *Transportation Research Part B: Methodological*, *40*(1), 35–53. https://doi.org/10.1016/j.trb.2005.01.003

Bhowmik, T., Eluru, N., Hasan, S., Culotta, A., & Chandra, K. (2024). Predicting hurricane evacuation behavior synthesizing data from travel surveys and social media. *Transportation Research Part C*,

165(July), 104753. https://doi.org/10.1016/j.trc.2024.104753
Bhowmik, T., Iraganaboina, N. C., & Eluru, N. (2024). A Novel Maximum Likelihood Based Probabilistic Behavioral Data Fusion Algorithm for Modeling Residential Energy Consumption. *PLoS ONE*, *Under revi*.
Bhowmik, T., Yasmin, S., & Eluru, N. (2022). Accommodating for systematic and unobserved heterogeneity in panel data: Application to macro-level crash modeling. *Analytic Methods in Accident Research*, *33*, 100202. https://doi.org/10.1016/J.AMAR.2021.100202
Burnett, J. W., & Kiesling, L. L. (2022). How do machines predict energy use? Comparing machine learning approaches for modeling household energy demand in the United States. *Energy Research and Social Science*, *91*(June), 102715. https://doi.org/10.1016/j.erss.2022.102715
Chun-sheng, Z., Shu-wen, N., & Xin, Z. (2012). Effects of household energy consumption on environment and its influence factors in rural and urban areas. *Energy Procedia*, *14*, 805–811. https://doi.org/10.1016/j.egypro.2011.12.1015
D'Agostino, D., Minelli, F., Minichiello, F., & Russo, N. (2024). The impact of the new EPBD "Green Houses" in the Italian building context: technical-economic analysis on improving the energy class of existing buildings. *E3S Web of Conferences*, *523*, 1–14. https://doi.org/10.1051/e3sconf/202452304005
Dixit, S. K., & Singh, A. K. (2022). Predicting Electric Vehicle (EV) Buyers in India: A Machine Learning Approach. *The Review of Socionetwork Strategies*, *16*(2), 221–238. https://doi.org/10.1007/s12626-022-00109-9
EIA. (2024a). *Degree days*. U.S. Energy Information Administration. https://www.eia.gov/energyexplained/units-and-calculators/degree-days.php
EIA. (2024b). *How much energy is consumed in U.S. buildings?* U.S. Energy Information Administration. https://www.eia.gov/tools/faqs/faq.php?id=86&t=1
Ellsworth-Krebs, K. (2020). Implications of declining household sizes and expectations of home comfort for domestic energy demand. *Nature Energy*, *5*, 20–25. https://doi.org/https://doi.org/10.1038/s41560-019-0512-1
Fan, J. L., Yu, H., & Wei, Y. M. (2015). Residential energy-related carbon emissions in urban and rural China during 1996-2012: From the perspective of five end-use activities. *Energy and Buildings*, *96*, 201–209. https://doi.org/10.1016/j.enbuild.2015.03.026
Heinonen, J., & Junnila, S. (2014). Residential energy consumption patterns and the overall housing energy requirements of urban and rural households in Finland. *Energy and Buildings*, *76*, 295–303. https://doi.org/10.1016/j.enbuild.2014.02.079
Iraganaboina, N. C., & Eluru, N. (2021). An examination of factors affecting residential energy consumption using a multiple discrete continuous approach. *Energy and Buildings*, *240*, 110934. https://doi.org/10.1016/j.enbuild.2021.110934
Jahan, M. I., Bhowmik, T., Borjigin, S. G., Lou, J., Ugwu, N. M., Niemeier, D. A., & Eluru, N. (2024). A Maximum Log-likelihood Based Data Fusion Model for Estimating Household's Vehicle Purchase Decision. *Transportation Letters: The International Journal of Transportation Research*, *Under revi*.
John, J. St. (2022). 5 charts that shed new light on how people charge EVs at home. *Canary Media*.
Kabli, A., Bhowmik, T., & Eluru, N. (2020). A multivariate approach for modeling driver injury severity by body region. *Analytic Methods in Accident Research*, *28*, 100129. https://doi.org/10.1016/J.AMAR.2020.100129
Kaitouni, S. I., Sangkyu, P., Mghazli, M. O., El Mansouri, F., Jamil, A., Ahachad, M., & Brigui, J. (2024). Design parameters influencing the energy performance and indoor comfort of net zero energy building "NZEB" designed for semi-arid urban areas: Digital workflow methodology, sensitivity analysis and comparative assessment. *Solar Energy*, *268*(December 2023), 112264. https://doi.org/10.1016/j.solener.2023.112264
Kuang, B., Schelly, C., Ou, G., Sahraei-Ardakani, M., Tiwari, S., & Chen, J. (2023a). Data-driven analysis of influential factors on residential energy end-use in the US. *Journal of Building Engineering*, *75*(March). https://doi.org/10.1016/j.jobe.2023.106947

Kuang, B., Schelly, C., Ou, G., Sahraei-Ardakani, M., Tiwari, S., & Chen, J. (2023b). Data-driven analysis of influential factors on residential energy end-use in the US. *Journal of Building Engineering*, *75*(March), 1–22. https://doi.org/10.1016/j.jobe.2023.106947

Lee, S. J., & Song, S. Y. (2022). Determinants of residential end-use energy: Effects of buildings, sociodemographics, and household appliances. *Energy and Buildings*, *257*, 111782. https://doi.org/10.1016/j.enbuild.2021.111782

Lu, Y., Gao, W., Kuroki, S., & Ge, J. (2022). Household characteristics and electricity end-use under dynamic pricing in the collective housing complex of a Japanese smart community. *Journal of Asian Architecture and Building Engineering*, *21*(6), 2564–2579. https://doi.org/10.1080/13467581.2021.1987244

Malla, S. (2022). An outlook of end-use energy demand based on a clean energy and technology transformation of the household sector in Nepal. *Energy*, *238*, 121810. https://doi.org/10.1016/j.energy.2021.121810

Minelli, F., Ciriello, I., Minichiello, F., & D'Agostino, D. (2024). From Net Zero Energy Buildings to an energy sharing model - The role of NZEBs in Renewable Energy Communities. *Renewable Energy*, *223*(November 2023), 120110. https://doi.org/10.1016/j.renene.2024.120110

NHTS. (2022). *National Household Travel Survey*. Federal Highway Administration. https://doi.org/https://nhts.ornl.gov/

RECS. (2020). *Residential Energy Consumption Survey*. U.S. Energy Information Administration. https://doi.org/https://www.eia.gov/consumption/residential/

Ren, Z., Foliente, G., Chan, W. Y., Chen, D., Ambrose, M., & Paevere, P. (2013). A model for predicting household end-use energy consumption and greenhouse gas emissions in Australia. *International Journal of Sustainable Building Technology and Urban Development*, *4*(3), 210–228. https://doi.org/10.1080/2093761X.2013.801801

Shimoda, Y., Sugiyama, M., Nishimoto, R., & Momonoki, T. (2021). Evaluating decarbonization scenarios and energy management requirement for the residential sector in Japan through bottom-up simulations of energy end-use demand in 2050. *Applied Energy*, *303*, 117510. https://doi.org/10.1016/j.apenergy.2021.117510

Shin, H.-S., Farkas, Z. A., & Nickkar, A. (2019). An analysis of attributes of electric vehicle owners' travel and purchasing behavior: The case of Maryland. *International Conference on Transportation and Development*, 77–90.

USAFacts. (2024). *What are the main sources of US greenhouse gas emissions?* USA Facts. https://usafacts.org/articles/what-are-the-main-sources-of-us-greenhouse-gas-emissions/

Viegas, J. L., Vieira, S. M., Melício, R., Mendes, V. M. F., & Sousa, J. M. C. (2016). Classification of new electricity customers based on surveys and smart metering data. *Energy*, *107*(2016), 804–817. https://doi.org/10.1016/j.energy.2016.04.065

Wales, T. J., & Woodland, A. D. (1983). ESTIMATION OF CONSUMER DEMAND SYSTEMS WITH BINDING NON-NEGATIVITY CONSTRAINTS. *Journal of Econometrics*, *21*, 263–285.

Wang, L., Lee, E. W. M., Hussian, S. A., Yuen, A. C. Y., & Feng, W. (2021). Quantitative impact analysis of driving factors on annual residential building energy end-use combining machine learning and stochastic methods. *Applied Energy*, *299*(December 2020), 117303. https://doi.org/10.1016/j.apenergy.2021.117303

Xie, Y., & Noor, A. I. M. (2022). Factors Affecting Residential End-Use Energy: Multiple Regression Analysis Based on Buildings, Households, Lifestyles, and Equipment. *Buildings*, *12*(538), 1–18.

Yildiz, B., Bilbao, J. I., Dore, J., & Sproul, A. B. (2017). Recent advances in the analysis of residential electricity consumption and applications of smart meter data. *Applied Energy*, *208*(October), 402–427. https://doi.org/10.1016/j.apenergy.2017.10.014

Yu, B., & Zhang, J. (2015). Modeling household energy consumption behavior: A comparative analysis. *Transportation Research Part D: Transport and Environment*, *39*, 126–140. https://doi.org/10.1016/j.trd.2015.07.001

Yu, B., Zhang, J., & Fujiwara, A. (2013). Integrated Analysis of Household Energy Consumption Behavior

Across the Residential and Transport Sectors. In A. Fujiwara & J. Zhang (Eds.), *Sustainable Transport Studies in Asia* (pp. 211–236). Springer.

Zou, B., & Luo, B. (2019). Rural household energy consumption characteristics and determinants in China. *Energy*, *182*, 814–823. https://doi.org/10.1016/j.energy.2019.06.048