An Open-Source Framework for Characterizing Urban Energy Models: Integrating Top-Down and Bottom-Up Methods to Predict Residential Buildings Characteristics

Preprint

Rawad El Kontar,[1,2] Joseph Robertson,[1]
Khanh Nguyen Cu,[1] Alexandra Grayson,[3] Jiazhen Ling,[1]
Hanna Sotiropoulos,[1] and Tarek Rakha[2]

1 National Renewable Energy Laboratory
2 Georgia Institute of Technology
3 University of California, Berkeley

Presented at the 2024 Summer Study on Energy Efficiency in Buildings
Pacific Grove, California
August 4-9, 2024

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

Conference Paper
NREL/CP-5500-90883
August 2024

An Open-Source Framework for Characterizing Urban Energy Models: Integrating Top-Down and Bottom-Up Methods to Predict Residential Buildings Characteristics

Preprint

Rawad El Kontar,[1,2] Joseph Robertson,[1]
Khanh Nguyen Cu,[1] Alexandra Grayson,[3] Jiazhen Ling,[1]
Hanna Sotiropoulos,[1] and Tarek Rakha[2]

Suggested Citation
El Kontar, Rawad, Joseph Robertson, Khanh Nguyen Cu, Alexandra Grayson, Jiazhen Ling, Hanna Sotiropoulos, and Tarek Rakha. 2024. *An Open-Source Framework for Characterizing Urban Energy Models: Integrating Top-Down and Bottom-Up Methods to Predict Residential Buildings Characteristics: Preprint*. Golden, CO: National Renewable Energy Laboratory. NREL/CP-5500-90883. https://www.nrel.gov/docs/fy24osti/90883.pdf.

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

Conference Paper
NREL/CP-5500-90883
August 2024

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

NOTICE

This work was authored in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the U.S. Department of Energy Office of Science. The views expressed herein do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

U.S. Department of Energy (DOE) reports produced after 1991 and a growing number of pre-1991 documents are available free via www.OSTI.gov.

Cover Photos by Dennis Schroeder: (clockwise, left to right) NREL 51934, NREL 45897, NREL 42160, NREL 45891, NREL 48097, NREL 46526.

NREL prints on paper that contains recycled content.

An Open-Source Framework for Characterizing Urban Energy Models: Integrating Top-Down and Bottom-Up Methods to Predict Residential Buildings Characteristics

Rawad El Kontar[12], Joseph Robertson[1], Khanh Nguyen Cu[1], Alexandra Grayson[3], Jiazhen Ling[1], Hanna Sotiropoulos[1], and Tarek Rakha[2]

[1]National Renewable Energy Laboratory
[2]Georgia Institute of Technology
[3]University of California, Berkeley

ABSTRACT

Bottom-up urban energy models are crucial for understanding current energy use patterns and informing design strategies. However, accurately characterizing these models to represent different communities remains a challenge due to the extensive data needed for simulating existing energy use behavior. This data includes information related to human activities and building characteristics, all of which correlate with socioeconomic factors. To overcome this challenge, we developed an automated framework that utilizes both top-down and bottom-up data, to predict unknown building and occupant characteristics that are needed for more accurate and equitable modeling and analytics. Our framework, integrated into the URBANopt district energy modeling platform, uses statistical data models from ResStock. URBANopt models co-located buildings and neighborhoods. At this scale there are data gaps in building characteristic data, such as materials, insulation, occupancy, income, and energy usage of the buildings. To address this data gap, we use ResStock data, representative at the census tract scale, and develop machine-learning and deep-learning techniques to disaggregate it to individual buildings. By mapping unique occupant, building and economic properties to URBANopt energy models, we gain detailed insights into the variability of building energy use across different neighborhoods. This insight helps deploy technologies for co-located buildings and supports targeted upgrades for communities with unique economic and demographic characteristics, ensuring energy equity. Accurate characterization of energy models allows us to develop equitable strategies tailored to diverse neighborhoods, whether underserved or affluent. Our automated framework streamlines energy modeling and provides a reliable tool for building energy characterization.

Introduction

In the United States, the residential sector consumes 16% of total energy and 55% of building energy (EIA 2023), presenting a significant opportunity for conservation. The U.S. government has implemented policies and research to enhance residential energy efficiency (EPA 2023). Accurate forecasting and identification of factors influencing consumption are crucial for

This report is available at no cost from the National Renewable Energy Laboratory at www.nrel.gov/publications.

effective energy management (González-Torres et al. 2021). Addressing climate change requires reducing energy use in buildings. The International Energy Agency (IEA) predicts that by 2040, buildings could be 40% more energy-efficient, primarily by reducing heating, cooling, and water heating energy (IEA 2019). However, developing tailored strategies for energy conservation is challenging due to data gaps. Complete, consistent, and accessible data is essential for precise energy forecasting and identifying impactful energy efficiency measures and building technology upgrades.

Residents define communities and neighborhoods, and they exhibit different energy use behaviors. The buildings they reside in also have varying characteristics, which correlate geospatially with the sociodemographic and economic factors of the residents. Consequently, different neighborhoods exhibit unique behaviors and require tailored energy measures and building upgrades to meet their specific needs. For instance, underserved neighborhoods with a high energy burden might be unable to afford costly technologies and may operate their buildings differently based on work schedules. Therefore, it is crucial to accurately characterize buildings when modeling neighborhood energy use and analyzing new energy efficiency measures. In this sense, accurate data for urban building energy modeling is vital for predicting energy needs, implementing energy-saving measures, and reducing urban carbon footprints (Kontokosta et al. 2019).

Urban energy modeling techniques fall into two categories: (1) top-down models, which utilize econometric or technological methods and aggregated data to generalize current trends, and (2) bottom-up models, which rely on data-driven methods or engineering physics to analyze energy use by studying individual components and their interactions. The bottom-up approach is more frequently used for predicting urban energy consumption. (Kavgic et al. 2010) (Swan 2009).

Recent developments in building energy modeling have leveraged both bottom-up and top-down approaches. For instance, ResStock (Wilson, 2017) was developed to represent the entire residential stock in the U.S., utilizing extensive datasets for U.S. residential buildings to create models. These buildings were characterized using a top-down approach, developing probability distributions for building characteristics from vast datasets (Public ResStock Dataset). These distributions were then used to characterize energy models across U.S. states, with the finest granularity achievable being the Census Public Use Microdata Area (PUMA) level. PUMAs vary in area, contain several census tracts, and have a minimum population of about 100,000. PUMA provides a detailed resolution for understanding building characteristics within a specific area, though it does not pinpoint exact building locations. Consequently, models based on this approach struggle to accurately simulate energy use below the PUMA resolution, limiting their effectiveness in testing technologies at the district or multi-building level.

Conversely, the bottom-up approach begins with localized information, such as building footprint vertices, to accurately model and calibrate energy use. Models like URBANopt™ (El Kontar et al. 2020) (Polly et al. 2016), tailored for district energy modeling, leverage geographic and characteristic data of buildings to precisely simulate districts and neighborhoods. However, these models depend on detailed information encompassing building envelope, building systems, occupant behavior, and economic factors which are often only partially known to URBANopt modelers. In the face of incomplete data, bottom-up models like URBANopt default to the

This report is available at no cost from the National Renewable Energy Laboratory at www.nrel.gov/publications.

International Energy Conservation Code (IECC) and American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) standards, modeling buildings based on Department of Energy (DOE) prototype specifications for various building types. This approach neglects the diverse energy-related characteristics of residential occupants and buildings, leading to a homogenization of building energy models. As a result, the energy demand profiles generated by these models, fails to capture the energy demand variability across different buildings and neighborhoods, and they also misrepresent energy peaks due to inaccurate coincident energy use behaviors.

This generalized approach highlights a critical limitation: it inadequately captures the variability and uniqueness inherent in individual buildings and neighborhoods. Consequently, this can skew predictions of energy demand, leading to inefficient planning and suboptimal energy conservation measures for the different neighborhoods with unique sociodemographic and economic factors. Residential neighborhoods exhibit distinct features and characteristics, ranging from building types, construction properties and building systems differences to variations in occupant behavior and sociodemographic and economic profiles. These characteristics significantly influence energy use within neighborhoods. To develop equitable energy retrofit strategies that cater to communities with diverse sociodemographic and economic statuses, it is crucial to consider and accurately represent occupant behaviors and economic characteristics in our models. A recent study by Palani et al. highlighted that occupant behavior, alongside sociodemographic and economic factors, are pivotal in driving building energy consumption (Palani et al., 2023).

Addressing this discrepancy necessitates the development of refined modeling methods that can accurately capture the variability of buildings and occupants' energy related characteristics. This approach ensures that developed technologies and policies are equitable and effectively targeted. Furthermore, enhancing the granularity and accuracy of bottom-up models, we can achieve a more realistic representation of energy demand profiles. Such improvements would enable more precise predictions of energy consumption and demand, facilitating the implementation of targeted energy efficiency and management strategies that are better aligned with the actual behavior and characteristics of diverse building types. This step forward is essential for advancing our ability to support energy-efficient urban districts and communities in and equitable manner.

This paper introduces a novel framework, integrated into the URBANopt district energy modeling platform, utilizes statistical data models from ResStock. Our methodology addresses the challenge of data scarcity to characterize bottom-up building energy models and introduce innovative methods to disaggregate census tract data to the level of localized, co-located buildings. It enables users with basic building information—such as type, age, and number of stories—to predict essential characteristics for energy modeling. These characteristics include envelope properties, orientation, window area, HVAC system types, efficiencies, occupant behaviors, and income. This innovation enhances the connection between high-level data and detailed, site-specific modeling, thereby improving the accuracy and utility of energy models for diverse communities characterized by distinct sociodemographic, economic, and geographical features.

This report is available at no cost from the National Renewable Energy Laboratory at www.nrel.gov/publications.

The framework includes methods to predict unknown characteristics of actual buildings in neighborhoods. We developed a combination of machine learning techniques, including the K-nearest neighbor algorithm, to find similar sampled buildings and impute the missing characteristics from these nearest neighbors. We also developed a neural network model to predict missing characteristics when only a few are known. The level of known data can vary from one building to another and from one district to another. For this reason, our analysis studies the effectiveness of the models with different levels of information. Additionally, the applicability of the models is analyzed based on complexity and computational time. These methods have been validated and compared to help users understand their applicability for characterizing district energy models.

The developed methods help modelers more accurately characterize district models and capture the variability of features across different neighborhoods. This ensures equity in energy analysis and decisions related to implementing new technologies and energy efficiency strategies tailored for specific neighborhoods with unique sociodemographic, economic, and behavioral characteristics. Accurate characterization is also crucial for deploying suitable technologies for co-located buildings and guiding targeted upgrades to enhance spatial energy efficiency and grid stability. The development and validation of a Baltimore area dataset exemplifies our framework's capability, highlighting its potential to refine energy modeling practices and offering a sophisticated tool for accurately characterizing building energy models.

Methodology

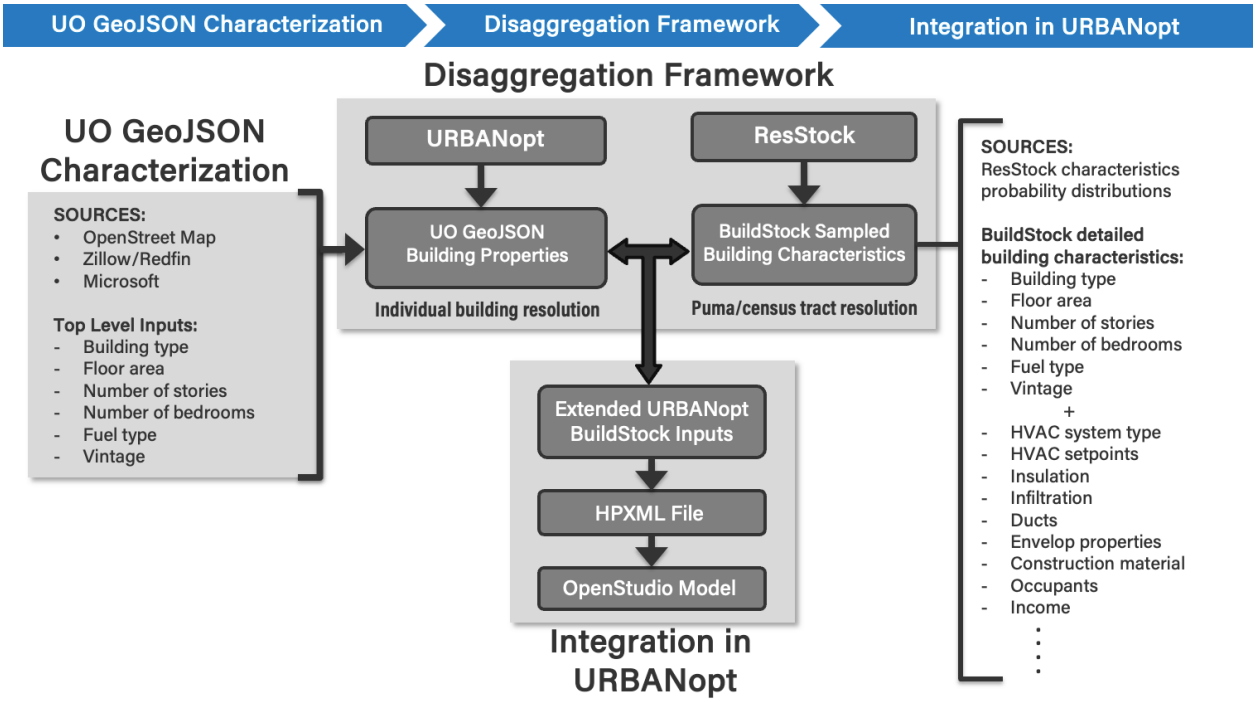

Figure 1. General Framework.

This report is available at no cost from the National Renewable Energy Laboratory at www.nrel.gov/publications.

Figure 1 illustrates the general framework for developing characterized URBANopt models that combine known building information with predicted missing characteristics from ResStock. Initially, users gather collocated buildings information for the district they want to model in URBANopt. This information can be sourced from various public databases such as OpenStreetMap, Redfin, Zillow, and Microsoft Building Footprint. These sources typically provide general details such as residential building type (multifamily, single family detached, single family attached, etc.), floor area, number of stories, number of bedrooms, vintage and in some cases, fuel type and building equipment information, although availability varies by location. The second step involves developing methods that use these initial known characteristics to predict other important characteristics for our energy models using the ResStock dataset. The ResStock data include detailed building characteristic information generated from all locations in the US. However, this data is synthetic data as it is generated using probability distribution from actual data collected at the PUMA resolution. For this reason, the prediction of unknown characteristic falls under the disaggregation framework since we are using the ResStock buildings data set, generated from data at the PUMA resolution, to predict unknown characteristics of collocated building defined in the URBANopt GeoJSON file. After predicting this set of characteristics for the energy model, in the last step, we integrate these characteristics into the URBANopt/OpenStudio framework to update the white box energy model with these crucial inputs and characteristics. In this paper, we briefly describe the 1st and 3rd step of the workflow while we will primarily focus on the details of the disaggregation framework (2nd part), where we develop and compare methods to predict unknown characteristics.

URBANopt GeoJSON file characterization

URBANopt is a bottom-up model that defines geometric and non-geometric inputs to characterize building energy models. The main input for a URBANopt project is a GeoJSON file, which is a standard file format for storing geographic information of geometries. This format has been extended to include building-related properties such as building type, number of stories, HVAC system type, vintage, and other top-level characteristics of a building. A user with limited information can often obtain some of these details from various open-source sources. For example, the longitude and latitude of the vertices of a building footprint can be obtained from OpenStreetMap, while other information like the number of stories, building type, and vintage might be sourced from platforms like Microsoft, Zillow, and Redfin. After gathering this limited information, users can fill in other detailed URBANopt inputs using OpenStudio measures or rely on DOE prototype building characteristics to fill in unknown characteristics and inputs. DOE prototype characteristics are defined for each building type and size category. However, these prototype buildings model each building with the same type and identical characteristics, which falls short in capturing the diversity of characteristics across buildings and neighborhoods since it uses a single archetype of characteristics to identically characterize all buildings of a similar type. Therefore, there is a need to fill in missing characteristics from datasets that capture the variability of characteristics and residential building behavioral inputs, such as ResStock.

Residential Building Energy Model Inputs and Integration in URBANopt

ResStock, on the other hand, is used to create a BuildStock CSV that includes sampled buildings from ResStock probability distribution datasets based on building characteristics. The BuildStock CSV includes multiple detailed building characteristics listed in the diagram and are

This report is available at no cost from the National Renewable Energy Laboratory at www.nrel.gov/publications.

mapped to energy model (OpenStudio/EnergyPlus) inputs using a mapper file that maps these characteristics to arguments for the model inputs. This format and characterization of inputs in the BuildStock CSV is different from the URBANopt GeoJSON file that exposes only a few inputs. Therefore, the integration process involves extending the URBANopt inputs and capabilities to include the BuildStock characteristics and map them to the energy model inputs. This process is done through an HPXML file that reads the characteristics, maps them to energy model inputs, and generates the detailed energy models. This integration step is separated from the rest of the workflow to allow users to develop different disaggregation frameworks or prediction models for the BuildStock characteristics and then easily implement them in a URBANopt model using this integration framework.

Disaggregation Framework: Predicting Unknown Characteristics

Disaggregating ResStock data to individual building URBANopt inputs involve predicting unknown characteristics essential for bottom-up URBANopt model inputs. This prediction process is complex and computationally demanding, as it requires the development of machine learning or AI models capable of inferring multiple missing characteristics from a set of initially known characteristics.

Therefore, we begin with a sensitivity analysis to identify the most important features for prediction. We acknowledge that some features have a minimal impact on the model's outputs. After assessing feature importance, we develop multiple methods to predict the unknown characteristics. Finally, we validate these methods and compare their results. Figure 2 illustrates this framework. Each component represented in the figure will be described in the subsequent paragraphs.

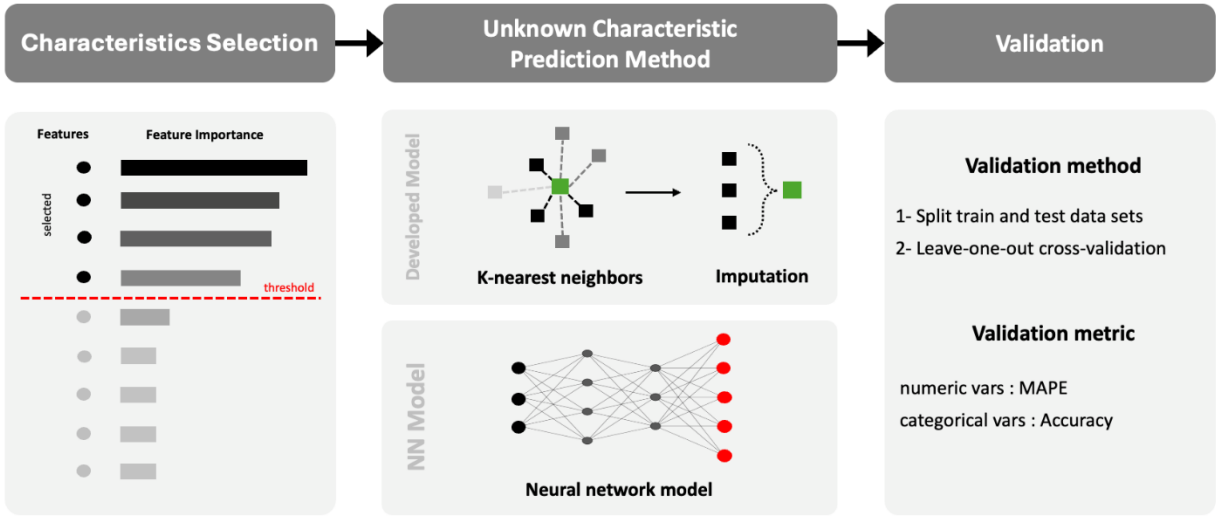

Figure 2. Framework to predict unknown characteristics.

Characteristics Selection

We conducted a sensitivity analysis to identify the most influential characteristics affecting energy use outputs. This involved using a random forest model to explore the relationship between

This report is available at no cost from the National Renewable Energy Laboratory at www.nrel.gov/publications.

various characteristics (X variables) and annual energy use (Y variable). We then extracted feature importance from this model, which ranks the input variables based on their significance in predicting the target variables.

The primary aim was to understand feature importance, so we fine-tuned the random forest model to accurately represent the relationship between input features and outputs before determining the importance of each feature. The top 20 features were identified as critical and will be the focus in the next step of our methodology. Figure 6 showcases the results from the sensitivity analysis, ranking the building characteristics from the ResStock dataset for the Baltimore area by importance.

Unknown characteristics prediction method

To predict missing building characteristics, we can develop models using a subset of known variables to estimate the full set of input features. For instance, with known data points such as building type, floor area, and heating fuel, the model's goal is to predict all other inputs effectively using the sampled buildings and characteristics from the ResStock data. One method involves developing a neural network AI model that predicts multiple outputs from given inputs. Such a model requires extensive computational time and data for training. Additionally, as the initial information can vary for each building, a new model must be developed and trained for each set of inputs, which is both computationally expensive and time-consuming.

Another approach is to find the closest match in the ResStock dataset or to use imputation techniques to estimate unknown variables. To streamline the process and improve accuracy, we developed a methodology utilizing the K-nearest neighbor (KNN) algorithm. This method identifies the nearest neighbors of a building with a few known characteristics and then applies imputation techniques to predict the missing characteristics based on these neighbors.

We apply this method to our ResStock dataset, which consists of both numeric characteristics, such as the number of bedrooms, and categorical characteristics, like insulation type and heating fuel type. Initially, we preprocess the data using a transformer to encode categorical data and standardize all inputs. This preprocessing is crucial for the subsequent application of machine learning and AI models. The encoders used in this step will also allow us to inversely encode the data, enabling us to present users with estimates of previously unknown building characteristics.

Developed KNN + Imputation Method:

In this method, users first define an URBANopt building feature using some known characteristics (building denoted in blue). With these inputs, we identify the K-nearest neighbors (KNN) within the ResStock dataset. The KNN algorithm employs the Hamming distance measure to determine distance among characteristic values. Distances are then weighted based on the results from the sensitivity analysis and closest distances are identified to get the nearest neighbors. The left panel of Figure 3 illustrates how the KNN algorithm calculated distances within ResStock buildings, selecting those with the closest match to the new URBANopt building. When multiple nearest neighbors are identified, we use an imputation technique to estimate the values of the missing inputs among these neighbors. For continuous variables, we calculate the mean, and for categorical variables, we determine the mode, using data derived from the pool of nearest neighbors. This method provides a nuanced estimation that effectively handles both types of data. The process is illustrated in the right-side diagram of Figure 3.

This report is available at no cost from the National Renewable Energy Laboratory at www.nrel.gov/publications.

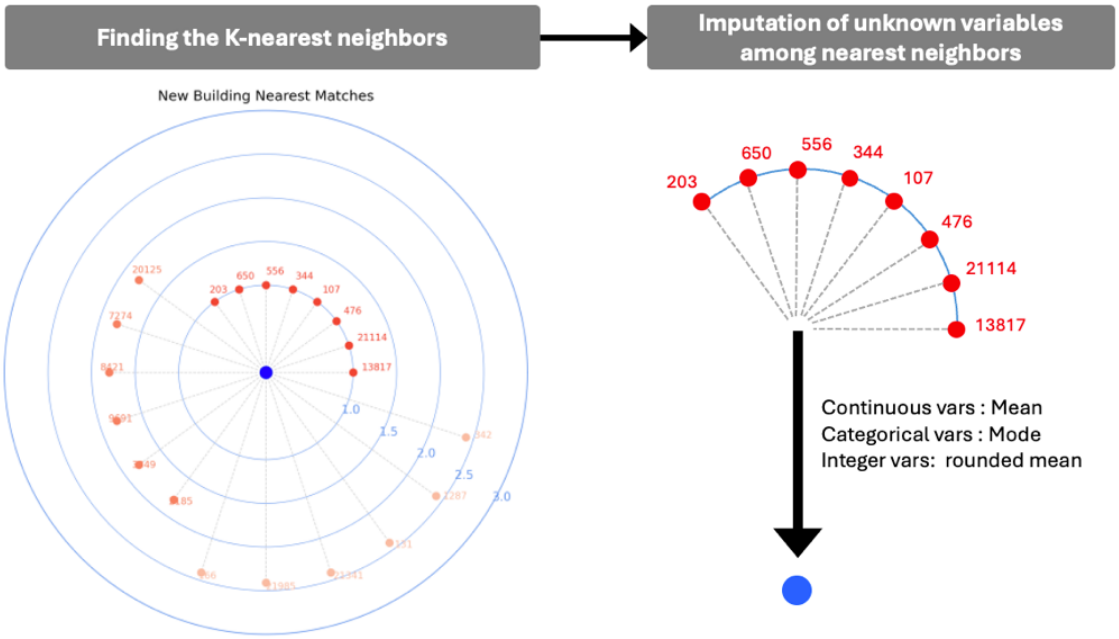

Figure 3. Developed Weighted KNN + Imputation techniques Model

Neural Network Model:

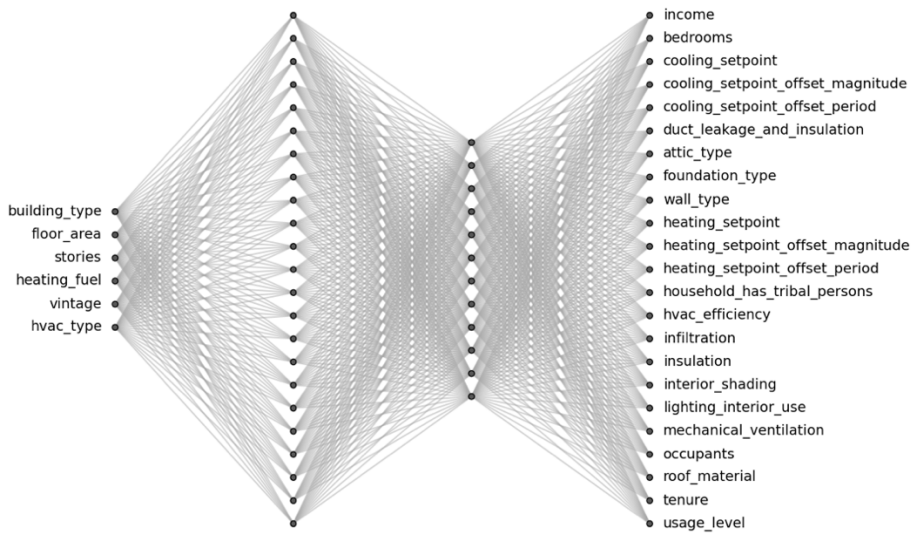

Figure 4. Developed Neural Network Model

In this section, we develop a Neural Network model to predict missing characteristics. We employed cross-validation to determine the optimal Neural Network structure and to tune the hyperparameters. Figure 4 illustrates and example the developed Neural Network model, which is trained to predict missing characteristics based on a few known inputs. The Neural Network model predict all other characteristics using the known characteristics. The chosen network architecture includes two hidden layers, with the first layer containing more nodes than the second, a design commonly referred to as a funnel or pyramid structure. This configuration is effective because it

This report is available at no cost from the National Renewable Energy Laboratory at www.nrel.gov/publications.

allows the network to initially form a broad representation of the input features, which it then refines into more abstract features in subsequent layers. Note that the level of information available from the inputs significantly impacts this method. As new inputs are introduced, we must restructure and train a new Neural Network model to accommodate the new inputs and accurately predict the missing variables.

Validation

In this step, validation is crucial to assess our model's performance and its effectiveness in predicting missing building characteristics. Validation also enables us to compare the two models, helping us determine which is best suited for use. Figure 5 illustrates the validation framework. In this framework, we randomly select a building and remove it from the ResStock data set, then mask unknown characteristics. We then use our developed models to predict these unknown characteristics. Finally, we validate the predicted characteristics against the true masked values. For validation, we use the Mean Absolute Percentage Error (MAPE) for numeric variables. The MAPE metric was chosen to account for magnitude differences across numeric variable values. For categorical variables, we present an accuracy metric that measures the percentage of correct predictions relative to the total number of predictions.

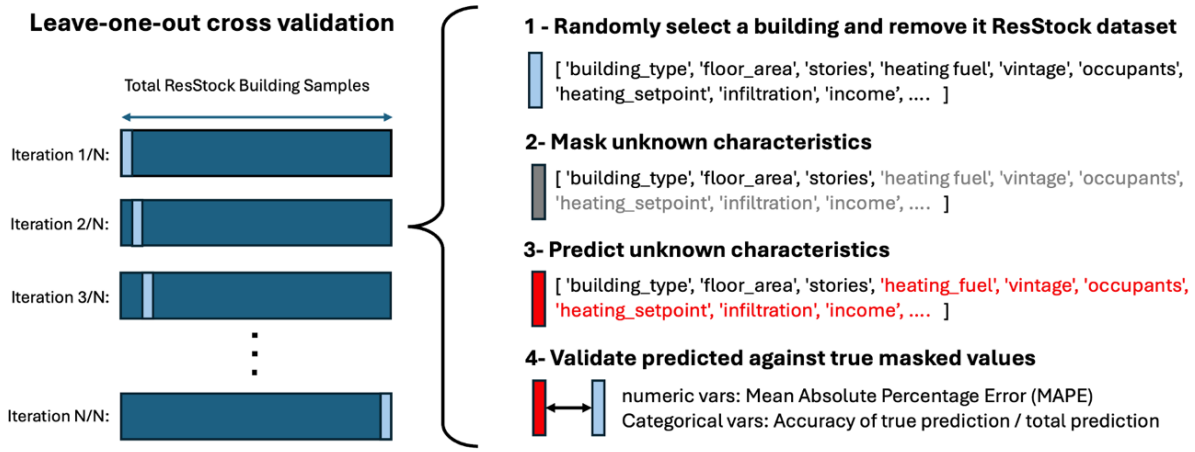

Figure 5. Validation framework.

Results

In this section, we utilized published ResStock data for the state of Baltimore (Present et al. 2024). This dataset includes sampled buildings along with their characteristics and simulated annual energy use. We followed the steps of the framework described in the disaggregation framework methodology section, beginning with a sensitivity analysis to identify the building characteristics that most influence annual energy use. Subsequently, we applied the developed models to predict missing characteristics. The first model combines the KNN and imputation techniques, and the second is the Neural Network model. In the final step, we validated and compared these models.

This report is available at no cost from the National Renewable Energy Laboratory at www.nrel.gov/publications.

Following the methodology steps, Figure 6 illustrates the results of the sensitivity analysis, highlighting how different building characteristics variably impact the annual energy use output. Based on the Figure, building type and floor area are the most prominent bars, indicating the highest impact on energy use. This suggests that the physical characteristics of a building, such as its function and size, are the most significant predictors of its energy consumption. Also, note that floor area and building type are highly correlated. The next most impactful factor is heating fuel, which is also linked to other building characteristics such as building vintage and building systems. Infiltration and insulation follow closely, showing that the quality of a building envelope is crucial in predicting energy use. Insulation and infiltration rates significantly impact heating and cooling loads, thus directly affecting energy consumption. The chart lists numerous other factors with smaller bars, indicating a less significant but still measurable impact on energy use. These include aspects like occupancy level, equipment presence and efficiency, income, building orientation, and various other building features and behaviors. It's important to note that the results of the sensitivity analysis are location dependent. For instance, envelope characteristics like infiltration and insulation may not be as impactful in areas with moderate temperatures, such as California, where occupancy levels and behaviors, along with equipment type and presence, might have a greater influence. Therefore, users should conduct the sensitivity analysis specific to the location of their project. Based on these feature importance results, we selected the top 20 characteristics. These will be used in our predictive models to estimate other variables, tailored to the specific location of each project.

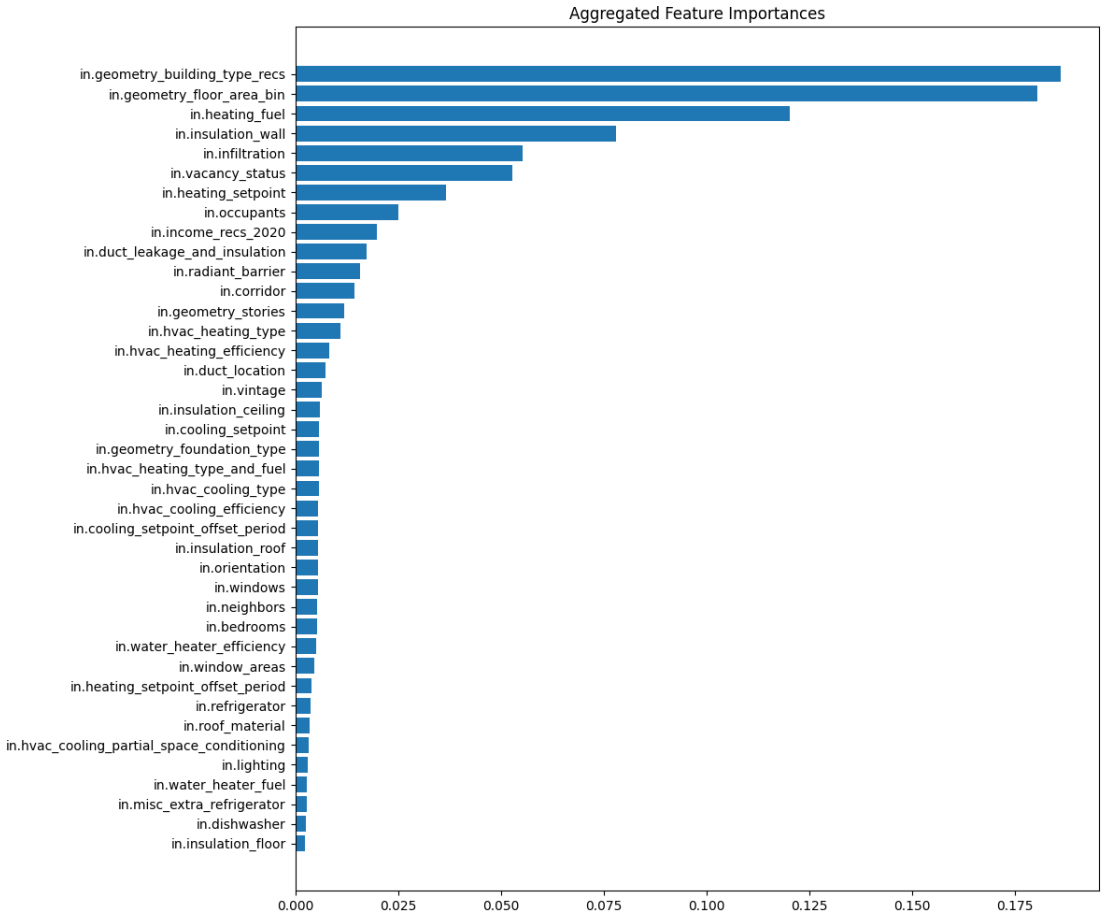

Figure 6. Building characteristics importance results.

This report is available at no cost from the National Renewable Energy Laboratory at www.nrel.gov/publications.

KNN and imputation method results:

For the initial application of our K-nearest neighbor (KNN) methodology, we predict the values of various variables upon receiving a new building entry. We identify the K-nearest neighbors within the dataset. In cases where multiple nearest neighbors are identified, we use statistical measures to infer the values of missing inputs. Specifically, we calculate the mean for continuous variables and the mode for categorical variables, using data derived from the pool of nearest neighbors. This approach allows for an estimation that effectively accommodates both types of data. For this analysis, we considered five known building characteristics, commonly found in datasets such as Zillow, Redfin, and OpenStreetMap. The known variables are as follows: building type, floor area, number of stories, heating fuel, and vintage.

To achieve this, we first preprocess the data by identifying and separating numeric and categorical variables in the dataset. We then encode the categorical variables using one-hot encoding, which preserves distance semantics. Next, we apply the previously developed method and evaluate the model using our established validation framework. The average accuracy for predicting the categorical variables is 77.7%, while the Mean Absolute Percentage Error (MAPE) for the numeric variables is 36%. Figure 7 show the results of this run, broken down by the predicted characteristics.

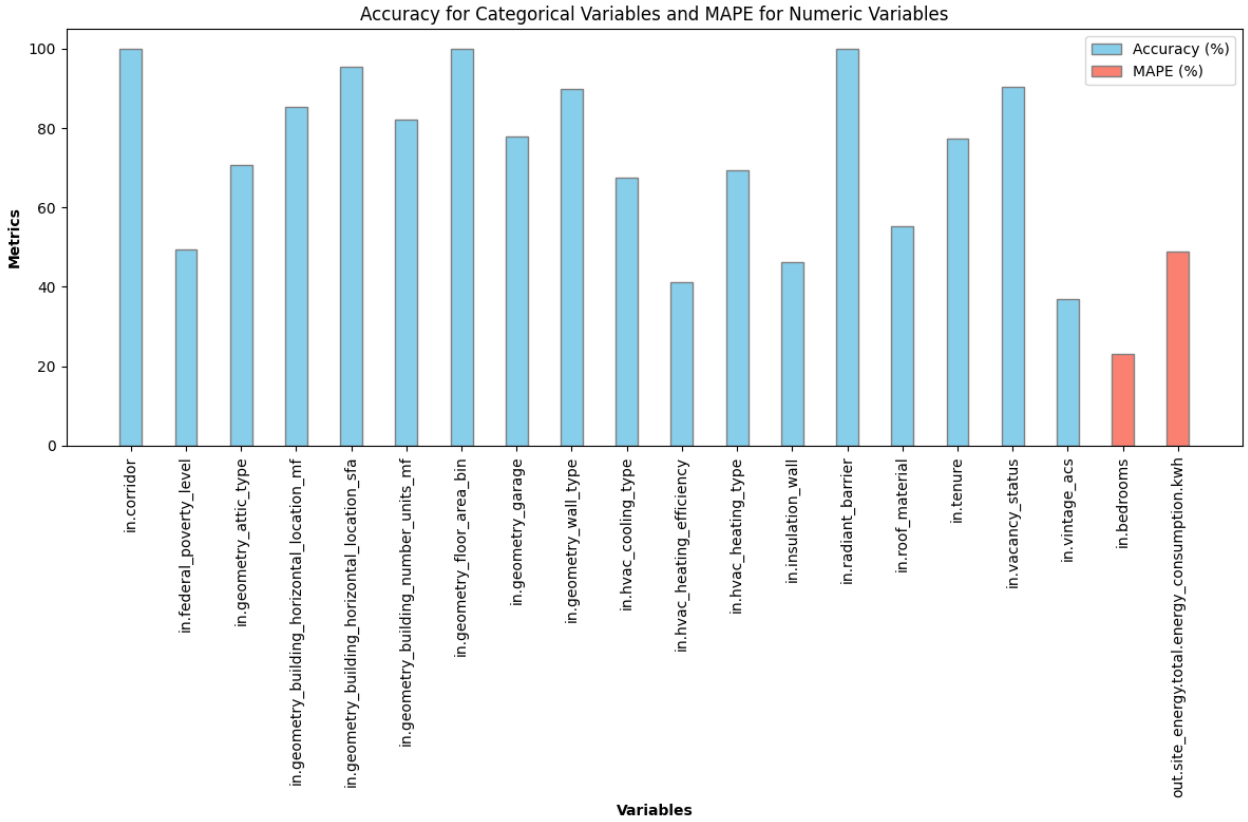

Figure 7. Method 1 validation metrics broken down by the predicted building characteristics.

This report is available at no cost from the National Renewable Energy Laboratory at www.nrel.gov/publications.

The results chart is divided into two sections comparing the accuracy of predictions for categorical variables (shown in blue) and the Mean Absolute Percentage Error (MAPE) for numeric variables (shown in red). The accuracy for categorical variables is consistently high, with most bars exceeding 60% accuracy. Some variables related to envelope construction even achieve more than 80% accuracy. This figure summarizes the performance of the developed method, indicating varying degrees of accuracy in predicting different characteristics and demonstrating relative reliability in predicting missing characteristics.

This method can be used by urban energy modelers who are focused on a specific district but have limited information about the buildings. Their goal is to predict certain characteristics required by the energy model, given a subset of known characteristics. In this use case, users might possess varying levels of known information about the buildings. To address this, we conducted an analysis where we incrementally added more known features to the prediction model and evaluated the model's performance with each addition. This approach helps us understand how the performance of the model is influenced by the level of information provided.

We constructed multiple iterations with varying levels of characteristics, starting with three known characteristics and adding one characteristic in each iteration until reaching nine known characteristics. The selection of these characteristics is based on the authors' expertise regarding which characteristics users are most likely to find from building surveys and requested information. The known characteristics are described in Table 1, and the results are illustrated in Figure 8.

Iteration 1	building_type, floor_area, geometry_stories
Iteration 2	building_type, floor_area, geometry_stories, heating fuel
Iteration 3	building_type, floor_area, geometry_stories, heating fuel, vintage_acs
Iteration 4	building_type, floor_area, geometry_stories, heating fuel, vintage_acs, occupants
Iteration 5	building_type, floor_area, geometry_stories, heating fuel, vintage_acs, occupants, heating_setpoint
Iteration 6	building_type, floor_area, geometry_stories, heating fuel, vintage_acs, occupants, heating_setpoint, infiltration
Iteration 7	building_type, floor_area, geometry_stories, heating fuel, vintage_acs, occupants, heating_setpoint, infiltration, median_income

Table 1. Iterations with different number of known characteristics.

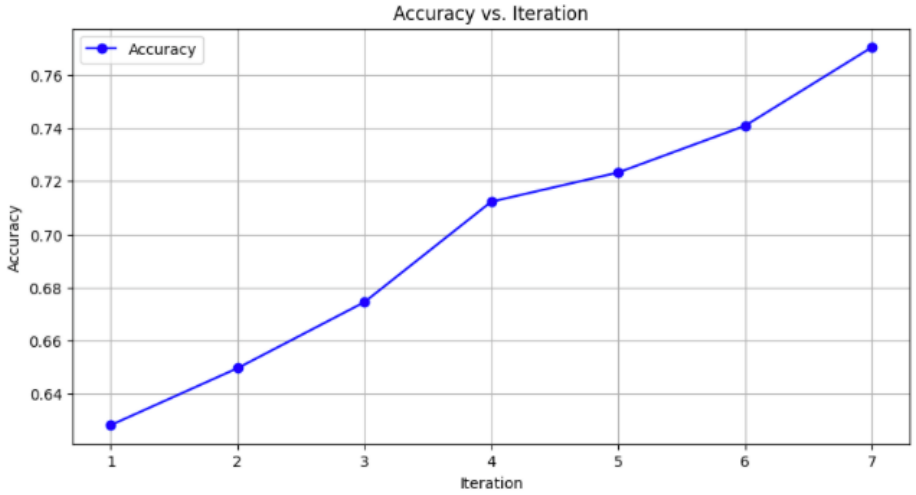

Figure 8. Accuracy vs Iteration.

This report is available at no cost from the National Renewable Energy Laboratory at www.nrel.gov/publications.

Initially, the model's accuracy starts at just above 65% and gradually increases to just above 80% by the seventh iteration, showing a positive trend. Table 1 lists the variables added at each iteration, starting with basic building attributes and progressively including more complex features such as heating setpoint and median income. The most notable increase in accuracy occurs at the 4rth iterations when we added the number of occupants to the known inputs. The graph indicates that as more relevant features are added, the model improves in its predictive capabilities, though the rate of improvement tends to diminish with each added variable. This trend may suggest an optimization in the selection of variables that balances accuracy with the efficiency and complexity of the model.

Artificial Neural Network method:

For the second method, we developed a neural network model. We used the same known inputs as in the first method (residential building type, floor area, number of stories, heating fuel, and vintage) and incorporated two hidden layers to model the relationships between these inputs and other unknown characteristics. All variables are transformed using the developed encoder that applies one-hot encoding for categorical variables and scaling for continuous variables. This model is tuned and trained using a random search to find the best hyperparameters and model structure. This model predicts the output layers, which are then reverse-encoded back to their original state. The validation framework is then employed to assess this model's performance.

Models' comparison:

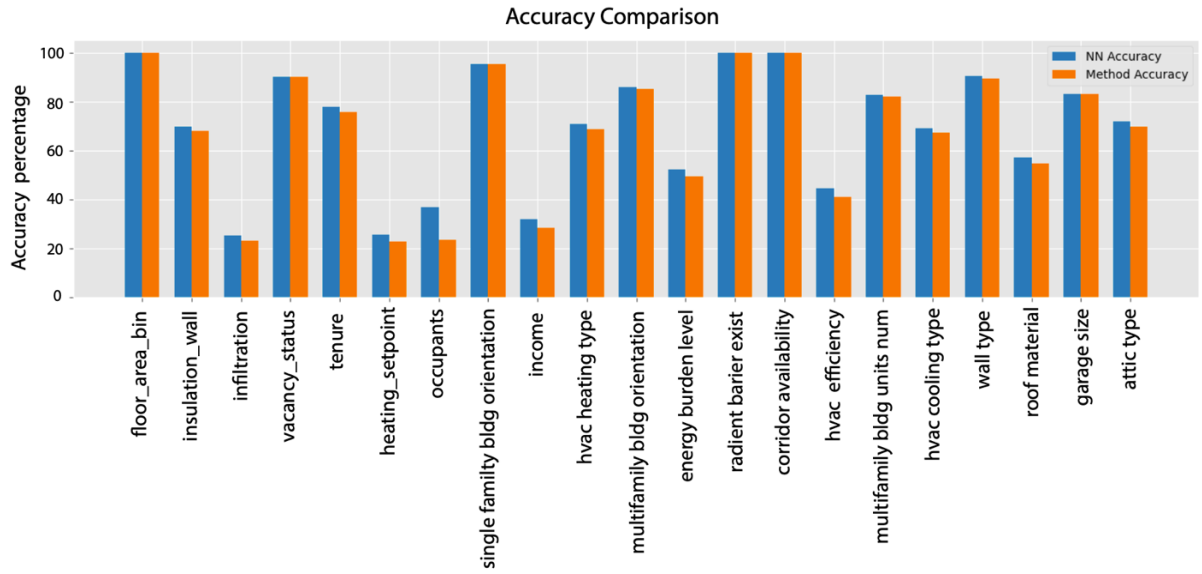

Figure 9. Comparison between method 1 and 2

Figure 9 presents an "Accuracy Comparison" between two predictive methods across various building characteristics. The accuracy is relatively similar for most characteristics, with the NN model exhibiting slightly higher accuracy overall. Notably, the NN model shows significantly higher accuracy in predicting the number of occupants. This figure demonstrates that the developed method achieves comparable performance to the NN model in predicting unknown

This report is available at no cost from the National Renewable Energy Laboratory at www.nrel.gov/publications.

characteristics, but with much less computational time. This advantage makes the KNN approach more generalizable, especially useful when users have varying levels of information for different buildings and wish to avoid constructing a new NN model as the number of inputs and outputs change. However, if the user prioritizes output accuracy and has the resources to support the computationally intensive NN model, then this model can be a viable option.

Discussion and Conclusion

Our framework integrates with the URBANopt district energy modeling platform, employing ResStock's statistical data to enhance energy modeling for unique neighborhoods. By addressing data gaps and disaggregating data to the neighborhood scale, our methods—using advanced machine learning and deep learning techniques—accurately map energy and economic properties onto building energy models.

Two models have been developed for predicting unknown building characteristics using known data. The first method uses a KNN + imputation approach with imputation techniques, while the second utilizes an NN model to learn relationships between multiple inputs and outputs. Both methods provide reliable predictions, which are valuable for energy modelers working with incomplete data. However, NNs, despite their ability to handle complex relationships, require significant computational resources for training and tuning. This becomes particularly challenging if input data changes, necessitating the development of a new model. In contrast, the KNN + imputation method is more adaptable and computationally efficient for varying information levels, making it advantageous in many scenarios. Conversely, when the level buildings data is uniform, the NN model is preferable as it shows better predications.

This effort leverages ResStock data, which encompasses a diverse set of building characteristics deemed crucial for equity-focused energy analytics. Characteristics related to occupant behavior and income are particularly significant and should be incorporated into modeling efforts. Our developed workflow enables energy modelers and analysts to consider these factors, which influence energy usage, in their analytical approaches. This integration starts with selecting relevant characteristics from the ResStock data and incorporating them into the model to assess their impact on energy consumption.

The developed energy model characterization framework achieves more accurate models that capture the variability of energy-related behaviors and properties across different buildings and neighborhoods. This leads to more precise energy analyses of neighborhoods, informing the deployment of targeted upgrades and efficient technologies. Consequently, it enhances energy efficiency in neighborhoods with unique economic and socio-demographic traits. Furthermore, our framework provides significant insights into equity by highlighting disparities in energy consumption and efficiency across different neighborhoods. By incorporating data on occupant behavior and income levels, the model identifies areas where residents may be disproportionately affected by high energy costs or where there is a greater need for energy-efficient technologies. This allows for more targeted interventions that can alleviate energy poverty and improve living conditions in underserved communities. Our methods contribute to a more equitable distribution of energy resources, ensuring that all neighborhoods benefit from advancements in energy efficiency.

Overall, this framework efficiently utilizes a variety of important building characteristics from ResStock datasets to develop comprehensive bottom-up energy models, offering a robust tool

This report is available at no cost from the National Renewable Energy Laboratory at www.nrel.gov/publications.

for sustainable and equitable urban energy solutions. By providing detailed insights into the energy dynamics of diverse neighborhoods, our approach supports the development of policies and initiatives that address the specific needs of different communities, ultimately fostering a more inclusive and resilient energy landscape.

Acknowledgement

This work was authored by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the U.S. Department of Energy Office of Science. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

References

U.S. energy consumption by source and sector, 2022, U.S. Energy Information Administration (EIA), 2023. https://www.eia.gov/energyexplained/us-energy-facts/.

Local Residential Energy Efficiency, (2023). https://www.epa.gov/statelocalenergy/local-residential-energy-efficiency .

M. González-Torres, L. Pérez-Lombard, J.F. Coronel, I.R. Maestre, A cross-country review on energy efficiency drivers, Applied Energy 289 (2021) 116681. https://doi.org/10.1016/j.apenergy.2021.116681.

International Energy Agency (IEA). Energy Efficiency: Buildings The global exchange for energy efficiency policies, data and analysis. 2019. https://www.iea .org/topics/energyefficiency/buildings/.

Kontokosta, C. E., Reina, V. J., & Bonczak, B. (2019). Energy Cost Burdens for Low-Income and Minority Households: Evidence From Energy Benchmarking and Audit Data in Five U.S. Cities. *Journal of the American Planning Association*, *86*(1), 89–105. https://doi.org/10.1080/01944363.2019.1647446

Kavgic M, Mavrogianni A, Mumovic D, Summerfield A, Stevanovic Z, Djurovic- Petrovic M. A review of bottom-up building stock models for energy consumption in the residential sector. Build Environ 2010;45:1683–97. https://doi.org/10. 1016/J.BUILDENV.2010.01.021.

Swan LG, Ugursal VI. Modeling of end-use energy consumption in the residential sector: a review of modeling techniques. Renew Sustain Energy Rev 2009;13:1819–35. https://doi.org/10.1016/j.rser.2008.09.033.

This report is available at no cost from the National Renewable Energy Laboratory at www.nrel.gov/publications.

Public ResStock Dataset : https://resstock.nrel.gov/datasets#largeee 2024.

Palani, Hevar, Juan Acosta-Sequeda, Aslihan Karatas, and Sybil Derrible. "The role of socio-demographic and economic characteristics on energy-related occupant behavior." *Journal of Building Engineering* 75 (2023): 106875.

Wilson, Eric J. *ResStock-Targeting Energy and Cost Savings for US Homes*. No. NREL/FS-5500-68653. National Renewable Energy Lab.(NREL), Golden, CO (United States), 2017.

Present, Elaina, Philip R. White, Chioke Harris, Rajendra Adhikari, Yingli Lou, Lixi Liu, Anthony Fontanini, Christopher Moreno, Joseph Robertson, and Jeff Maguire. *ResStock Dataset 2024.1 Documentation*. No. NREL/TP-5500-88109. National Renewable Energy Laboratory (NREL), Golden, CO (United States), 2024.

Polly, Ben, Chuck Kutscher, Dan Macumber, Marjorie Schott, Shanti Pless, Bill Livingood, and Otto Van Geet. "From zero energy buildings to zero energy districts." *Proceedings of the 2016 American council for an energy efficient economy summer study on energy efficiency in buildings, Pacific Grove, CA, USA* (2016): 21-26.

El Kontar, Rawad, Benjamin Polly, Tanushree Charan, Katherine Fleming, Nathan Moore, Nicholas Long, and David Goldwasser. *URBANopt: An open-source software development kit for community and urban district energy modeling*. No. NREL/CP-5500-76781. National Renewable Energy Lab.(NREL), Golden, CO (United States), 2020.

This report is available at no cost from the National Renewable Energy Laboratory at www.nrel.gov/publications.