KIEAE Journal, Vol. 26, No. 4, Aug. 2026, pp.5-12.

Check for updates

KIEAE Journal

Korea Institute of Ecological Architecture and Environment

140

A Comparative Analysis of Influencing Factors in the Residential Energy Consumption Survey 2009, 2015, and 2020

Jonghoon Ahn*

* *Associate Professor, School of Architecture and Design Convergence, Hankyong National Univ., South Korea (architectism@hknu.ac.kr)*

ABSTRACT

Purpose: Despite being a high-quality dataset that has evolved through continuous methodological improvements since the 1970s to capture the diverse aspects of U.S. residential buildings, the Residential Energy Consumption Survey remains underutilized in effective analysis. The primary objective of this study is to evaluate how geographic, climatic, physical, demographic, and infrastructural factors influence total annual household energy and energy use intensity. **Method:** The framework integrates thermodynamic principles, statistical multivariate linear regression, and data mining techniques to model linear behaviors, and structural interaction effects. **Result:** Degree days, and total conditioned floor area consistently emerge as the primary determinants governing annual residential energy consumption across all three surveys. Although newer homes exhibit improved insulation and building envelope standards, spatial expansion and increased appliance plug loads frequently offset per-unit efficiency gains. Analysis of the 2015 dataset highlights a paradigm shift from purely statistical models toward physics-based engineering approaches alongside a overall decline in average household consumption. The 2020 dataset underscores teleworking as an influential behavioral variable, where each additional remote work day per week increased annual household consumption. Additionally, a statistical analysis on the 2020 confirms that transitioning space and water heating from fossil fuels to electric heat pump systems serves as the single most critical leverage point for reducing site energy demand. In conclusion, residential energy consumption is governed by a structural hierarchy dominated by climate and spatial scale, while post-pandemic lifestyle shifts and heating electrification are reshaping modern load dynamics.

KEYWORD

Residential Building
Energy Use
RECS
Influencing Factor
Statistical Analysis

ACCEPTANCE INFO

Received Aug. 2, 2026
Final revision received Aug. 7, 2026
Accepted Aug. 13, 2026

1. Introduction

1.1. Research Purpose

The building sector accounts for a major share of the nation's primary energy consumption and greenhouse gas emissions in USA. Understanding and predicting the energy consumption patterns of residential buildings within this sector is the most essential prerequisite for improving energy efficiency, ensuring grid stability, and establishing climate change response policies. The 2009, 2015, and 2020 Residential Energy Consumption Survey (RECS) data, collected under the auspices of the U.S. Energy Information Administration (USEIA), is one of the most comprehensive and reliable datasets available for identifying the micro- and macro-level drivers of residential energy demand [1,2]. Since its first implementation in 1978, the RECS applies the Complex Multistage, area-probability sample design technique to provide precise microdata for over 12,000 households that statistically represent over 110 million residential households across the United States.

The primary purpose of this report is to examine in depth how the geographic, climatic, physical, demographic, and infrastructural independent variables included in the provided RECS 2009 microdata affect the dependent variable, total annual household energy use, in terms of the extent and intensity of their influence [1,2].

1.2. Research Framework

This study combines thermodynamic principles, statistical multiple linear regression analysis (MLR), and behavioral economics perspectives to thoroughly examine not only the direct effects of individual variables but also the multi-dimensional interactions among variables. Beyond four-dimensional, fragmented correlation analyses, it develops and derives the second- and third-order ripple effects including the rebound effect, where the optimization of the building environment does not directly lead to reduced energy consumption through an integrated narrative prose approach that separates base load and variable load in energy consumption.

pISSN 2288-968X, eISSN 2288-9698
http://dx.doi.org/10.12813/kieae.2026.26.4.005

2. Literature Review

2.1. Energy Use Intensity

Energy Use Intensity (EUI) is a primary quantitative benchmark used in building science and energy policy to measure a building's operational energy efficiency. It expresses a building's annual total energy consumption relative to its total gross floor area, providing a normalized, size-independent metric for comparing energy performance across heterogeneous structures. Mathematically, EUI is calculated as the ratio of total annual energy consumption to the gross floor area [3,4]:

$$EUI = \frac{E_{annual}}{A_{groos}} \quad \text{(Eq. 1)}$$

where: E_{annual} is the sum of all energy consumed on-site over a continuous 12-month period (e.g., electricity, natural gas, fuel oil, district thermal energy), A_{gross} is the total gross conditioned (or indoor) floor area of the building.

Thermal Units per square foot per year commonly used in U.S. building codes, ENERGY STAR Portfolio Manager, CBECS and RECS datasets, and SI System is a standard across European, Asian, and international research frameworks. Simply, they can be converted by use of the Conversion factor: 1kBtu/sf= 3.1546kWh/m^2=11.356MJ/m^2.

In academic literature and policy applications, a crucial distinction is made between Site EUI and Source EUI. The Site EUI Measures the energy consumed at the building location (the utility meter level). And it focuses Evaluates operational efficiency, the Heating, Ventilation, and Air Conditioning (HVAC) performance, occupant behavior, and thermal envelope integrity, but does not account for off-site generation inefficiencies or transmission losses. The source EUI ncorporates total primary energy demand, factoring in off-site extraction, generation, and distribution losses required to deliver energy to the site. And it focuses Evaluation of the comprehensive environmental and raw resource impact of the building. It expressed by multiplying site energy end-uses by national or regional primary energy conversion factors.

By dividing by floor area, the EUI allows direct comparison between buildings of vastly different footprints and it serves as a standard continuous target variable in data-driven modeling to eliminate geometric scale bias. It forms the baseline for green building certification standards, urban energy benchmarking ordinances, and net-zero energy building verification.

2.2. RECS Data

To analyze the highly heterogeneous and complex phenomenon of residential energy consumption, a clear understanding of the statistical structure under which the data were collected is essential. Rather than relying on Simple Random Sampling (SRS), the RECS dataset employs a complex survey design involving stratification and clustering [2~5]. While each row in this dataset represents an individual household with a unique identifier (DOEID), the final sample weights (NWEIGHT) must be applied to ensure these households accurately represent the target population, all U.S. households [2~5]. Furthermore, calculating the variance and relative standard errors (RSEs) of statistical estimates requires sophisticated methodological techniques. Applying conventional Ordinary Least Squares (OLS) variance estimation directly ignores the intra-cluster correlation induced by cluster sampling, thereby posing a significant risk of Type I error through overstated statistical significance (p-values) [2~6]. To mitigate this bias, analysts utilize replicate weight methods, such as the Jackknife (JK1) or Fay's Balanced Repeated Replication (BRR) technique [2~6]. By iteratively fitting the model across 60 to 96 replicate weights, researchers can rigorously establish confidence intervals for the effects of explanatory variables on annual total household energy consumption (TOTALBTU) [2~6]. This statistical rigor serves as a critical safeguard to eliminate structural sampling bias when estimating regression coefficients for key covariates, such as climate, square footage, and housing unit type.

The dependent variable TOTALBTU, which serves as the focal outcome measure across all analyses, represents an absolute energy consumption index that aggregates all primary and secondary energy sources consumed by a household over a one-year period, converted into British Thermal Units (BTU). Energy is consumed across diverse physical units and formats, including electricity (kWh), natural gas (USENG, measured in CCF or therms), liquefied petroleum gas (LPG/propane), fuel oil, and wood. Integrating these disparate sources into a single continuous dependent variable for regression analysis requires a rigorous thermodynamic unit-conversion mechanism [7~9].

In the case of electricity consumption, 1 kilowatt-hour (kWh) corresponds thermodynamically to approximately 3,412.14 BTU. Within the dataset, electricity is disaggregated into specific end-use estimates such as space heating (KWHSPH), space cooling (KWHCOL), water heating (KWHWTH), refrigeration (KWHRFG), and other appliances (KWHOTH) before being aggregated into total BTU [7~9]. Similarly, volume-based fuels like natural gas are converted into BTU by applying national standards or regionally weighted conversion factors, such as 1.003 therms per CCF [7~9].

The TOTALBTU metric denotes "site energy" that is, the energy directly consumed at the residential boundary, excluding the thermodynamic generation, transmission, and distribution losses incurred off-site [10]. Consequently, households relying primarily on electricity exhibit an inherent structural tendency toward lower TOTALBTU values relative to those utilizing on-site combustion fuels.

3. Research Method

The Pearson Correlation Coefficient, foundational to modern statistical analysis, is a parametric measure utilized to quantify the strength, direction, and probability of the linear association between two continuous random variables [11,12]. Mathematically, the coefficient is defined as the covariance of two variables divided by the product of their respective standard deviations [11,12]. The computed coefficient, denoted as γ, is strictly bounded within a dimensionless range of −1.0 to +1.0. A value of +1.0 indicates a perfect positive linear trajectory, −1.0 denotes a perfect negative linear trajectory, and a value approaching 0 implies the complete absence of any linear relationship [13]. In the context of building energy prediction and multivariate modeling, conducting a Pearson correlation analysis serves as a mandatory and rigorous preliminary step before advancing to complex regression algorithms. Deploying correlation analysis allows researchers to empirically execute 'feature selection' systematically identifying and retaining only those independent architectural or climatic variables that exhibit a statistically significant linear driving force on the dependent target variable. Furthermore, a comprehensive correlation matrix is instrumental in diagnosing severe multicollinearity among the independent predictors themselves [14,15]. By filtering out statistically insignificant noise and isolating redundant predictors, Pearson correlation analysis ensures the mathematical stability, robust variance estimation, and ultimate interpretability of the subsequent MLR models [14,15]. Mathematically, the sample Pearson correlation coefficient (γ) between two continuous variables, X and Y, is calculated by dividing their covariance by the product of their standard deviations. The fundamental formula is expressed as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad \text{(Eq 2)}$$

where: "n" is the total number of observations, "x_i" and "y_i" are the individual sample values of variables X and Y, respectively, "$\bar{x}$" is the sample mean of variable X, "$\bar{y}$" is the sample mean of variable Y.

These models evaluate building energy performance by isolating primary drivers and estimating their corresponding weights. Establishing their predictive accuracy and reliability requires rigorous statistical validation. While standard deviation describes data spread, Ordinary Least Squares (OLS) regression and Analysis of Variance (ANOVA) isolate specific parameter relationships. Namely, the Root Mean Squared Error (RMSE) evaluates residual variance, whereas the coefficient of determination (R^2 or adjusted R^2) measures the variance explained by the model's predictors [16,17].

$$\begin{aligned} &\textit{R-squared } (R^2) \\ &\quad \equiv 1-(\textit{The sum of squared of residuals})/ \\ &\quad\quad (\textit{The total sum of squares}) \end{aligned} \qquad \text{(Eq 3)}$$

$$RMSE = \textit{squared root}\ ((\sum_{t=1}^{n}(\hat{y_t} - y_t)^2)/n \qquad \text{(Eq 4)}$$

where "n" is the sample size, "$\hat{y}_t$" represents the predicted values, "y_t" is the observed dependent variable, and bar is its mean.

In order to allow fair comparison across different scales, the Coefficient of Variation of the Root Mean Square Error (CvRMSE) normalizes the error against the empirical mean [18~20].

$$\textit{CV of RMSE} = RMSE/\hat{y} \qquad \text{(Eq 5)}$$

4. Results and Discussion

4.1. Comprehensive Analysis of Factors Influencing the Annual Total Household Energy Consumption

The statistical results derived from the Pearson Correlation Analysis to identify factors highly influencing the TOTALBTU are indicated in Table 1.

Residential energy consumption is fundamentally the result of a physical defense mechanism aimed at maintaining indoor thermal comfort against thermodynamic loads imposed by an uncontrollable external environment. Consequently, climate and geographical characteristics are exogenous variables that account for the largest share of variance in TOTALBTU. The HDD65 and CDD65 variables included in the survey data represent heating and cooling degree days, respectively. They serve as absolute metrics of climate load, calculated as a weighted average from

Table 1. Rank of the statistical coefficient for variables

Rank	Variable Name (Code)	Coefficient (γ)
1	Number of Stories (STORIES)	0.568
2	Total Conditioned Area (TOTSQFT_EN)	0.529
3	Number of Windows (WINDOWS)	0.508
4	Total Rooms (TOTROOMS)	0.365
5	Bedrooms (BEDROOMS)	0.351
6	Heating Degree Days (HDD65)	0.338
7	Bathrooms (NCOMBATH)	0.209
8	Type of Housing Unit (TYPEHUQ)	-0.286
9	Cooling Degree Days (CDD65)	-0.283
10	Insulation Adequacy (ADQINSUL)	-0.058
11	Main Heating Equipment (EQUIPM)	-0.054

nearby weather stations based on a standard baseline temperature of 65˚F (approximately 18.3℃). Considering internal heat gains (from human bodies, appliances, etc.) and solar radiation, 65˚F represents the theoretical thermal balance point where no additional heating or cooling is required. If the daily mean temperature falls below 65˚F, the difference accumulates into HDD65; if it exceeds 65˚F, it accumulates into CDD65. According to machine learning and multiple regression analysis studies evaluating variable importance (VI), HDD65 and CDD65 exhibit explanatory powers of 20.6% and 6.2%, respectively, in predicting total energy consumption, exerting an overwhelmingly greater influence than any other variable. Comparing specific households within the dataset clearly demonstrates the impact of these climate variables. For example, a household located in a severely cold climate with an HDD65 of 6,231 (DOEID 3552) records a TOTALBTU reaching 1,096,083, whereas a household in a mild region with an HDD65 of only 834 (DOEID 8172) stays at a TOTALBTU level of 9,110. Because heating loads involve a significantly larger indoor-outdoor temperature difference compared to cooling loads, operating heating equipment accounts for the largest portion of TOTALBTU across most U.S. climate zones. Consequently, households in northern regions with high HDD65 values inevitably record a higher baseline energy consumption compared to those in southern regions. Alongside HDD65 and CDD65 which represent actual weather conditions for a single year (2009) the dataset includes the HDD30YR and CDD30YR variables, reflecting the 30-year average climate from 1981 to 2010. Rather than dictating immediate operational rates in a given year, these long-term climate indicators act as the "structural decision-making background" upon which builders and homeowners previously made choices regarding building design, insulation thickness (ADQINSUL), and HVAC equipment capacity (EQUIPM). Furthermore, deviations between long-term and short-term climate indicators can serve as metrics demonstrating how inefficiently a housing unit responds when extreme weather events driven by climate change occur.

The REGIONC (four Census regions) and DIVISION (nine division sub-regions) variables serve as proxies for multi-layered infrastructural differences that extend beyond mere climate variations, such as the stringency of building energy codes adopted by state governments, local electricity generation costs, and natural gas pipeline penetration rates. On a more granular level, the METROMICRO (Metropolitan/Micropolitan distinction) and UATYP10 (Urban Area Type) variables capture differences in housing typography driven by the degree of urbanization. Metropolitan areas (METRO), characterized by high land prices and population densities, feature a higher concentration of multi-family housing and apartments. Compared to non-metropolitan areas (MICRO/RURAL) dominated by single-family homes, these multi-family units feature smaller average floor areas per household and a higher proportion of shared exterior walls, structurally lowering the TOTALBTU per individual household. In regression models, the interaction terms between urbanization and regional variables exert a statistically significant effect on total energy consumption (p-value〈0.0001), supporting the strong correlation between spatial density and energy efficiency.

If climate factors determine the intensity of the thermal load imposed on a housing unit, a building's physical area and geometric structure dictate its capacity and surface area for accepting that load—thereby either amplifying or mitigating TOTALBTU. Among the various area metrics recorded in the dataset, TOTSQFT_EN is the most direct and powerful internal determinant of energy consumption. This variable does not merely represent a building's gross square footage; rather, it sums the total space where energy is actively consumed—encompassing primary living quarters (such as living rooms and bedrooms) as well as basements, finished attics, and heated/cooled garages that receive heating or cooling. Thermodynamically, as a building's volume increases, the volume of air that must be conditioned to reach a target temperature grows linearly. Consequently, TOTSQFT_EN exhibits a highly significant positive correlation with TOTALBTU, possessing the highest explanatory power in regression models after climate variables. The TOTHSQFT (heated area) and TOTCSQFT (cooled area) variables in the dataset break down the specific zones where residents actually operate heating or cooling within the total area, serving as key parameters that maximize load prediction precision in detailed energy modeling, such as for KWHSPH (space heating electricity) or KWHCOL (air conditioning electricity).

The type of housing unit (TYPEHUQ) variable classifies housing types into five categories. The influence of this variable on TOTALBTU can be explained by the physics of the surface-to-volume ratio, which governs heat loss in buildings. Even for housing units with TOTSQFT_EN, a single-family detached home has its ceiling, floor, and all four exterior walls exposed to cold or hot ambient air, maximizing energy loss via thermal transmittance. Conversely, an apartment in a building with 5 or more units shares its ceiling, floor, and adjacent walls with neighboring units, drastically reducing the thermal exchange surface area. Research demonstrates that single-family detached homes consume 54% more energy for space heating and 26% more energy for cooling compared to multi-family units. Similarly, the STORIES variable, which indicates a building's vertical footprint, can be interpreted in a similar context. Higher-story buildings exhibit a lower roof-to-floor area ratio, which reduces heat loss through the roof and alters internal air circulation patterns via the stack effect, exerting a complex influence on space heating load characteristics.

Within the physical envelope of a building lies the dynamic factor of occupant behavior. In the RECS dataset, variables such as TOTROOMS (total number of rooms), BEDROOMS (number of bedrooms), and NCOMBATH (number of full bathrooms) go beyond simple architectural partitioning; they serve as critical proxy variables reflecting household size and socioeconomic status. A higher count of BEDROOMS and NCOMBATH statistically correlates with a greater number of occupants residing in the housing unit. A larger household inevitably leads to an increased frequency of showers and baths, directly driving sharp rises in domestic hot water energy consumption, such as electricity for water heating (KWHWTH) or natural gas for water heating (UGWATER). In variable importance analyses using machine learning models, household size (number of household members) was identified as the primary governing factor, accounting for 66.45% of the variance in water heating energy consumption. Furthermore, multi-person households increase the operational frequency of appliances such as washing machines, clothes dryers, and dishwashers. They also maintain a higher inventory of personal electronic devices (e.g., televisions and computers), driving a substantial increase in baseline electrical loads such as ELFOOD (cooking electricity) and ELOTHER (miscellaneous appliance electricity). Prior research applying structural equation modeling (SEM) demonstrated that the indirect effect of household demographic characteristics on total energy consumption, mediated through housing scale and room counts, is more than four times greater than its direct effect. Consequently, while variables like TOTROOMS exhibit strong multicollinearity with TOTSQFT_EN, they independently capture the intensity of household living patterns, contributing as statistically significant explanatory variables in the TOTALBTU regression model.

The building envelope serves as the primary barrier that confines thermal energy indoors against extreme external climates. The RECS dataset contains detailed indicators for evaluating a building's thermal resistance, including construction year (YEARMADERANGE), exterior wall material (WALLTYPE), roof material (ROOFTYPE), window characteristics (WINDOWS), and insulation adequacy (ADQINSUL). YEARMADERANGE variable reflects not only the physical aging of a building, but also the stringency of the building energy codes enforced at the time of its design. Following the global oil shocks of the 1970s, energy regulations in the United States were progressively tightened. According to RECS data, only 52% of homes built before 1990 feature high-efficiency multi-pane windows, whereas this proportion jumps to nearly 80% for housing built after 2000. Reductions in window thermal transmittance (U-value) and improvements in wall insulation performance have dramatically improved EUI. A granular analysis of statistical regression models, however, reveals a frequent empirical counter-trend: newer homes often exhibit higher TOTALBTU levels than older ones. While seemingly paradoxical, this phenomenon is well-explained by the rebound effect. Although newer housing units boast significantly higher efficiency per unit area, their overall floor area (TOTSQFT_EN) has expanded considerably compared to older stock. Furthermore, modern homes come standard with high-capacity central air conditioning systems and house multiple televisions alongside large-capacity household appliances. Consequently, energy savings achieved through superior building envelope performance (high ADQINSUL quality) are fully offset by absolute spatial expansion and increased appliance plug loads (APKWH). Because of this offsetting interaction, YEARMADERANGE achieves its true statistical significance in regression models when incorporated as an interaction term alongside floor area variables rather than as a standalone metric.

The primary heating fuel type (FUELHEAT) and the specific equipment used to convert that fuel into heat (EQUIPM) serve as direct mechanical variables determining TOTALBTU to meet thermal setpoints and defend against external building loads. Looking at the primary heating fuel (FUELHEAT) distribution across U.S. households, natural gas (USENG) is the most dominant at approximately 49%, followed by electricity at 34% and fuel oil at 6%. This fuel choice is not merely a matter of consumer preference; it dictates baseline physical efficiency. Boiler and furnace systems that burn natural gas or fuel oil rely on site-combustion methods, meaning that even with high Annual Fuel Utilization Efficiency (AFUE) ratings, their operational efficiency typically ranges between 70% and 95%. In other words, generating 100 units of heat requires inputting 110 to 140 BTUs worth of gas or fuel oil.

Conversely, modern electric heat pump technology operates by transferring heat from ambient air or geothermal sources rather than generating it directly through combustion, achieving a Coefficient of Performance (COP) between 200% and 300%. Consequently, households specifying electric heat pumps under their equipment variable (EQUIPM) hold a statistical advantage in yielding lower final TOTALBTU values compared to those relying on conventional boilers. This mechanical disparity explains why HVAC equipment type ranks among the most impactful explanatory variables in regression analyses, second only to floor area metrics. The ELWARM, ELCOOL, ELWATER, and ELFOOD variables in the survey dataset function as binary flags (1 or 0) indicating whether a specific household utilizes electricity for a given end use. These flags make it possible to identify how energy is distributed within a residence. For instance, space heating (ELWARM, UGWARM) and space cooling (ELCOOL) represent variable loads that fluctuate directly with outdoor temperature. In contrast, water heating (ELWATER, UGWATER), cooking (ELFOOD), and miscellaneous appliances (ELOTHER) form the baseline structural loads that remain steady throughout all four seasons. In modern residential environments—where household sizes vary and home electronics have reached saturation—the combined share of these active baseline load flags within total TOTALBTU continues to increase steadily.

The effects of the individual variables analyzed thus far do not occur in isolation in the real world. To synthesize these factors for predicting TOTALBTU and inferring causal mechanisms, researchers adopt sophisticated Multiple Linear Regression (MLR) and statistical cross-validation frameworks. Residential energy consumption data (TOTALBTU) inherently exhibits a right-skewed distribution. This skewness stems from extreme outliers, such as massive estates spanning tens of thousands of square feet or deeply impoverished housing with severely degraded insulation. To satisfy the linear regression assumptions of normality and homoscedasticity, taking the natural logarithm of the dependent variable, yielding (ln (TOTALBTU)), is the standard methodological approach. Following this transformation, each regression coefficient converts from a simple absolute BTU increment into a direct measure of elasticity, representing the percentage (%) change in total energy consumption per unit change in the independent variable. An optimized regression model incorporates cross-interaction effects, significantly improving both explanatory power (R^2) and information criteria:

1) HDD65 * TOTSQFT_EN: For larger housing units, as outdoor winter temperatures drop and HDD rise, energy consumption increases at a noticeably steeper slope (acceleration rate).
2) MAJORITY * HHINC2: Demographic factors, such as majority-white neighborhood status (MAJORITY), interact with household income (HHINC2) to shape housing choices. Because high-income households possess the financial capital to install top-tier energy-efficient systems alongside larger living spaces, the relationship exhibits a non-linear knee point where income growth does not translate infinitely into BTU increases.
3) REGIONC * UATYP10: Dense urban apartments in the Northeast and sprawling single-family suburban homes in the West exhibit vastly different thermodynamic footprints even at identical square footages; combining geographic variables thus yields high statistical significance. In deriving these models, a large absolute value for a regression coefficient does not automatically imply greater feature importance, as individual variables operate on vastly different scales.

4.2. Key Characteristics and Conclusions of the RECS 2015

Multiple regression analyses and empirical studies utilizing the RECS 2015 dataset reveal several core methodological and statistical shifts compared to 2009 data:

1) Pivot from Statistical Modeling to Engineering Approaches: While RECS analyses from 1980 through 2009 relied heavily on statistical regression models to estimate end-use energy consumption, the 2015 dataset marked a major methodological transition toward physics-based engineering modeling informed by collected building and climate data. This represents a fundamental paradigm shift toward building physics in energy estimation.
2) Continued Dominance of Heating Systems and Floor Area: Multiple regression models controlling for climate zones confirm that HVAC equipment type and total floor area (TOTSQFT_EN) remain the primary determinants of total energy consumption (TOTALBTU). Conversely, socioeconomic variables (such as income) and building age, which held substantial weight in 2009 models, exhibited diminished contribution margins. Factors such as education level and water heater type showed no statistically significant effect. However, the presence of solar systems demonstrated a statistically significant positive effect in select regions.
3) Reduction and Stabilization of Average Consumption: Average household energy consumption declined markedly

from approximately 89.6 MMBtu (89.6 million BTU) in 2009, establishing a downward, stabilized trend by 2015.

4) Overall Conclusion: The 2015 data analysis demonstrates that the core drivers of residential energy consumption have increasingly concentrated around physical baseline conditions: climate, floor area, and primary HVAC equipment. The indirect influence of occupant demographics and socioeconomic backgrounds has relatively waned, driven by rising baseline appliance efficiencies and stricter building codes.

4.3. Key Characteristics and Conclusions of the RECS 2020

The RECS 2020 dataset reflects unique statistical dynamics shaped by an unprecedented global pandemic and accelerating energy transition trends:

1) Emergence of Teleworking as a Behavioral Variable: The rapid expansion of remote work during the COVID-19 pandemic induced structural shifts in residential energy loads. Multilevel modeling reveals that among low-income and top-income brackets, each additional day of teleworking per week increased annual household energy consumption by approximately 871 thousand BTUs. Statistically, prolonged indoor occupancy acts as a powerful driver, elevating both space conditioning and baseline electrical loads simultaneously.
2) Elucidating Electrification Effects via Machine Learning: Applying CatBoost algorithms and SHAP sensitivity analyses to 2020 data confirms that heating degree days (HDD65), total floor area, and the adoption of electricity for space and water heating remain the top national determinants of total consumption. Crucially, utilizing electricity over traditional fossil fuels for space heating emerged as the single most critical leverage point for dramatically lowering site energy consumption, strongly supporting the systemic benefits of heat pump adoption.
3) Parity Between Electricity and Natural Gas Shares: At the national level, household energy consumption shares reached near-parity between electricity (47%) and natural gas (45%). Average household consumption held steady at 76.8 MMBtu, solidifying a sustained reduction relative to 2009 levels.
4) Overall Conclusion: Analysis of the 2020 dataset indicates that while building thermal properties and equipment efficiency have reached a mature baseline, evolving lifestyle patterns, such as remote work, introduce new dynamic variances into residential energy loads. Crucially, data-driven machine learning models demonstrate that electrification (transitioning from fossil fuels to electric systems) stands as the single most potent technical and policy leverage point for achieving absolute energy demand reductions.

5. Conclusion

Cross-analyzing the statistical and time-series determinants of annual total household energy consumption using the U.S. Energy Information Administration's 2009, 2015, and 2020 RECS datasets reveals that residential energy demand is governed by a multi-layered hierarchy of factors whose influence evolves over time.

1) Environmental Compulsion: Climate variables, epitomized by heating and cooling degree days, represent the top-tier, uncontrollable factor governing the single largest share of variance in total energy consumption. Climate dictates the absolute magnitude of the thermal load imposed on a home.
2) Physical and Geometric Amplification Factors: Heated and cooled floor area and housing type consistently form the second core axis of dominance across 2009, 2015, and 2020. Single-family detached structures and larger floor areas serve as primary drivers of thermal loss and load expansion.
3) Equipment and Structural Efficiency Factors: The physical efficiency of space conditioning equipment, such as gas boilers versus electric heat pumps, creates a decisive variance in final consumed annual total household energy consumption. The 2020 analysis particularly demonstrates that transitioning to electric-based systems exerts a dramatic downward push on overall energy consumption.
4) Demographic and Behavioral Factors: While room and bedroom counts historically served as proxies reflecting household size and appliance usage density to drive up baseline loads, recent data highlights emerging lifestyle patterns as key direct sources of consumption variance.
5) Evolution of Determinants Over Time: As confirmed by the 2015 and 2020 datasets, while building envelope performance and thermal system efficiencies have steadily upgraded across the board, absolute climate and spatial drivers continue to dominate. Simultaneously, new lifestyle variables like remote work are introducing fresh dynamics into energy consumption, and the electrification of heating sources has emerged as the most distinct driver for absolute site-energy reduction.

Individual household energy consumption patterns are not merely a function of how modern an appliance is; they are the output of a complex structural equation tightly combining climate, spatial scale, building geometry, and socioeconomic stratification. To achieve true nationwide annual total household energy consumption reductions and carbon neutrality, policymakers must guard against continuous spatial expansion offsetting efficiency gains, actively accelerate electrification, and implement flexible micro-level policies tailored to evolving occupant behaviors.

However, a major limitation of this study is the lack of a clear logical justification for removing outliers during data preprocessing and relying on off-the-shelf applications for missing value imputation and outlier removal. Consequently, the statistical validity of the benchmark model remains unverified. In order to address these shortcomings, a follow-up study will refine the preprocessing framework to establish a logically sound, highly valid multivariate regression model for predicting the energy use intensity of both expanded and newly constructed residential buildings.

References

[1] USEIA, Residential energy consumption survey: What's new in our home energy use?, 2011, https://www.eia.gov/consumption/residential/reports/2009/overview.php, 2026.05.08.

[2] USDOE, Residential energy consumption survey: 2009 RECS survey data, 2013, https://www.eia.gov/consumption/residential/data/2009/index.php?view=microdata, 2026.05.08.

[3] DATA.GOV, Building energy data book: Building databook, 2026, https://catalog.data.gov/dataset/buildings-energy-data-book, 2026.07.01.

[4] J. Ahn, Development of energy performance metrics for airport terminal buildings using multivariate regression modeling (Doctoral dissertation), Raleigh, NC: North Carolina State University, 2016.

[5] NYSERDA, Market characterization methodology report - Residential energy consumption survey, https://www.nyserda.ny.gov/-/media/Files/Publications/PPSER/Program-Evaluation/2017ContractorReports/LMI-Methodology-Rpt---Res-Energy-Consumption-Survey.pdf, 2026.07.04.

[6] S. Zimmer et al., Exploring complex survey data analysis using R: A tidy introduction with {srvyr} and {survey}, https://dokumen.pub/exploring-complex-survey-data-analysis-using-r-a-tidy-introduction-with-srvyr-and-survey-9781032302867-9781032306353-9781003305996.html, 2026.07.04.

[7] USEIA, Codebook, https://www.eia.gov/consumption/residential/data/2015/xls/codebook_publicv4.xlsx, 2026.05.04.

[8] USEIA, RECS Data, https://www.eia.gov/consumption/residential/data/2009/csv/recs2009_public.csv, 2026.05.04.

[9] USEIA, Residential energy consumption survey (RECS) 2015 consumption and expenditures technical documentation summary, https://www.eia.gov/consumption/residential/reports/2015/methodology/pdf/2015C&EMethodology.pdf, 2026.05.08.

[10] C. Han et al., Empirical assessment of the relationship between teleworking frequency and residential energy use in the U.S.: Multilevel modeling and time use comparisons, Environmental Research: Infrastructure and Sustainability, 5(2), 2025, 025018.

[11] J. Cohen et al., Applied multiple regression/correlation analysis for the behavioral sciences, 3rd ed., Mahwah, NJ: Lawrence Erlbaum Associates, 2003.

[12] J. Rodgers, W. Nicewander, Thirteen ways to look at the correlation coefficient, The American Statistician, 42(1), 1988, pp.59-66.

[13] J. Benesty et al., Pearson correlation coefficient, In: Noise reduction in speech processing, Berlin, Heidelberg: Springer, 2009, pp.1-4.

[14] D. Montgomery, E. Peck, G. Vining, Introduction to linear regression analysis, 5th ed., Hoboken, NJ: John Wiley & Sons, 2012.

[15] I. Jolliffe, Principal component analysis, 2nd ed., New York, NY: Springer, 2002.

[16] J. Ahn, S. Cho, D. Chung, Development of a statistical analysis model to benchmark the energy use intensity of subway stations, Applied Energy, 179, 2016, pp.488-496, https://doi.org/10.1016/j.apenergy.2016.06.065.

[17] J. Ahn, A benchmark methodology to assess the energy performance of train station complexes, KIEAE Journal, 19(6), 2019, pp.5-12, http://dx.doi.org/10.12813/kieae.2019.19.6.005.

[18] C. Kim, H. Shin, J. Ahn, Energy performance analysis of airport terminal buildings by use of architectural, operational information and benchmark metrics, Journal of Air Transport Management, 83, 2020, https://doi.org/10.1016/j.jairtraman.2020.101762.

[19] S. Cho et al., Methodology for energy strategy to prescreen the feasibility of ground source heat pump systems in residential and commercial buildings in the United States, Energy Strategy Reviews, 18, 2017, pp.53-62, https://doi.org/10.1016/j.esr.2017.09.012.

[20] G. Casella, G. Berger, Statistical inference, Pacific Grove, CA: Duxbury, 2002.