Deriving high-fidelity residential building archetypes and typical usage patterns from national energy use surveys to enhance “initial guesses” for Urban Building Energy Model (UBEM) inputs

Chengxuan Li[1,*], Timur Dogan[1]
[1]Environmental Systems Lab, Cornell University, Ithaca NY, United States.
*Corresponding Author: cl2749@cornell.edu

Abstract

Urban Building Energy Models (UBEMs) play a crucial role in improving energy efficiency and supporting decarbonization efforts across regions. However, scalability remains a challenge due to the lack of building-level information on construction, system properties, and usage patterns. This study presents a novel method for deriving detailed residential building archetypes from open urban datasets and the national Residential Consumption Survey (RECS), introducing a modular framework for archetype classification, assignment, and characterization. As a case study for residential buildings in climate zone 6A in the US, this study defines 41 sub-archetypes in total, enabling the creation of 270 unique sub-archetype combinations. Deployed in a 5R1C model, the archetypes effectively improve the “initial guesses” for building energy model inputs and enhance modeling accuracy and applicability. Potentially generalizable to all climate zones and building typologies in the US, the proposed workflow for UBEM archetype generation supports decision-making in building retrofit planning and accelerates building stock decarbonization.

Key Innovations

- Decoupling building archetypes into construction, systems, and energy loads sub-modules allows for versatility and better represents the diversity in the building stock.
- Usage of building archetypes enriches open urban data with stock level knowledge and improves “initial guesses” for UBEM inputs.
- Assigning feasible ranges to building archetype variables during characterization balances model simplicity with real-world variability.
- The proposed archetype workflow better synthesizes RECS dataset and is scalable to all U.S. homes.

Practical Implications

The usage of building archetypes in UBEMs could improve initial guesses for modeling inputs. Representing characterized archetype variables with feasible ranges instead of singular values enhances the robustness of archetype variables and the reliability of energy simulations.

Introduction

Buildings are significant contributors to global energy consumption and greenhouse gas (GHG) emissions, accounting for approximately one-third of global energy use and emissions related to building operations (United Nations Environment Programme, 2022). The building sector is a key target for achieving energy efficiency and decarbonization goals. As urban populations grow and the demand for higher living standards increases, cities worldwide are adopting ambitious targets to mitigate climate change and improve resiliency. For the building sector, these initiatives often focus on retrofitting existing buildings to enhance energy efficiency and incorporating renewable energy systems.

To understand how the existing building stock performs individually and as a whole, we need Urban Building Energy Models (UBEMs) to support prioritizing retrofitting strategies and assessing building upgrade scenarios (Kastner and Dogan, 2024; Reinhart and Cerezo Davila, 2016). By integrating bottom-up building energy simulations with urban data, UBEMs offer a bottom-up simulation framework capable of analyzing building energy consumption across entire cities or neighborhoods.

Archetype-based Urban Building Energy Models (UBEMs)

Although UBEMs are typically capable of delivering detailed energy performance predictions at the building level, they often rely on extensive and precise input data, which is commonly available only through laborious manual surveys for individual buildings. Typically, openly accessible urban datasets only provide high-level building attributes and aggregated demographic information. The absence of building-level records on envelope characteristics, heating and cooling system types, and user activity necessitates informed assumptions based on stock-level knowledge to bridge the gap.

A building stock is characterized by stunning internal variations and heterogeneity; however, based on inter-building similarity, it can be categorized into building types, reference buildings, or building archetypes (Ali et al., 2019). Typologies refer to the function of buildings; reference buildings are typically the “average” representation of building groups identified through expert knowledge; and building archetypes are derived from an unsupervised classification of the building stock based on patterns and internal similarities of the building stock using data-mining techniques. In state-of-the-art UBEMs, building archetypes are particularly useful for researchers and energy modelers to compensate information “gaps” in the local building stock or local

https://doi.org/10.26868/25222708.2025.11336

UBEM based on generalized archetypes derived from national building stock surveys (see Figure 1). With the “known attributes” about each building, archetypes allow for improved confidence in the “initial guess” inputs for energy models.

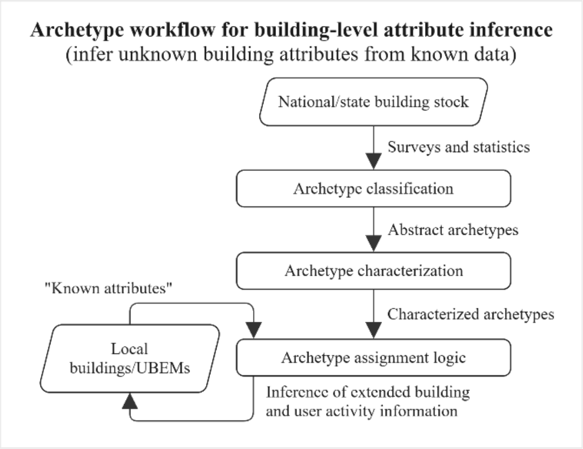

Archetype workflow for building-level attribute inference
(infer unknown building attributes from known data)
National/state building stock
Surveys and statistics
Archetype classification
Abstract archetypes
Archetype characterization
"Known attributes"
Characterized archetypes
Local buildings/UBEMs
Archetype assignment logic
Inference of extended building and user activity information

Figure 1: Workflow using archetypes to infer extra information about buildings

Towards a Modular, Data-Driven Archetype Framework

While archetype-based UBEMs offer a practical means of simulating urban-scale building energy use, existing methods often rely on simplified prototypes or expert-defined reference buildings that lack flexibility and adaptability to individual building properties and distinct local conditions. State-of-the-art archetype frameworks define archetypes at the whole-building level, limiting their capacity to represent diverse combinations of construction features, system configurations, and energy use behaviors. The desire to reduce the total number of archetypes for the sake of simplicity and tractability often sacrifices the ability to represent unique and meaningful variations across building typologies, envelope constructions, HVAC systems, and usage patterns. When multiple dimensions of a building's characteristics are bundled into a single, monolithic archetype, it becomes difficult to explain or interpret their mixed effect on energy outcomes. A modular, combinatory approach—separating construction, systems, and energy loads into distinct sub-archetypes—can better reflect real-world diversity and enhance model explainability.

Additionally, most current archetype applications rely on deterministic point values for simulation inputs, even though these inputs are only informed “initial guesses” about typical performance or characteristics. In practice, these parameters involve uncertainty and are context-dependent, particularly at the urban scale. Representing archetype attributes as bounded ranges or probability distributions rather than fixed values enables a more realistic starting point for model calibration. It also accommodates data uncertainty and supports advanced methods such as Bayesian calibration. While some recent UBEM studies have adopted feasible-range or probabilistic representations of archetypes (Sokol et al., 2017, 2017). They are often demonstrated as one-off case studies about specific cities or housing types, with limited transferability or standardization.

In the context of UBEMs in the United States, this limitation is especially apparent. Although rich national energy datasets like the Residential Energy Consumption Survey (RECS) (U.S. Energy Information Administration, 2023) provide comprehensive coverage of building construction, systems, and occupant behaviors, there is currently no standardized, generalizable approach to generate RECS-based archetypes. For example, a notable UBEM effort in New York State (Shen et al., 2021) used fixed building geometries per typology derived from DOE prototype buildings (U.S. Energy Information Administration, 2023), treating RECS data primarily as calibration targets. This underutilizes the potential of RECS to inform bottom-up variations at the individual building level. A scalable framework that translates RECS data into assignable modular archetypes would significantly improve fidelity, interpretability, and transferability of UBEM inputs across the U.S. housing stock. This research, therefore, aims to address the identified gaps in the following aspects:

1. Decoupling building archetypes into interpretable, combinable sub-archetypes to account for building-level variations in construction, systems, and loads;
2. Representation of archetype attributes as feasible ranges to better support calibration and uncertainty quantification in the building energy models created;
3. Usage of RECS in a reproducible, scalable manner to assign archetypes to individual buildings based on openly available urban data.

Methods

In response to the identified gaps, we propose a modular archetype approach generalizable to all climatic regions in the U.S. as specified in RECS. This approach decouples or decomposes traditional building archetypes into three sub-archetypes: (building) construction, (heating and cooling) systems, and energy loads, allowing for combinatory configurations that improve representation of heterogeneity with a manageable number of definitions.

Urban data collection and “known attributes” extraction

Data collection and preprocessing are the initial and most crucial steps to build an archetype-based UBEM model, as they help researchers determine the scope of available information or the range of “known attributes” about each building, to identify the information “gaps” for RECS-derived archetypes to bridge. For the purpose of demonstration, our workflow is deployed on a UBEM for the residential building stock of a city in the northeastern U.S. The workflow, as illustrated in Figure 3, begins with extracting building geometry, tax parcel data, and assessment information from the county's ESRI feature server. Spatial operations are applied to enrich building datasets by joining additional attributes, such as thermal systems, vintage, and improvement records, to the building geometries. Site GIS layers, including census demographic data, rural-urban definitions, and income

https://doi.org/10.26868/25222708.2025.11336

information, are incorporated through left joins and area-based extrapolation to reflect population and household attributes for the region.

The validation dataset consists of 4 consecutive years of monthly billed electricity and energy consumption at the meter level. Supplied with only coarse address text files, this workflow adopts geocoding, a method of probabilistically associating text-based addresses or place names with corresponding geographic coordinates (Roongpiboonsopit and Karimi, 2010) to align billed energy consumption to spatial building records.

Apart from building and occupant information, effective UBEMs require high-quality local weather data. The National Solar Radiation Database (NSRDB) (Sengupta et al., 2018) provides an application programming interface (API) to access typical and historical weather data. This study utilizes this web API to retrieve Actual Meteorological Year (AMY) inputs for the UBEM.

Using building archetypes to enrich open urban data

As a result of the urban data collection process, we have analyzed urban attributes with consistent availability and have identified 12 “known attributes” as shown in Figure 2 to assist the building-level archetype assignment. The “known attributes” include “housing typology” as defined by RECS: *mobile homes, single-family detached homes, single-family attached homes, homes in apartment buildings with 2-4 units, and homes in apartment buildings with more than 5 units*. We therefore introduce an archetype classification, assignment, and characterization pipeline to enrich energy modeling inputs based on the high-fidelity information in nationwide energy consumption surveys.

As the demonstration site is located in Köppen-Geiger climate zone 6A (Beck et al., 2018), this paper presents the residential building archetype workflow for this climate zone only for simplicity. and the associated national energy consumption survey is RECS, with which archetypes will be derived, characterized, and assigned to buildings based on the 12 “known attributes” shared by both open urban data and RECS to assist the assignment of archetypes (see Figure 2).

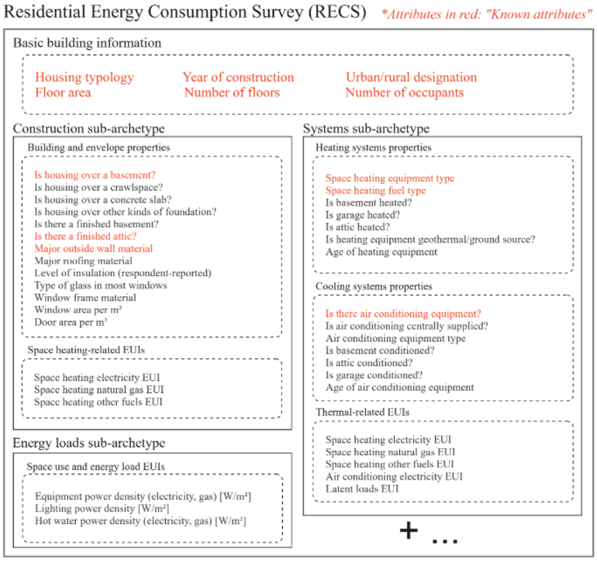

Residential Energy Consumption Survey (RECS)
*Attributes in red: "Known attributes"
Basic building information
Housing typology
Floor area
Year of construction
Number of floors
Urban/rural designation
Number of occupants
Construction sub-archetype
Building and envelope properties
Is housing over a basement?
Is housing over a crawlspace?
Is housing over a concrete slab?
Is housing over other kinds of foundation?
Is there a finished basement?
Is there a finished attic?
Major outside wall material
Major roofing material
Level of insulation (respondent-reported)
Type of glass in most windows
Window frame material
Window area per m²
Door area per m²
Space heating-related EUIs
Space heating electricity EUI
Space heating natural gas EUI
Space heating other fuels EUI
Energy loads sub-archetype
Space use and energy load EUIs
Equipment power density (electricity, gas) [W/m²]
Lighting power density [W/m²]
Hot water power density (electricity, gas) [W/m²]
Systems sub-archetype
Heating systems properties
Space heating equipment type
Space heating fuel type
Is basement heated?
Is garage heated?
Is attic heated?
Is heating equipment geothermal/ground source?
Age of heating equipment
Cooling systems properties
Is there air conditioning equipment?
Is air conditioning centrally supplied?
Air conditioning equipment type
Is basement conditioned?
Is attic conditioned?
Is garage conditioned?
Age of air conditioning equipment
Thermal-related EUIs
Space heating electricity EUI
Space heating natural gas EUI
Space heating other fuels EUI
Air conditioning electricity EUI
Latent loads EUI
+ ...

Figure 2: RECS attributes and "known attributes"

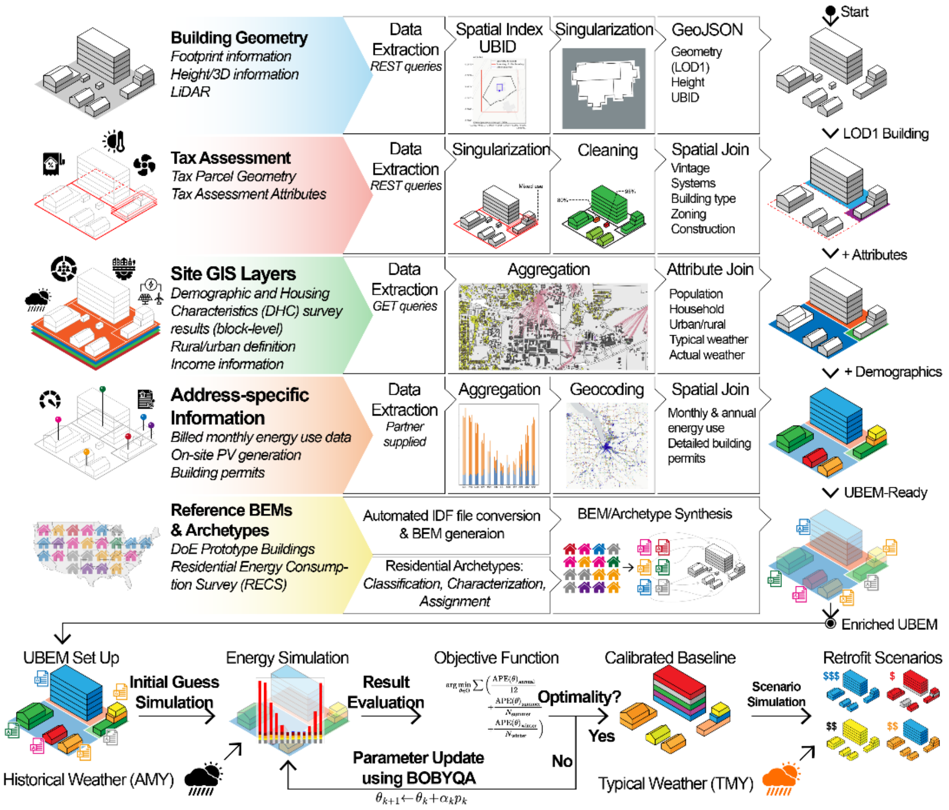


Figure 3: Archetype UBEM workflow: data collection, UBEM formulation, archetype application and calibration

https://doi.org/10.26868/25222708.2025.11336

Separation of building archetype into sub-archetypes

Traditional building archetypes with a limited number of 20-30 often fail to adequately represent the diversity of real-world combinations of building physical properties, building systems, and energy use behaviors. While such archetypes may encourage interpretability, they exclude important variations in energy performance and usage patterns. This study thus separates the original concept of unitary building-level archetypes and decouples them into specialized sub-archetypes according to the "attribute section" definition in RECS. Sub-archetypes are defined for 3 themes: building construction and envelope properties (construction sub-archetype); heating, cooling and ventilation systems (systems sub-archetype); and user-behavior related space use and energy loads (energy loads sub-archetype). Detailed descriptions for the sub-archetype attributes and "known attributes" are described in Figure 2. Thus, the archetypes applied to each building in the UBEM will be a combination of a construction sub-archetype, a systems sub-archetype, and an energy loads sub-archetype. This ensures modularity and flexibility and could represent heterogeneous building stock characteristics with fewer sub-archetype definitions.

The generation of each sub-archetype follows a classification–characterization–assignment workflow as discussed in the following section.

Clustering and determining the optimal number of clusters

To best describe the variability within each typological group, a stratified clustering approach is adopted. Housing typology and climate zone are regarded by existing research to have significant impacts on household energy use patterns in terms of both intensity and fuel type (Shen et al., 2021). As described earlier, RECS defines 5 housing typologies, and the initial step of sub-archetype generation thus involves a stratified subdivision of the RECS dataset based on climate zone and typology into several subsets to be further clustered.

For the clustering step, two clustering algorithms are compared: a K-Means clustering and a Gaussian Mixture Model (GMM) based clustering with Expectation Maximization (EM) algorithm (Murphy, 2014).The optimal number of clusters is determined by the combination of an "elbow method" based clustering performance evaluation, and the possibility of predicting the correct cluster label from the "known attributes."

K-means divides observations to k clusters and minimizes the sum of distances of the observation samples to the centroid of the cluster it belongs to, expressed as

$$\mathrm{argmin}_{C_k} \sum_{k=1}^{K} \sum_{x_i \in C_k} | x_i - \mu_k |^2 \qquad (1)$$

where C_k are clusters, μ_k is the centroid of cluster k, and x_i is a data point in cluster C_k, and this research uses the Euclidean distance $|x_i - \mu_k|^2$. In K-means, the measure of fit is the sum of squared distances (SSD) between each data point and its cluster centroid, defined as

$$\mathrm{SSD} = \sum_{k=1}^{K} \sum_{x_i \in C_k} | x_i - \mu_k |^2 \qquad (2)$$

where C_k represents the clusters, μ_k is the centroid of cluster k, and x_i are the data points in cluster C_k.

Another clustering algorithm with a probabilistic nature is the Gaussian Mixture Model (GMM). The objective is to maximize the likelihood of the data given the mixture model, expressed as

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(x \mid \mu_k, \Sigma_k \right) \qquad (3)$$

where $p(x)$ is the probability density function, π_k is the weight of the k-th Gaussian component (with $\sum_{k=1}^{K} \pi_k = 1$), $\mathcal{N}\left(x \mid \mu_k, \Sigma_k \right)$ is the Gaussian distribution for cluster k, μ_k is the mean vector, and Σ_k is the covariance matrix. GMM employs the Expectation-Maximization (EM) algorithm to estimate these parameters iteratively. To determine the optimal number of clusters in GMMs, the Bayesian Information Criterion (BIC) is used:

$$\mathrm{BIC} = -2 \cdot \ln(\mathcal{L}) + p \cdot \ln(n) \qquad (4)$$

where $\mathcal{L}$ is the likelihood of the data given the model, p is the number of parameters in the model, and n is the number of data points.

Due to the need to assign archetypes to buildings with only "known attributes" available, the optimal choice for the number of clusters also relies on the ability to predict the cluster assignment from the "known attributes." Thus, apart from evaluating clustering performance using the elbow method, the cluster prediction accuracy from "known attributes" in open urban data serves as an additional evaluative criterion. Multiple machine learning models, including Logistic Regression, Gradient Boosting, Random Forest, and Support Vector Machine (Murphy, 2014) are employed and compared.

Random Forest is an ensemble learning method that combines the predictions of multiple decision trees. Each tree is trained on a bootstrap sample of the data, and a subset of features is randomly selected at each split. The final prediction is obtained through majority voting:

$$y = \mathrm{argmax}_c \frac{1}{N} \sum_{i=1}^{N} P_i(c|x) \qquad (5)$$

where N is the number of trees, and $P_i(c \mid x)$ is the predicted probability of class c by tree i.

Gradient Boosting is an ensemble method that builds a strong predictive model by combining the outputs of multiple weak learners sequentially. At each iteration, the model minimizes a loss function by fitting a new tree to the residuals of the previous model. The prediction for a data point x is given as:

$$f(x) = \sum_{m=1}^{M} \alpha_m h_m(x) \qquad (6)$$

where $h_m(x)$ is the output of the m-th weak learner, α_m is its weight, and M is the total number of trees.

Logistic Regression is a linear model that predicts the probability of a data point belonging to a specific cluster using a sigmoid function, outputting the probability

$$P(y = 1|x) = \frac{1}{1+e^{-(w^T x + b)}} \qquad (7)$$

where w is the weight vector, b is the bias, and x represents the input features. The decision boundary is determined by the linear combination of the features. The cluster label is assigned based on the highest probability.

Support Vector Machine identifies an optimal hyperplane to separate data into clusters. For linearly separable data, the decision boundary is defined as:

$$f(x) = w^T x + b \tag{8}$$

where w is the normal vector to the hyperplane, b is the bias, and x is the input vector.

2 evaluation criteria are used for model selection: prediction accuracy and log loss. Prediction accuracy measures the proportion of correct assignment:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{9}$$

Sub-archetype Characterization

For categorical attributes, we adopt a frequency-based, deterministic characterization, which refers to the process of identifying dominant or representative features within each sub-archetype cluster. This is typically achieved by analyzing the frequency of attribute occurrences or identifying a centroid/medoid as a representative entity within the cluster. Subsequently, each attribute of the resultant sub-archetypes is quantified and translated for standard UBEM input. For example, heating and cooling system categories are translated to the corresponding system specifications with systems parameters and Coefficient of Performance (COP) for typical residential buildings extracted from Building America House Simulation Protocols (Hendron and Engebrecht, 2010).

Apart from the deterministic categorical variables, other variables, either categorical or continuous, typically involve a higher degree of uncertainty during characterization. For example, construction materials are specified as construction names or material categories in RECS, but would have varied thermal performance in different regions, constructions, building styles and vintages. For such variables, they are characterized with a local average value for "initial guess" with a feasible range indicating the maximum and minimum values for each attribute under typical conditions. Specifically, assumptions related to wall and roof u-values and their feasible range are based on typical construction material specifications used in local projects.

Similarly, continuous variables in the energy loads sub-archetypes exhibit high uncertainty and values spread across a wide range, such as equipment power density, lighting power density, gas power density, and hot water power usage. This renders single-valued, deterministic characterization approaches profoundly inappropriate. In a similar fashion, they are characterized with the "initial guess" represented by the median of all values in RECS, with the feasible range roughly based on the 5th and 95th percentile of all associated records in RECS.

Sub-archetypes Assignment

Sub-archetypes are assigned using the best-performing machine learning model trained for the cluster prediction task in the sub-archetype classification stage. Using "known attributes", the model predicts the most possible construction sub-archetype, systems sub-archetype, and energy loads sub-archetype for each building modeled.

Deploying Archetypes in UBEMs

For the UBEM simulation, a 5R1C resistor-capacitor model described in the EN ISO 13790 standard is used to produce accurate hourly building energy simulation at reduced computational costs (Kastner and Dogan, 2024). The sub-archetypes are defined in separate JSON files as complementary input data in addition to the mandatory geometry, building information, and weather input for the UBEM model. Such a versatile configuration also allows the sub-archetype to be potentially reused for common BEM tools such as EnergyPlus.

UBEM Calibration

We quantify UBEM accuracy by the difference between simulated electricity/gas consumption and metered consumption. With the "initial guess" for each archetype variable and the respective feasible range, we have formulated the calibration of the model inputs as an optimization problem with the objective to minimize the difference between simulated and billed energy use, expressed as

$$\theta^* = \arg\min_{\theta \in \Theta} \sum \left(\frac{|E_{\text{simulated}}(\theta) - E_{\text{measured}}|}{E_{\text{measured}}} \right) \tag{10}$$

Where $E_{\text{simulated}}(\theta)$ represents the predicted energy consumption for a specific fuel category (electricity, gas) in a particular season based on the simulation model, while E_{measured} corresponds to the actual billed energy use. The calibration parameters, $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$ represent the archetype inputs, and they are constrained within a feasible search space Θ, as prescribed in the corresponding archetype variable, defined in the form of

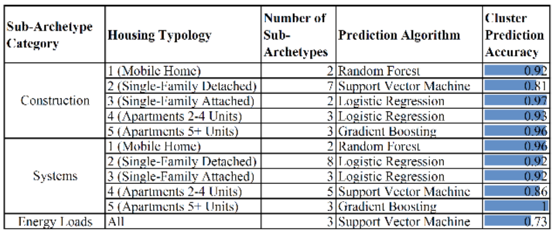
$$\theta_i^{\text{lower}} \leq \theta_i \leq \theta_i^{\text{upper}} \forall i \in \{1, \ldots, n\} \tag{11}$$

For all calibration runs, we used the Bounded Optimization BY Quadratic Approximation (BOBYQA) (Powell, 2009) method to find a local minimum.

Results

Sub-archetype counts and prediction performance

From the sub-archetype classification workflow, we have identified 41 sub-archetypes, comprising 17 construction sub-archetypes; 21 systems sub-archetypes; 3 energy loads sub-archetypes. Combining all sub-archetypes, it is possible to define 270 combinations. Figure 4 describes the optimal clustering algorithm, number of clusters, prediction model and performance for each sub-archetype category, while Figure 5 and Figure 7 compare performances for different numbers of clusters.



Sub-Archetype Category	Housing Typology	Number of Sub-Archetypes	Prediction Algorithm	Cluster Prediction Accuracy
Construction	1 (Mobile Home)	2	Random Forest	0.92
	2 (Single-Family Detached)	7	Support Vector Machine	0.81
	3 (Single-Family Attached)	2	Logistic Regression	0.97
	4 (Apartments 2-4 Units)	3	Logistic Regression	0.93
	5 (Apartments 5+ Units)	3	Gradient Boosting	0.96
Systems	1 (Mobile Home)	2	Random Forest	0.96
	2 (Single-Family Detached)	8	Logistic Regression	0.92
	3 (Single-Family Attached)	3	Logistic Regression	0.92
	4 (Apartments 2-4 Units)	5	Support Vector Machine	0.86
	5 (Apartments 5+ Units)	3	Gradient Boosting	1
Energy Loads	All	3	Support Vector Machine	0.73

Figure 4: Optimal sub-archetype counts and prediction method for each sub-archetype category

https://doi.org/10.26868/25222708.2025.11336

Construction Sub-archetypes

Table 1: List of construction sub-archetypes

Type	Specifications
1	Standard Vinyl Window Mobile Home
	Standard Metal Window Mobile Home
2	Siding Façade Home
	Siding Façade, Finished Basement, Big WWR
	Wood Façade Home
	Siding Façade, Finished Basement, Mod. WWR
	Siding Façade Home with Metal Roof
	Poorly Insulated Shingles Façade Home
	Well-Insulated Siding Façade Home
3	Attached Home with Moderate WWR
	Attached Home with Big WWR
4	Apartment with Big WWR
	Apartment with Moderate WWR
	Wood Façade Apartment
5	Apartment with Metal Windows (Electric Heating)
	Apartment with Metal Windows (Gas Heating)
	Apartment with Wood Windows

17 construction sub-archetypes are defined (see Table 1), with 2 for mobile homes (housing typology 1), 7 for single-family housing detached (2), 2 for single-family housing attached (3), 3 for apartments in buildings with 2-4 units (4), and 3 for apartments in buildings with 5+ units (5). The construction sub-archetypes highlight distinct wall and roof materials for non-apartment buildings, whilst for apartments, window-related properties contribute to the significant inter-cluster differences.

Systems Sub-archetypes

Table 2: List of systems sub-archetypes

Type	Specifications
1	Gas Furnace, Central AC Mobile Home ()
	Oil Furnace, Wall AC Mobile Home
2	Propane Furnace, Central AC Home
	Typical-Efficiency Gas Furnace, Central AC Home
	Oil Furnace, Wall AC Home with Gas
	Low-Efficiency Gas Furnace, Wall AC Home
	Heat Pump, Central AC Home
	Oil Furnace, Wall AC Home with No Gas
	Electric Heating, Wall AC Home
	High-Efficiency Gas Furnace, Central AC Home
3	Old Gas Furnace, Central AC Attached Home
	Typical Gas Furnace, Central AC Attached Home
	Propane Furnace, Wall AC Attached Home
4	Oil Furnace, Wall AC Apartment
	Gas Steam Heater, Wall AC Apartment
	Gas Furnace, Central AC Apartment
	Electric Heating, Wall AC Apartment
	Oil Steam Heater, Wall AC Apartment
5	Gas Furnace, Central AC Multi-Apartment
	Propane Steam Heater, Wall AC Multi-Apartment
	Electric Heating, Wall AC Multi-Apartment

21 systems sub-archetypes are defined (Table 2), with 2 for mobile homes (1), 8 for single-family housing detached (2), 3 for single-family housing attached (3), 5 for apartments in buildings 2-4 units (4), 3 for apartments in buildings with 5+ units (5). The systems sub-archetypes make a clear distinction between heating equipment types.

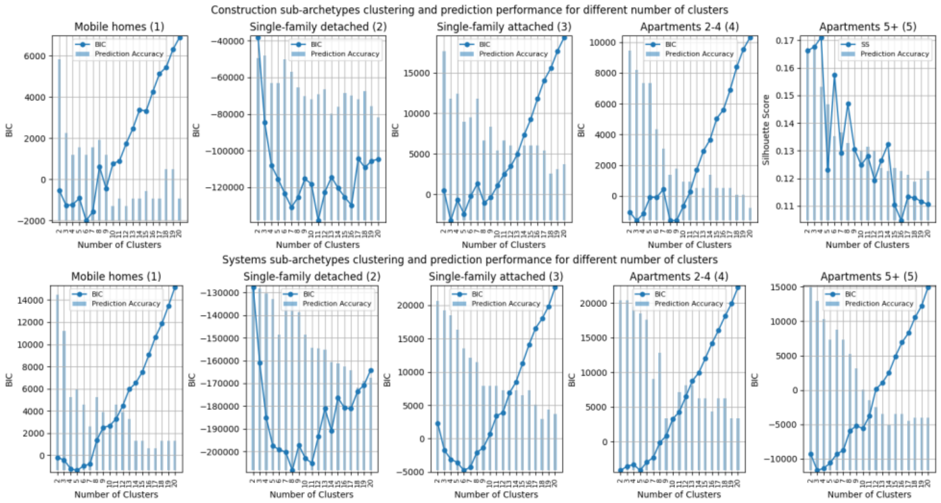
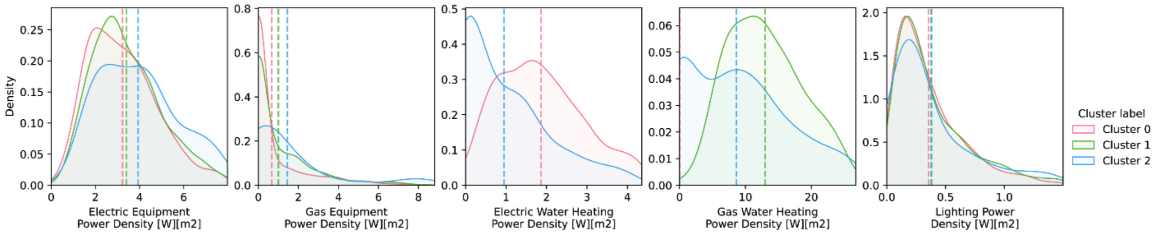


Figure 5: Construction and systems sub-archetypes clustering and prediction performance



label	5%	25%	median	75%	95%	5%	25%	median	75%	95%	5%	25%	median	75%	95%	5%	25%	median	75%	95%	5%	25%	median	75%	95%
0	1.172	2.002	3.127	4.338	7.405	0.000	0.000	0.000	0.601	3.924	0.413	1.101	1.956	3.098	6.205	0.000	0.000	0.000	0.000	0.000	0.061	0.142	0.283	0.517	1.162
1	1.317	2.314	3.258	4.632	8.174	0.000	0.000	0.000	1.628	4.314	0.000	0.000	0.000	0.000	0.000	4.859	8.905	13.063	18.546	30.054	0.065	0.157	0.294	0.584	1.549
2	1.566	2.516	3.957	5.432	8.029	0.000	0.964	11.363	21.312	46.812	0.000	0.000	0.711	1.604	3.919	0.000	0.000	8.059	15.041	27.530	0.057	0.151	0.282	0.717	3.111

Figure 6: Continuous values distributions for 3 attributes for all 3 energy loads sub-archetypes

https://doi.org/10.26868/25222708.2025.11336

Energy Loads Sub-archetypes

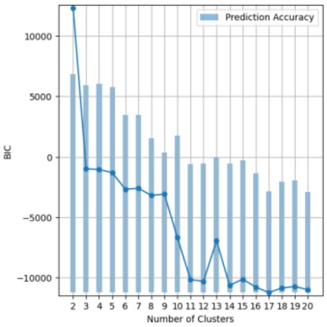


Figure 7: Energy loads sub-archetypes clustering and prediction performances

3 energy loads sub-archetypes are defined. Although the reduced number of energy loads sub-archetypes may have difficulty accounting for the variety of user behavior within residential buildings, this is adequately addressed by the feasible-range-based characterization for each attribute. The per-cluster distributions for all three continuous attributes for each energy loads sub-archetype are presented in Figure 6.

UBEM Performance

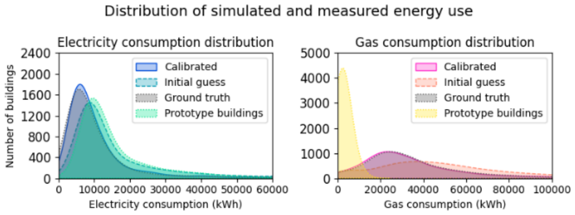


Figure 8 Comparison of the distribution of energy consumption between various inputs and ground truth

To assess the effectiveness of the archetype-based approach, we have compared the following settings against metered annual electricity and gas use: simulation with DOE prototype buildings, archetype-informed “initial guesses,” and the calibrated “initial guesses,” As illustrated in Figure 8.

Discussions

Stratified clustering makes archetypes interpretable

The results confirm that stratified clustering improves archetype interpretability, ensuring that the generated sub-archetypes accurately reflect the diversity in building stock characteristics while maintaining practical interpretability and usability for UBEM applications (see Table 1 and Table 2). The identified 41 sub-archetypes effectively capture variations in construction, systems, and energy loads, providing a structured approach to characterizing the complexity inherent in residential buildings in climate zone 6A. The integration of machine learning-based classification methods further enhances the accuracy of archetype assignment, reinforcing the utility of this framework in large-scale urban simulations.

Modular, high-fidelity archetypes are readable and reusable encodings of the complex building stock

The high-fidelity archetypes derived in this study encode well the variability within the residential building stock. By decomposing traditional archetypes into three distinct sub-archetype categories, this method allows for modular representation of residential energy behaviors and system configurations, as confirmed by Figure 9: Assignment of sub-archetypes for the residential building stock in the studied area which showcases the assigned sub-archetypes per theme for all residential buildings in the studied area. This modularity not only enhances interpretability but also aligns with scalable UBEM workflows. The combination of deterministic and range-based characterization techniques can address uncertainty in urban-scale energy modeling.

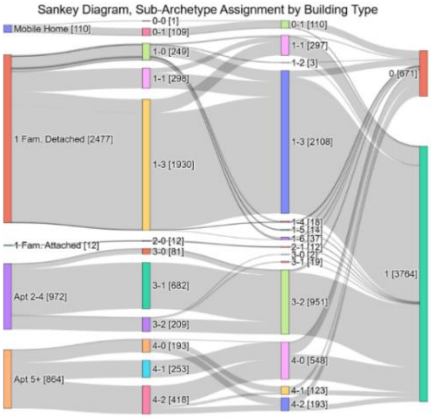


Figure 9: Assignment of sub-archetypes for the residential building stock in the studied area

Archetypes are summaries of stock-level knowledge and could be extrapolated to improve initial building energy model inputs (“initial guesses”)

Validating the simulation results against validation energy consumption data, we have demonstrated that our archetype workflow effectively extrapolates building-level knowledge from stock-level data and compensates for the “gaps” by serving as better-informed “initial guesses” for energy model inputs compared to DOE prototype buildings. The proposed workflow enriches UBEM input assumptions by reducing reliance on overly generalized, deterministic “defaults.” The validation results show that initial energy simulations with archetypes yield better alignment with validation data compared to DOE prototype buildings, especially for gas (Figure 8). Most gas consumption in this region is due to thermal loads, indicating that the “initial guesses” are capable of correctly inferring envelope properties and conditioning system specifications, particularly important for retrofit scenario simulations where the inference of building baseline qualities impacts the feasibility and effectiveness of potential building refurbishments and system upgrade measures. For example, the archetypes infer the age of gas boilers, water heating efficiency, or the types of glazing. They are essential indicators of the eligibility for policy incentives (Dogan et al., 2025). Meanwhile, for electricity consumption, the improvement over prototype buildings is comparably minor, due to the inherent stochasticity related to user behavior and household demographics, which could not be abundantly reflected in either RECS or open urban data.

https://doi.org/10.26868/25222708.2025.11336

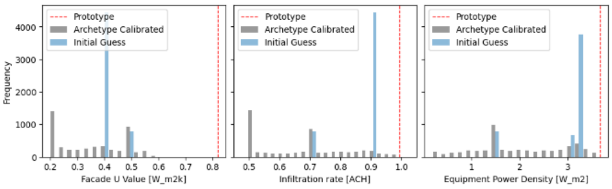


Figure 10 Prototype building, archetype "initial guesses," and calibrated input value distributions

Further, the "deep dive" into the comparison of prototype building properties, archetype characteristics ("initial guesses"), and calibrated parameters demonstrates the necessary "correction" of the prototype building when applied to a local context of the studied building stock as prototypes are assuming a generally worse case of thermal performance and energy consumption in general (Figure 10). These results further highlight the advantage of archetype-informed initial simulation performance over traditional prototype assumptions.

Conclusion

UBEM emerges as a tool for simulating urban building energy use and has served as the decision-making support for building retrofit scenarios and stock-level decarbonization pathways. A key challenge is the acquisition of building-level information for the building stock. To address this challenge, this study introduced a novel method for generating residential building archetypes from RECS to enrich open urban data and bridge the "gaps" in building information inputs for energy models. These archetypes are defined across construction, systems, and energy loads sub-archetypes for flexible, combinatory assignment. Using range-based characterization, this method enhances UBEM performance based on "initial guess" parameters, resulting in more accurate UBEM simulations when calibrated. This dual-source open data integration ensures reusability and scalability across the entire U.S., enabling broader impact. Furthermore, the non-deterministic, ranged characterization of sub-archetype attributes may welcome a probability-based characterization and Bayesian calibration methods (Cerezo et al., 2017; Sokol et al., 2017). This work represents a significant step towards scalable, data-driven, automatable UBEMs and would contribute significantly to the decision-making process for building stock decarbonization.

References

Ali, U., Shamsi, M.H., Hoare, C., Mangina, E., O'Donnell, J., 2019. A data-driven approach for multi-scale building archetypes development. Energy and Buildings 202, 109364. https://doi.org/10.1016/j.enbuild.2019.109364

Beck, H.E., Zimmermann, N.E., McVicar, T.R., Vergopolan, N., Berg, A., Wood, E.F., 2018. Present and future Köppen-Geiger climate classification maps at 1-km resolution. Sci Data 5, 180214. https://doi.org/10.1038/sdata.2018.214

Cerezo, C., Sokol, J., AlKhaled, S., Reinhart, C., Al-Mumin, A., Hajiah, A., 2017. Comparison of four building archetype characterization methods in urban building energy modeling (UBEM): A residential case study in Kuwait City. Energy and Buildings 154, 321–334. https://doi.org/10.1016/j.enbuild.2017.08.029

Dogan, T., Li, C., Tseng, H.M., Su, A.J., Kastner, P., 2025. A bottom-up urban building energy model for evaluating thermal load electrification measures. Journal of Building Performance Simulation 1–28. https://doi.org/10.1080/19401493.2025.2536261

Hendron, R., Engebrecht, C., 2010. Building America House Simulation Protocols (Revised) (No. NREL/TP-550-49246, DOE/GO-102010-3141, 989422). https://doi.org/10.2172/989422

Kastner, P., Dogan, T., 2024. Towards Auto-Calibrated UBEM Using Readily Available, Underutilized Urban Data: A Case Study for Ithaca, NY. Energy and Buildings 317, 114286. https://doi.org/10.1016/j.enbuild.2024.114286

Murphy, K.P., 2014. Machine Learning - A Probabilistic Perspective, Adaptive Computation and Machine Learning. MIT Press, Cambridge.

Powell, M.J.D., 2009. The BOBYQA algorithm for bound constrained optimization without derivatives.

Reinhart, C.F., Cerezo Davila, C., 2016. Urban building energy modeling – A review of a nascent field. Building and Environment 97, 196–202. https://doi.org/10.1016/j.buildenv.2015.12.001

Roongpiboonsopit, D., Karimi, H.A., 2010. Comparative evaluation and analysis of online geocoding services. International Journal of Geographical Information Science 24, 1081–1100. https://doi.org/10.1080/13658810903289478

Sengupta, M., Xie, Y., Lopez, A., Habte, A., Maclaurin, G., Shelby, J., 2018. The National Solar Radiation Data Base (NSRDB). Renewable and Sustainable Energy Reviews 89, 51–60. https://doi.org/10.1016/j.rser.2018.03.003

Shen, P., Wang, Z., Ji, Y., 2021. Exploring potential for residential energy saving in New York using developed lightweight prototypical building models based on survey data in the past decades. Sustainable Cities and Society 66, 102659. https://doi.org/10.1016/j.scs.2020.102659

Sokol, J., Cerezo Davila, C., Reinhart, C.F., 2017. Validation of a Bayesian-based method for defining residential archetypes in urban building energy models. Energy and Buildings 134, 11–24. https://doi.org/10.1016/j.enbuild.2016.10.050

United Nations Environment Programme, 2022. 2022 Global Status Report for Buildings and Construction: Towards a Zero-emission, Efficient and Resilient Buildings and Construction Sector. Nairobi.

U.S. Energy Information Administration, 2023. 2020 Residential Energy Consumption Survey (RECS) Data.

https://doi.org/10.26868/25222708.2025.11336