Applied Energy 253 (2019) 113550

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

An integrated data-driven framework for urban energy use modeling (UEUM)

Narjes Abbasabadi[a,*], Mehdi Ashayeri[a], Rahman Azari[a], Brent Stephens[b], Mohammad Heidarinejad[b]

[a] Illinois Institute of Technology, College of Architecture, 3360 S. State Street, Chicago, IL 60616, United States
[b] Illinois Institute of Technology, Department of Civil, Architectural, and Environmental Engineering, 3201 S. Dearborn Street, Chicago, IL 60616, United States

HIGHLIGHTS

- Proposes an integrated data-driven framework for urban energy use modeling (UEUM).
- UEUM employs machine learning to model urban building and transportation energy.
- The framework is demonstrated using Chicago as a case study.
- Predicts multi-scale urban energy use with acceptable accuracy.
- Examines the relative contribution of urban socio-spatial factors on energy use.

ARTICLE INFO

Keywords:
Urban energy use modeling
Data-driven
Building operational energy
Transportation energy
Urban socio-spatial patterns
Machine learning

ABSTRACT

Many urban energy use modeling tools and methods have been developed to understand energy use in cities, but often have limitations in aggregating across multiple scales and end-uses, which adversely affects accuracy and utility. Increased data availability and developments in machine learning (ML) provide new possibilities for improving the accuracy and complexity of urban energy use models. This paper presents an integrated framework for urban energy use modeling (UEUM) that localizes energy performance data, considers urban socio-spatial context, and captures both urban building operational and transportation energy use through a bottom-up data-driven approach. The framework employs ML techniques for building operational energy use modeling at the urban scale with a travel demand model for transport energy use prediction. The framework is demonstrated using Chicago as a case study because it has significant variations in urban spatial patterns across its neighborhoods and it provides publicly available data that are essential for the framework. Results for Chicago suggest that, among the tested algorithms, k-nearest neighbor shows the best overall performance in terms of accuracy for a single-output model (i.e., for building or transportation energy use separately) and artificial neural network algorithm is the most accurate for the integrated model (i.e., building and transportation energy use combined). Exploratory analysis demonstrates that the urban attributes examined herein explain 41% and 96% of the variance in building and transportation energy use intensity, respectively. The UEUM framework has the potential to aid designers, planners, and policymakers in predicting urban energy use and evaluating robust theories and alternative scenarios for energy-driven planning and design.

1. Introduction

With increasing population growth and urbanization [1], and with greenhouse gas (GHG) emission reductions becoming a global priority, cities are pushing for sustainability now more than ever. Cities consume over two-thirds of primary energy resources and are responsible for more than 70% of GHG emissions worldwide [2]. The building sector is the typically single largest contributor to urban energy use and emissions, followed by the transportation sector. In the United States, the combined building and transportation sectors account for around 69% energy related emissions [3]. Moreover, building and transportation energy performance are interrelated at different levels [4–5] and interlinked with the urban spatial and socioeconomic context [6], which can greatly impact urban energy use and associated emissions. In this

* Corresponding author.
E-mail address: nabbasab@iit.edu (N. Abbasabadi).

https://doi.org/10.1016/j.apenergy.2019.113550
Received 25 February 2019; Received in revised form 23 June 2019; Accepted 13 July 2019

context, integrated energy models at the urban scale are needed to better inform designers, planners, and policymakers of current urban energy demand patterns and to provide stakeholders with tools to predict the energy and environmental impacts associated with different development scenarios. However, there are a limited number of methods and tools to do so, which often have limitations in providing a realistic representation of urban energy use and aggregating across multiple scales [7–9] and end-uses [4] crucial to urban environments.

The methods and tools for urban energy use modeling can be grouped as: (1) *top-down* models that apply econometric or technological approaches, use aggregated data, and generalize the status quo; and (2) *bottom-up* models that use data-driven or engineering physics-based techniques to examine urban energy use by understanding the behavior of its components and their inter-related interactions [10–11]. The latter is the most commonly used approach for urban energy use prediction [11]. However, shortcomings of engineering-based bottom-up studies, which apply simulation to predict urban energy use (and can be powerful for simulating individual buildings with independent heating, ventilation, and air-conditioning (HVAC) systems), stem from relying on inadequate archetypes that do not represent realistic variations of buildings across the city and simplification of context and system data for energy modeling at urban scale [12–14]. Conversely, data-driven based bottom-up methods rely on actual empirical energy data and can more accurately represent urban energy use, although the reliability of their results depends on the availability and quality of data and explanatory variables in the model [15–17]. The limitations of the existing literature mainly arise from applied aggregated data or generalized linear models which do not allow energy characterization at the individual building level.

In addition, common approaches to urban energy modeling have been limited in comprehensiveness towards defining, understanding, and estimating urban energy use. Urban energy use modeling is often reduced to its building operational energy component, often leaving transportation and/or key uncertainties in system interactions unaccounted for. This happens mainly because transportation energy modeling is associated with a high level of complexity and uncertainty due to the fact that a transportation energy model depends on multi-dimensional factors of land use, and technological and behavioral aspects [18–19]. There also exists a limited number of studies that predict urban energy performance through integrated analysis of networks of buildings in urban or neighborhood contexts. In many energy studies, the urban microclimate, as well as the interactions between individual buildings and the city, are often overlooked, although these factors have been shown to impact the accuracy of operational energy use estimations at both building and urban scales [20]. Urban morphological structure and building characteristics such as height impact urban microclimate and urban heat island effect [21–23], shading, and wind flow [24–25], which in turn impact urban building energy demand.

Also, previous studies tend to explore only a limited number of aspects that can affect urban energy demand. There is an extensive body of research that examines how urban energy use in cities is influenced by major urban attributes such as urban morphology and density [26–27], building characteristics [28–30], occupant features and behaviors [31–34], socioeconomic factors [35–38], and the impact of urban density on transport energy use [30,39] via effects on human mobility and travel modes and distances. However, there are limited numbers of studies that examine the importance of these various potential determinants of urban energy use and their relationships a comprehensive approach. Without a comprehensive approach, the results of these studies are often not reliable or generalizable for policy-making applications. The development of advanced data-driven artificial intelligence methods, specifically those based on machine learning (ML) methods, combined with advances in big data and open data initiatives adopted by most of the cities across the world, provide new possibilities for improving the accuracy and complexity of urban energy use models [40,41]. ML methods provide the opportunity to understand and manage urban energy use and to reveal the importance and complex behaviors of different urban socio-spatial energy determinants. However, their predictive and explanatory performance and accuracy have not been fully understood yet.

To address some of these limitations, this research proposes an integrated urban energy use modeling (UEUM) framework that localizes energy performance measurements, considers urban socio-spatial context, and captures both urban building operational and transportation energy use through a bottom-up data-driven approach. The framework employs ML techniques for building operational energy use modeling at the urban scale with a travel demand model for transport energy use prediction. It tests the most promising ML algorithms to achieve the most accurate model. The proposed framework views the city not as a collection of individual buildings, but rather as a network of connected buildings. It produces Geographic Information Systems (GIS) based visualizations that enable a realistic and multi-scale model representing building, block, neighborhood and city levels, to communicate and visualize the predicted urban building and transportation energy use. Chicago is used as a case study to test the framework because it has significant variations in urban spatial patterns across its neighborhoods and because it has committed to long-term emission reduction goals [42] and thus provides publicly available data that are essential for conducting an accurate urban energy analysis with UEUM. The framework is also applicable for other cities with similar datasets.

2. Methods

This section presents the UEUM framework as a four-step model, as shown in Fig. 1. The first phase is the *Pattern Extraction* phase, which studies urban socio-spatial patterns with a certain level of detail from available datasets to extract local variables and contextualize the model. The second phase is the *Prediction* phase, which estimates urban energy use data, including both building and transportation energy end-uses, using various machine learning models with numerous local variables. The third phase, the *Analysis* phase, uses the model results to explain the relative contribution of each variable on urban energy use. Finally, the *Visualization* phase generates visualizations of the predicted urban energy use across multiple levels of city, neighborhood, block and individual buildings.

As Fig. 2 illustrates, this multi-scale energy use analysis and visualization presents a more accurate and holistic image of urban energy use and the associated impacts of different spatial patterns, design decisions, and energy efficiency policies. It also contributes to a low-carbon urban transformation by identifying energy performance patterns and addressing potential problems and improvement strategies at the scales of buildings, blocks, neighborhoods, and entire cities. For example, a building-level analysis using disaggregated data provides a model of operational energy of buildings that can be used for energy efficiency management and retrofit purposes, while a census block level analysis can inform how existing or future alternative scenarios of design and planning such as changes in building height or density influence urban energy use. And neighborhood- and urban-scale modeling aims to improve understanding of urban energy use in a manner that can inform decision-making regarding urban morphological and spatial patterns such as zoning, land use, accessibility, and other planning policies that can affect the city structure and subsequent building operational and transportation energy end-uses. Further, moving from the block to neighborhood or urban scale allows for capturing inter-building effects on energy consumption [43] and for identifying patterns of energy use on a scale at which urban transformations mostly take place and where funding opportunities for energy efficiency investments are often available [44].

2.1. Urban energy use, definition and variables

In this research, urban energy use is defined as the combination of

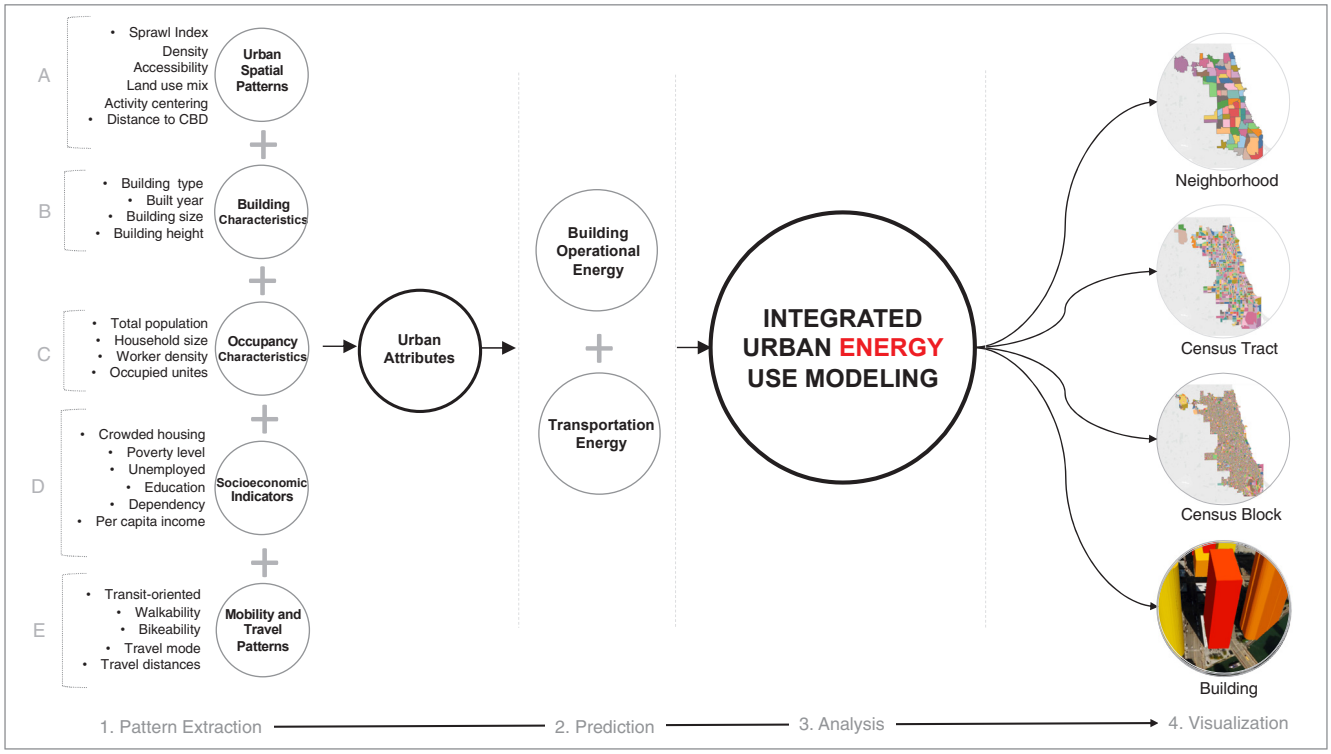

Fig. 1. Conceptual framework of the integrated data-driven UEUM.

operational building and transportation energy use at city scale. Urban transportation energy use is presented here as the amount of energy per household required on a daily and/or annual basis for various modes of travel including car and public transit. Urban building energy use is defined as the annual energy used for heating, cooling, water heating, appliances, electric plug loads, lighting, and all other building energy end-uses. Building operational energy use intensity (EUI), as the most common metric [45], is used to describe urban building energy use per unit of floor area (kBtu/ft^2 or kWh/m^2). However, it should be noted that the effectiveness of the EUI metric (kBtu/ft^2 or kWh/m^2) has limitations for measurement and benchmarking purposes, especially when occupancy and building type vary significantly [46]. An occupant-adjusted EUI metric (e.g., kBtu/ft^2/person-hour or kWh/m^2/person-hour) is considered to provide a more accurate assessment [47], but there are also limitations regarding occupancy data availability and accuracy that prevent usage of this metric.

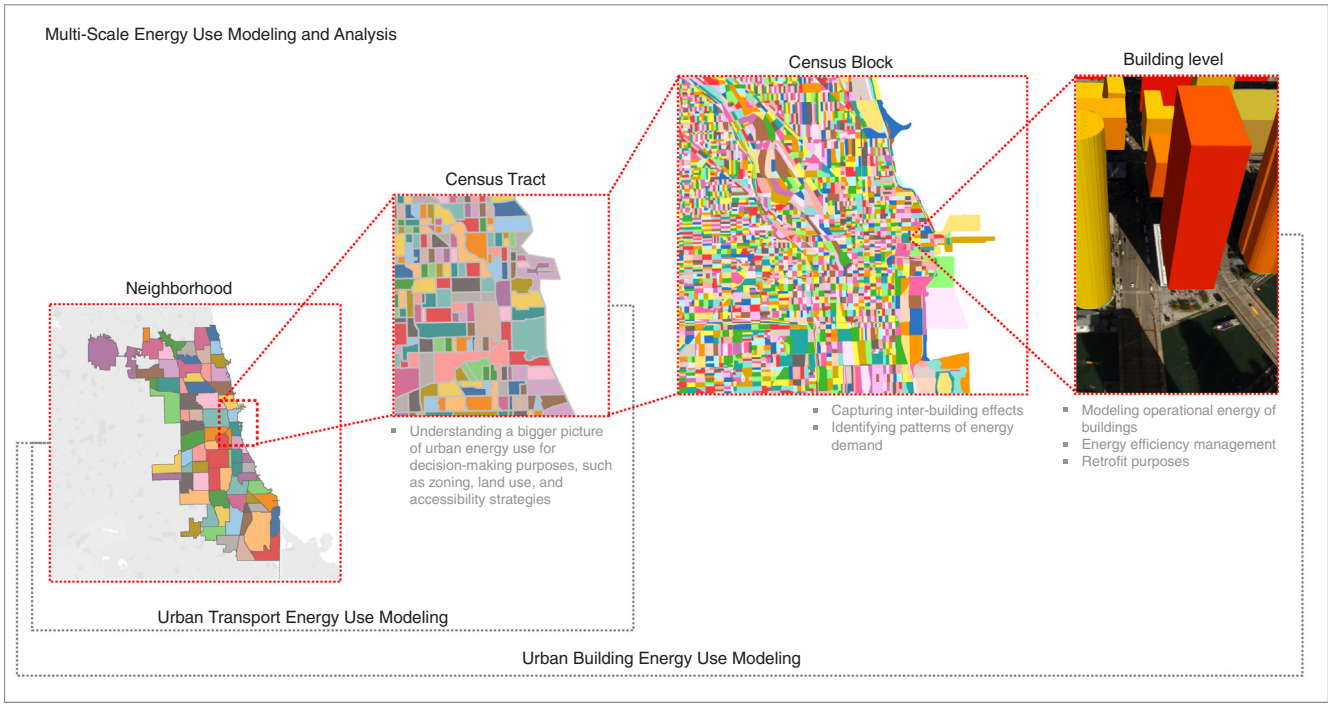

Fig. 2. Multiple scales of UEUM.

Table 1
Key determinants and variables incorporated in the urban energy model.

	Category	Variable	Unit
Independent Variables	Building Characteristics	Building height (number of floor)	Number of floors
		Building size (gross floor area)	ft^2 (m^2)
		Building type	–
		Year built	–
	Urban Attributes	Sprawl index (urban density, accessibility, connectivity, and land use mix)	–
		Distance to CBD	mi (km)
	Occupancy Characteristics	Household size	–
		Worker density	–
		Percentage of occupied units	–
		Weekly working hours	h
		Total number of occupants	–
	Socioeconomic Indicators	Unemployment (% of unemployed persons aged 16 years or older)	–
		Education (% of persons aged 25 years or older without a high school diploma)	–
		Income (per capita)	–
		Dependency (% of under 18-year-olds or over 64-year-olds population)	–
		Poverty (% of households below poverty level)	–
		Crowded housing (% of occupied units with more than one person per room)	–
	Mobility & Travel Patterns	Mode of travel	–
		Travel distance	mi (km)
		Transit-oriented	–
		Walkability	–
		Bikeability	–
Dependent Variables	Building Operational energy use	Site EUI	kBtu/ft^2/year (kWh/m^2/year)
	Transportation energy use	Transportation EUI (of household)	kBtu/HH/year (kWh/HH/year)

The urban building and transportation energy use determinants related to the scope of this research are classified into five main categories: (a) *Building Characteristics (BC),* consisting of variables such as building number of floors, building size, building type, and year built; (b) *Urban Attributes (UA)*, consisting of urban density, connectivity, accessibility and land use mix, which are presented through urban sprawl index [48] and distance to the Central Business District (CBD) representing the location; (c) *Occupancy Characteristics (OC)*, including household size, worker density, percentage of occupied units, weekly working hours, and total number of occupants; (d) *Socioeconomic Indicators (SI)* representing factors such as unemployment, education, dependency, income, poverty level, and crowded housing, and (e) *Mobility and Travel Patterns (MTP),* which include factors such as mode of travel and travel distance and neighborhood features representing transit-oriented, walkability, bikeability variables. Table 1 shows the list of key variables of interest in the model for urban building and transport energy use.

2.2. The UEUM workflow

Fig. 3 shows the UEUM workflow, which proceeds into the following steps:

2.2.1. Data preparation

First, data preparation was conducted, which includes coupling the data required for the framework and cleaning and processing the data. The UEUM database is built upon a merged Urban Attributes and Energy (UAE) dataset representing physical characteristics of buildings, urban socio-spatial patterns, building operational energy use data, and travel and mobility data. The major datasets used in UAE dataset contain geo-referenced data which allow for construction of an integrated dataset for energy use analysis at building-level. Data processing is the next step after locating data which includes several statistical techniques to clean data, handle missing data [49], test the model regarding normality [50], non-parametric analysis [51], multicollinearity [52], and feature selection [53]. These methods will be explained in more detail in Section 2.3, which describes the application of the framework to the city of Chicago. Both building and transportation EUI data are transformed to the natural logarithm, i.e., Log Site EUI (in kBtu/ft^2/year or kWh/m^2/year) and Log Transportation EUI (in kBtu/household (HH)/year or kWh/HH/year) to properly normalize the distributions and decrease the variability of data.

2.2.2. Pattern extraction

The available data is studied to detect and combine local properties representing contextualized urban patterns. The k-means clustering algorithm, which represents an efficient clustering method [54], was applied to define the typology (archetype) of buildings from the real urban context. Clustering enables grouping data into multiple classes in which the inner group similarity and inter-group dissimilarity of objects must be maximized. The maximization or optimization process can be performed using expectation maximization (EM) or k-fold algorithms. The k-means clustering allows grouping buildings with certain similarities for example, the building height typologies or different building sizes based on their energy consumption behavior [55]. The k-means has limitations such as it is sensitive to outliers in which determining the number of clusters may change the outcomes significantly. k-means is also a spatially-dependent algorithm in which rearrangement of data can affect the results for the same data.

2.2.3. Prediction

The prediction phase computes energy performance of urban buildings and transportation end-uses, which encompasses the model training, validation, comparison, and prediction. Several statistical and ML methods were tested and compared in terms of their energy predictive performance including Multiple Linear Regression (MLR), Nonlinear Regression (NLR), Classification and Regression Trees (C&RT), Random Decision Forest (RDF), k-Nearest Neighbors (k-NN), and Artificial Neural Networks (ANNs) which are among reliable and applicable data-driven methods [41,56–58]. For transportation energy prediction, a travel demand model is also developed as the bases for the transportation energy data.

In previous studies, the MLR model has commonly been used for both building and transportation energy use modeling because it is easily interpretable and computationally efficient [41]. The coefficient in MLR is interpreted as how a unit of change in an independent variable leads to changes in the unit of dependent variable. However, it has limitations to capture non-linear and complex patterns and is also sensitive to outliers [59,60]. This study tested the MLR algorithm as a commonly used method and for the purposes of comparing with several

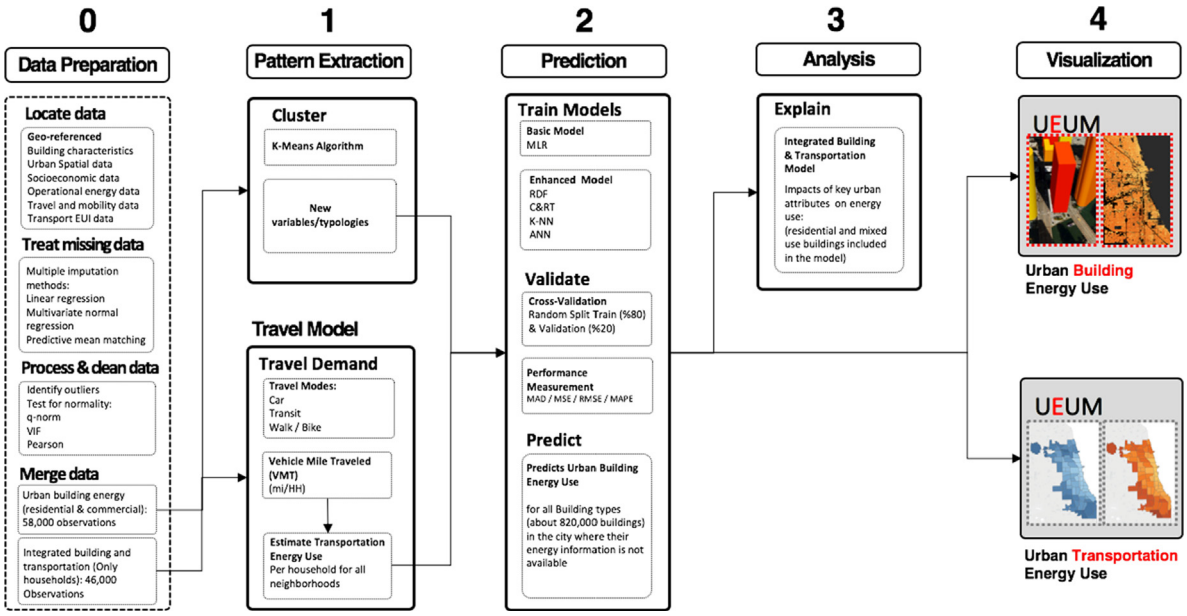

Fig. 3. UEUM workflow.

ML algorithms used for urban energy use modeling.

In this research, the ANNs [61] method, as a well-established ML algorithm, was tested and compared with other methods for effectiveness. ANNs is based on computational networks inspired from biological neural systems and calculates the tasks vaguely with no pre-assumptions about that system. ANNs generate and connect a network of input, hidden, and output nodes which has a remarkable capability in capturing complex nonlinear patterns between output and input variables [62]. Each neuron is made up of the summation function, activation function, bias, weights, inputs, and outputs. In this research, the Multi-Layer Perceptron of Artificial Neural Networks (MLP-ANN) model was used. The topology of the network comprises of 22 explanatory variables in including continuous, binary, and categorical neuron types with a single hidden layer, and a single node in the output layer. The activation function of the hidden layer and output layer are *tanh* and *identity*, respectively. In this article, an ensemble approach (51 networks) for training the model was implemented and the network with the highest performance was selected for the prediction.

k-NN [63] is another algorithm tested for effectiveness in this research. Similar to ANNs, k-NN [63] stands as a versatile ML technique that is used for solving both classification and regression problems on non-linear data. However, k-NN has been rarely applied for urban energy use modeling so far. The k-NN's learning process is implemented through memorizing the training data rather than using the discriminative functions. This characteristic enables k-NN to discover unseen data via the training dataset for the most similar k-neighbor instances, while it makes the model to be considered as a lazy learner among all ML methods. Meanwhile, k-NN needs a large number of instances to provide more accurate results. Another algorithm tested in this article is C&RT [64], which is one of the intuitive models that can be used for both classification and regression problems. CR&T's similarity to the human reasoning process enables them to act as white-box models since the training process can be easily visualized. C&RTs are classifier trees when the response variables can only take discrete values (categorical) and are regressor trees when the target parameters can take real numbers. C&RTs can also solve the missing value problems. RDF [65] algorithm was also tested, which is a flexible, high-speed execution, and easy-to-use machine learning regressor and classifier model. Indeed, RDF model builds up a set of decision-trees based on a random and independent selection of subsets of the data in growing the trees and then merge all of them together. The number of trees in this model was set as 100 and the number of variables for nodes are optimized in the train process. A distinctive difference between RDF and the traditional data-driven techniques is that RDF processes the data with making no prior assertions on the structure of the data or correlations between x and y variables. RDF is also less sensitive against spatial autocorrelation as well as multicollinearity problems.

2.2.4. Validation

Cross-validation (CV) [66], which has been shown to be an effective method [58,67], was applied to calculate the accuracy of the model prediction results. A 5-fold CV was used and the regularization parameter was tuned on the validation set to achieve the best prediction performance based on splitting data into train (80%) and test (20%). The standard predictive performance metrics including the Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and R-squared (R^2) were calculated as formulated in the following Eqs. (1)–(5) [58,68] to compare the models:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=0}^{n} |y_{predict,i} - y_{actual,i}| \tag{1}$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=0}^{n} (y_{predict,i} - y_{actual,i})^2 \tag{2}$$

$$\mathrm{RMSE} = \frac{1}{n}\sum_{i=0}^{n} \sqrt{\frac{\sum_{i=0}^{n} (y_{predict,i} - y_{actual,i})^2}{n}} \tag{3}$$

$$\mathrm{MAPE}(\%) = \frac{1}{n}\sum_{i=0}^{n} \left|\frac{y_{predict,i} - y_{actual,i}}{y_{actual,i}}\right| \times 100 \tag{4}$$

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=0}^{n} (y_{predict,i} - y_{actual,i})^2}{\sum_{i=0}^{n} (y_{actual,i} - \bar{y}_{actual})^2} \tag{5}$$

where $y_{actual,i}$ and $y_{predict,i}$ denote the actual and predicted EUI at ith observation, $\bar{y}_{actual}$ denotes the average EUI, and n denotes the entire number of observation within the dataset.

Finally, the models are compared, and the best model is proposed for a predictive model to compute the energy use for an output of sample model and an integrated explanatory model of building and transportation end-uses. The models were validated against actual energy data and compared against peer simulation and data-driven models from the literature.

2.2.5. Analysis

The analysis phase focuses on how key urban attributes affect urban energy use using the integrated model.

2.2.6. Visualization

GIS-based 2D and 3D visualizations were developed to communicate the results of urban building and transportation energy use modeling and analysis at multiple scales of individual building, block, neighborhood, and city scales. R software [69] was used for urban energy use statistical computing and graphics.

2.3. Application of the UEUM methodological framework to the City of Chicago

The UEUM framework was applied for Chicago, IL and its 77 communities (i.e., neighborhoods) that define the geographical divisions of Chicago. This study used communities (i.e. neighborhoods) as boundaries because the City of Chicago and Census data are mainly available for these community areas. The present research includes 820,606 buildings across different neighborhoods in the city which their building characteristics information is available in the current GIS-based dataset for Chicago [70].

2.4. Construction of urban attributes and energy (UAE) dataset

The UAE dataset for Chicago is constructed by merging several datasets as listed in Table 2 and explained in further detail into the following section:

2.4.1. Geo-Referenced building data

The publicly-available Chicago Building Footprint (CBF) dataset [70] was used as a geo-referenced building dataset which represents a compilation of building characteristics and geographic data for Chicago. The CBF dataset provides building-level data that contains the spatial characteristics such as building type, building number of floors, Gross Floor Area (GFA), year built, and location. However, this dataset does not contain data on other variables such as building renovation year and building Floor Area Ratio (FAR) which are reported in similar datasets for other cities such as the New York City's Primary Land Use Tax Lot Output (PLUTO) dataset [71]. This study also used zoning districts data [72] and property tax data [73] from the Assessor's Office for gaining further information such as property type and age. Despite of the CBF limitations, this dataset is the only and the most complete available dataset for building characteristic data in Chicago, and therefore, this research combined several datasets to complete the missing data across different datasets and build a dataset that would have data on both physical characteristics and operational energy use of buildings.

The merged dataset represents three building types including residential, commercial, and industrial. Residential buildings are represented by single-family, multi-family with less than seven units, and multi-family with seven or more units. Table 3 lists the type and frequency of buildings in the sample. Table 4 presents the descriptive information related to building characteristics and occupancy features for the buildings in the sample.

2.4.2. Geo-referenced urban attributes data

The sprawl index is used as an indicator of multiple aspects of urban attributes, including density, accessibility, and land use mix. The geo-referenced urban attributes dataset in this research includes the Urban Sprawl dataset [74] for the United States developed by Ewing [48,75] prepared for National Cancer Institute [74], which is based on the 2010 Census data and is available at four data levels: census tract, urbanized areas, Metropolitan Statistical Area (MSA), and county level. The sprawl index here represents: *(a) Density*, measured by several variables including population density of the census tracts area, the percent of the population living in low-density and medium- to high-density areas, the weighted density showing the density around the center of the Metropolitan Statistical Area (MSA), urban density within total built-upon land, and employment density; *(b) accessibility*, quantified by street network variables including average city block size, percent of urban blocks, average length of street blocks, the density of street intersections, and percent of different intersections defined as the street connectivity score; *(c) The land use mix*, captured via combining two variables including the job balance of total population and the job type mix of census block groups, and the Walkability score at census tract level; *(d) Activity centering*, quantified through the proportion of population and employment size in block groups, the ratio of amount of jobs and population of the CBD; and the speed of the population density declines from CBD [48]. In this research, the census tract level was used, as it is the smallest subdivision, approximately equivalent to a neighborhood, including a population of 2500–8000 [76].

As shown in Table 5, the sprawl index in Chicago neighborhoods varies from a minimum value of 87.46 to a maximum of 188.45, with a mean of 132.92 and standard deviation of 10.44. Larger values of sprawl index represent a higher degree of compactness and connectivity similar to those experienced in downtown areas. For example, Riverdale represents the most spread-out among Chicago neighborhoods with the lowest sprawl index (87.46) while the Loop represents the most compact and connected with the highest sprawl index (188.45). In this research, the mobility factors incorporated in the model represents neighborhood walkability, bikeability, and transit-oriented indices which are extracted from [77].

2.4.3. Socioeconomic indicators data

In order to provide a quantitative analysis of the relationship between socioeconomic patterns on two main components of urban energy use (building and transportation) at the same time, socioeconomic

Table 2
The main datasets used to construct the UAE dataset.

Type of data used in constructing the UAE dataset	Data source
Geographical location and building characteristics	• Chicago building footprints (CBF) dataset [70] • Chicago boundaries and zoning districts [72] • Property tax data from the Assessor's Office [73]
Sprawl index	• Urban Sprawl data for the Unites States [74]
Energy data	• Chicago Energy Benchmarking dataset [79] (2717 buildings greater than 50,000 ft^2) • Chicago Energy Usage dataset [80] (65,378 buildings of all sizes)
Transportation data	• Chicago Regional Household Travel Tracker Survey by CMAP [81] • Fuel Economy data [82] by the U.S. Department of Energy (DOE) • Average passenger transportation energy intensity per mile travel from the U.S. Department of Transportation (DOT) Bureau of Transportation Statistics (BTS) [83]
Mobility factors including neighborhood walkability, bikeability, and transit-oriented indices	• Walkability, bikeability, transit-oriented score [77]
Socioeconomic indicators	• Socioeconomic indicators dataset [78]

Table 3
Building type, sub-type and their frequency in the dataset.

Building Type	Building Subtype						
	Commercial	Municipal	Industrial	Multi 7+	Multi < 7	Single Family	Total
Residential	0	0	0	2192	19,213	25,506	46,911
Commercial	4864	154	0	1652	4609	0	11,279
Industrial	0	0	15	0	0	0	15
Total	4864	154	15	3844	23,822	25,506	58,205

Table 4
Building characteristics and occupancy features for the buildings in the dataset.

Variable	Observations	Mean	Standard Deviation	Min	Max
Building Height	58,205	1.87	2.20	1	110
Building Size (Gross Floor Area)	58,205	35,820 (ft^2) 3328 (m^2)	116,948 (ft^2) 10,865 (m^2)	300 (ft^2) 28 (m^2)	6,143,038 (ft^2) 570,707 (m^2)
Year Built	58,205	1935.27	31.81463	1852	2014
Total Occupants	58,205	83.90	84.65	0	3000
Average Household Size	58,205	2.34	1.39	0	9
Occupied Unit Percentage	58,205	87%	13%	0%	100%

Table 5
Sprawl index at the census tract level in Chicago.

Variable	Observations	Mean	Standard Deviation	Min	Max
Sprawl Index	58,205	132.92	10.44	87.46	188.45

indicators dataset [78] was used, which represents six socioeconomic factors of public health significance including income, employment, education, dependency, poverty level, and crowded housing level as discussed in Section 2.1. This dataset provides the data for Chicago communities for the years 2008–2012.

2.4.4. Building operational energy use data

The operational energy use dataset for the purpose of this research was constructed by merging of two unique datasets including Chicago Energy Benchmarking (2016) [79] and Chicago Energy Usage (2010) [80] datasets. The Chicago Energy Benchmarking dataset contains a building identifier number with an exact geographical location which allows for investigation at actual building-level. The data in this dataset is self-reported by building owners. The dataset contains annual energy utilization, energy star, and GHG emission data for buildings of different types. This dataset is reported in 2017 and includes energy utilization data for 2717 buildings in Chicago which represent about 1% of Chicago buildings and around 20% of building energy use in Chicago.

This study coupled the Chicago Energy Benchmarking dataset with the Chicago Energy Usage (2010) dataset, which provides energy use data for 65,378 buildings of all building sizes and various types including residential, commercial, and industrial buildings in Chicago. This dataset represents 68% of overall electrical usage and 81% of all gas consumption in Chicago in 2010. The Chicago Energy Usage (2010) data can complement other spatial and energy datasets as it contains many relevant variables including the census block, population, physical characteristics, floor area, average stories, average building age, and occupancy features. However, this dataset displays buildings at census block level, without providing a further geographic identifier. The present study used this dataset because it provides data for buildings smaller than 50,000 ft^2, which are not included in the Chicago Energy Benchmarking datasets. The merged building operational energy use dataset represents all building sizes in Chicago and is large enough to be used for developing representative statistical models.

Site EUI (expressed in kBtu/ft^2/year or kWh/m^2/year) is used as the main metric for energy use in this research because it allows for comparison of buildings across various neighborhoods and across energy datasets. Table 6 tabulates the Site EUI values for the building cases in the UAE dataset. The histograms in Fig. 4 illustrate the distribution of site EUI and Log site EUI. The figures show the site EUI data are approximately lognormally distributed. Thus, building EUI data are transformed to the natural logarithm, i.e., Log Site EUI, to properly normalize the distributions and decrease the variability of data.

2.4.5. Travel demand (TD) and transportation energy use (TEU) data

The Urban Transportation Energy (UTE) model is developed using a variety of inputs including travel demand information and transportation EUI data for various modes of travel. We use the latest Chicago Regional Household Travel Tracker Survey [81] for travel demand modeling. This dataset is a detailed travel and activity survey conducted by the Chicago Metropolitan Agency for Planning (CMAP) in 2007–08. The dataset includes daily travel information conducted in either 1-day or 2-day surveys for 23,808 individuals who resided in 10,552 households in the northeastern Illinois region (e.g., Cook, Lake, DuPage, Kane, Kendall, Grundy, McHenry, and Will Counties). The surveys have been weighted in 2016 by CMAP to remain consistent with Census estimates and represent the travel made by the households and the population in the region. The weighted surveys included the supplemental survey weights and corrected distance traveled values.

The dataset includes required travel details such as mode choice, trip purpose, and distance, as well as demographic information. The detailed data include mixed-mode trips with car, bus, subway, and rail transit. It also provides manufacturing information about the vehicles (e.g., year, make, model). The distance calculations are based on the actual location information provided in the survey responses. However,

Table 6
Summary statistics of building site EUI used in the model.

Variable	Observations	Mean	Standard Deviation	Min	Max
Building site EUI	58,205	67.29 (kBtu/ft^2) 212.28 (kWh/m^2)	30.01 (kBtu/ft^2) 94.68 (kWh/m^2)	10.65 (kBtu/ft^2) 33.60 (kWh/m^2)	540.00 (kBtu/ft^2) 1703.48 (kWh/m^2)

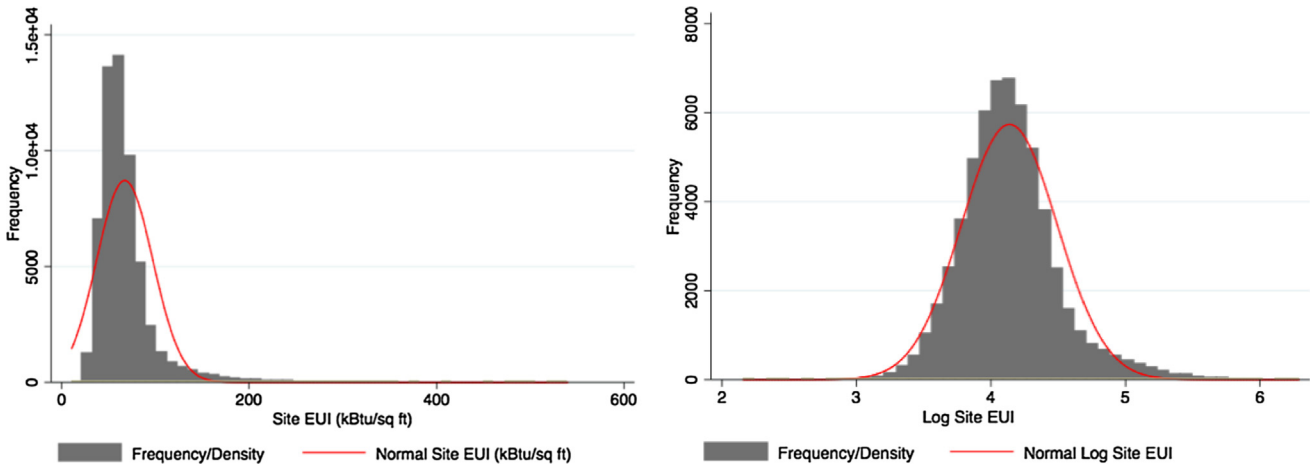

Fig. 4. Histograms of site EUI and Log site EUI (i.e., natural logarithm of site EUI) overlaid with a normal distribution curve.

in the public dataset the location data is anonymized by the centroid of the resident census tract. The distance value represents the direct/straight distance between the origin and destination locations (not a street network-based distance) for each trip segments, measured in miles. The total value of distance traveled by each household is estimated by summation of all single or mixed-mode trips per trip mode for each person in the household.

The mode-based transportation energy use per household for light-duty vehicles and private cars was estimated via the Average Miles Per Gallon (MPG) for automobiles from Fuel Economy dataset reported by the U.S. DOE [82], and for public transit through using average passenger transportation energy intensity per mile travel from the U.S. DOT Bureau of Transportation Statistics (BTS) which provides the annual average transport energy intensity per passenger through the National Transportation Statistics database [83]. This approach uses annual statistics, such as fuel or electricity which is the most available public transportation EUI factor calculated on an annual gross average basis, to estimate the transport energy use per passenger-mile. Table 7 and Fig. 5 show summary statistics and distributions of the estimated transportation energy use in Chicago in kBtu/household (HH)/year or kWh/HH/year. Again, the transportation EUI data are transformed to the natural logarithm, i.e., Log Transportation EUI, to be consistent with building site EUI.

2.4.6. Data cleaning and processing

2.4.6.1. Process & clean data. The available building, urban, and energy datasets contain considerable amount of errors, missing values, and outliers as extreme and beyond the normal range of observations with respect to independent or dependent variables. In the first step, three groups of errors in the dataset were visually identified and dropped from the dataset: (a) cases with multiple ID numbers; (b) cases with zero or null values; and (c) cases with multiple entries. In the next phase, the potential extreme outlier cases were detected through statistical tests. Then the influence of those potential outliers was evaluated on individual regression factors and the outliers with substantial effect were removed.

In order to identify potential outliers, residuals and standardized residuals were predicted and the residuals beyond the range of -2 and $+2$, as the threshold, were considered as indicators for likely outliers [84]. It should be noted, however, that $\pm$ 2 threshold was not treated as an absolute threshold. The Cook's Distance test [85] was used to assess the influence of the outliers on overall regression results. The Cook's Distance tests the influence of a given case on all the fitted values. A high Cook's Distance reflects high residual and its leverage. The Cook Distance value of less than 1 indicates that the influence is not significant [85] and the cases with such values were not dropped from the dataset, even though they may have been identified as outliers in previous steps. Also, the influences of individual cases on individual regression coefficients were assessed through the DFBETAS test using Eq. (6) [86]:

$$|dfb| > \sqrt{N} \tag{6}$$

where N denotes number of observations.

The final UAE dataset was constructed by merging several datasets including CBF dataset, urban attributes dataset, Chicago Energy Benchmarking dataset, Chicago Energy Usage dataset, travel demand, and the estimated transportation energy use dataset. The merged UAE dataset contained data on 58,205 building cases of all types, residential and non-residential, in which data for transportation model was available at household level, leaving 46,843 observations in the

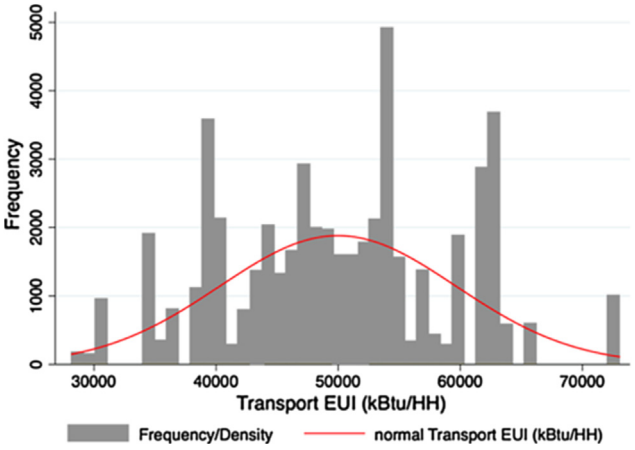

Fig. 5. Histogram of transportation EUI (per household) with a normal distribution curve overlaid.

Table 7
Summary statistics of estimated Transportation EUI (per household) in kBtu/HH/year (kWh/HH).

Variable	Observations	Mean	Standard Deviation	Min	Max
Transportation EUI	46,845	49,975 (170,516)	9684 (33,041)	28,171 (96,121)	73,035 (249,194)

transportation model dataset. It should be noted that the travel distances of two standard deviations higher than the mean were eliminated from the dataset which represent less frequent long-distance trips.

2.4.6.2. Treat missing data. We tested several statistical methods to treat missing data and maximize the available information, including Multiple Imputation (MI) with predictive mean matching algorithm [49]. This method analyzes data and replaces missing values according to valid frequency inference. For example, the CBF dataset has an extensive amount of missing data such as 'number of floors' information. To treat this missing piece of information, the predictive mean matching method was applied to 'guess' the missing values by considering available information on several variables including location which provide (x,y) coordination, neighborhood, building footprint, Gross Floor Area, year built, building type, and number of units.

2.4.6.3. Model normality test. The model normality test was performed to evaluate the model in terms of normal distribution and the shape of residuals and distribution of errors. The main limitations of regression models stem from their performance dependency on explanatory variables, and size and consistency of training data. The model accuracy is affected by insufficient or mis-specified core independent variables in the prediction or training model. Otherwise, it may lead to over-fitting or generalization beyond training range. We used Normal Quantile–Quantile (Q-Q) plot for detecting normal distribution which is an effective and mostly used normality test [50]. Fig. 6 illustrates the results of the normal Q-Q test on our constructed merged Urban Spatial and Energy (USE) dataset for building EUI (Fig. 6a) and transportation EUI (Fig. 6b). A slight deviation from normality is observed in the tail. We interpret the model in this research as a model that its normal distribution is still acceptable however not as a perfectly distributed. As explained in previous sections, ± 2 were not treated as an absolute outlier threshold and the Cook's Distance test was used to assess the influence of the outliers on overall regression results. Those observations were kept in the model mainly because excluding them meant eliminating some classes of building height (98-, 100, and 110-story buildings) that were important to the objectives of this research.

Table 8
Selected variables in the model and the variance inflation factor (VIF) values.

	Original Variable Name	Abbreviation	VIF
X1	Sprawl index	SPI	1.64
X2	Multifamily building with less than 7 units	MFL7	binary
X3	Multifamily building with more than 7 units	MFG7	binary
X4	Single family house	SF	binary
X5	Height	HGT	1.71
X6	GFA	GFA	1.75
X7	Year built	YRB	1.21
X8	Total occupants	TOC	7.15
X9	Total units	TUT	7.04
X10	Percentage of occupied units	POU	1.27
X11	Average household size	AHS	1.72
X12	Per capita income	PCI	7.01
X13	Percentage aged over 16 unemployed	P16U	5.99
X14	Percentage of aged over 25 without high school diploma	P25D	9.97
X15	Percentage aged under 18 or over 64	P1864	7.68
X16	Percentage of households below poverty	PHBP	5.11
X17	Percentage of housing crowded	PHC	7.36
X18	Daily VMT	VMT	2.40
X19	Distance to CBD	CBD	4.87
X20	Transit-oriented score	TOS	6.32
X21	Bikeability score	BIS	4.77
X22	Walkability sore	WKS	5.18

2.4.6.4. Non-parametric analysis. The non-parametric analysis which assumes a non-normal distribution was also performed to test the hypothesis for the equality of the building site EUI distribution across all building typologies. The Kruskal-Wallis H test [51], a multiple-sample type of Wilcoxon test [87], was conducted to determine if the building site EUI was significantly different across for example, building height typologies. Kruskal-Wallis H test revealed statistically significant variance in site EUI (kBtu/ft^2) across height typologies. Chi-squared for the model with all building types was shown to be 8799.672

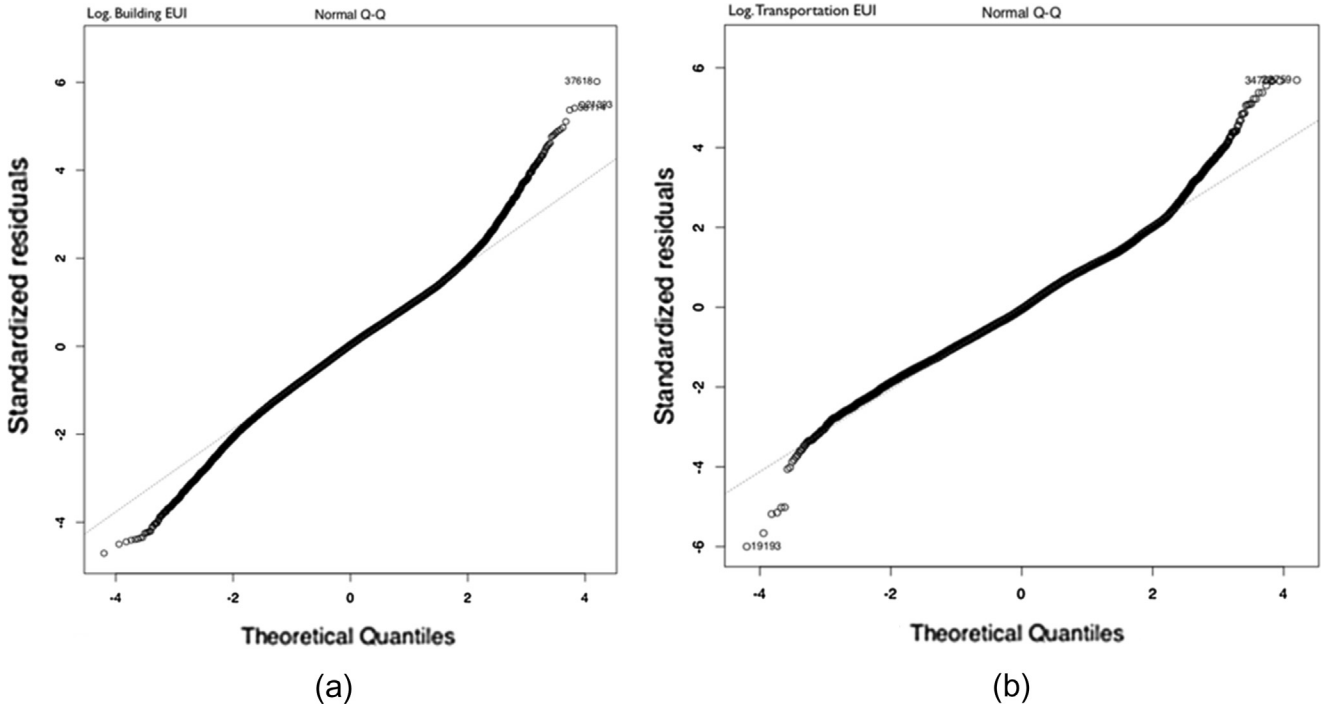

Fig. 6. Normal Q-Q plot: The results of Normal Q-Q test for (a) building EUI and (b) transportation EUI of our constructed merged Urban Spatial and Energy (UAE) dataset.

(p = 0.0001). Chi-squared for the model with only residential mix-use buildings was shown to be 10487.339 (p = 0.0001). The results suggest that in both the model for all building types and the model for residential mix-use buildings, there is significant difference in site EUI across height typologies.

2.4.6.5. Feature selection. The Variance Inflation Factor (VIF) [88] as a feature selection technique was applied to test the model regarding multicollinearity and to select variables in the model. Table 8 shows the results of VIF. The VIF represents the variance inflation of variable and 1/VIF represents tolerance as VIF value greater than 10 or 1/VIF values lesser than 0.1 may need additional analysis. The results of multicollinearity test in which all VIF values are under 10 and acceptable to be used in the model.

2.5. Urban energy use modeling

The predictive model is first implemented to estimate urban building energy use for both residential and non-residential buildings for which their energy consumption is not available. The model is applied using the merged data of 58,205 observations to predict the building EUIs for 820,606 out-of-sample buildings of all types in Chicago whose building characteristics are available in the CBF GIS based-dataset but whose energy data are not available. It also develops a travel demand model to estimate transportation energy use per household. Next, an integrated building and transportation model is developed to model the two main components of urban energy use simultaneously and to explore the contributions of each variable in the model, including key urban attributes such as building characteristics, urban spatial pattern, occupancy characteristics, and mobility and travel patterns. Since data for the urban transportation energy model is available only at the household level, and residential and non-residential buildings show different energy behaviors, the predictive and explanatory models are outlined separately. Fig. 7 presents the workflow of the data-driven urban energy use prediction model.

2.5.1. Urban building energy use modeling

The urban building energy use modeling is executed to predict energy use values using regression function as a supervised problem from response variables as discussed above. The urban building operational EUI, UBEUI, is predicted using Eq. (7):

$$UBEUI_i = f(\beta_1 BC_i, \beta_2 UA_i, \beta_3 \ OC_i) \tag{7}$$

where *BC*, *UA*, and *OC*, denote Building Characteristics (including building height, building type, gross floor area, and year built), Urban Attributes (including sprawl index), and Occupancy Characteristics (including average household size, worker density, operating hours, total number of occupants, total number of units, percentage of occupied units).

The machine learning methodology is first used to identify the relationships between the independent variables, including building characteristics, urban attributes, and occupancy characteristics, and the dependent variable, including site EUI, as trained on the UAE dataset. The patterns of mathematical relationships of the trained dataset are then used to predict energy use for building cases out of sample for which energy data are not available. The accuracy of the model could be increased by including more related independent variables. However, variables that are not well disaggregated may increase the margins of error in the model. In this framework, occupant behavior-related variables are excluded due to the lack of data availability.

2.5.2. Travel demand modeling and transportation energy use prediction

The transportation energy prediction model was developed in two steps: (a) travel demand modeling and (b) transportation energy prediction. The mileage traveled by households was determined by summing all segmented trips from origin to destination based on different transit-based modes including auto, bus, subway, rail commuter for each person in the household. Once the travel mile per person for each journey was calculated, the daily travel mileage of each household, each census tract level, and each neighborhood were totaled. For the two-day surveys, travel miles were summed and the average per day was determined. In the weighted dataset for the single day surveys, an equal value was applied while for the two-day surveys if both days were weekdays then the survey weights were with a value equal to 0.5. For two-day surveys including only one weekday, the weekend day information was excluded from the analysis and the weekday data was applied a weight value equal to one. The total TM values per household is formulated in the Equation (8):

$$TM_{hh} = \sum_{n=1}^{\infty} \sum_{i=1}^{\infty} \mathrm{TMi} \tag{8}$$

where TM_{hh}, n, i, and TMi denote daily Travel Miles per household, number of individuals in household, mode of travel, and daily travel miles per individual per mode of travel, respectively.

Transportation EUI was then estimated for different modes of travel across various neighborhoods. The urban transportation EUI is predicted through incorporating miles traveled, fuel economy of different modes of travel, and energy intensity factors per mode of travel. This method which has been employed by previous studies yields an effective transport energy estimation [27,89]. We adopted this method to estimate urban transportation EUI, UTEUI, for each neighborhood in kBtu/HH using the Equation (9):

$$UTEUI_i = \sum_{n=1}^{\infty} \mathrm{TM} * \mathrm{TEUI} \tag{9}$$

where *UTEUI*, TM and TEUI denote the Urban Transportation EUI per household, Travel Miles (for different modes), and Transportation Energy Intensity per mile of travel, respectively.

Once the daily transportation energy use per person for each mode was determined, the daily transportation energy use of each household was estimated. The daily transportation energy use per household for each census tract was estimated by summing all the total value of transportation energy of households and dividing by number of sampled households in the census tract. In the next level of analysis and mapping, the average values of the transportation energy use per household centroid of census tract were calculated and assigned to the associated neighborhoods as the median transport energy use of each neighborhood. Daily energy use values were scaled to annual use with a factor of 261 because of the exclusion of travels on weekends. The energy use of non-motorized modes of travel including walking and biking was assumed to be 0.

2.5.3. Integrated building and transportation energy model

Next, the building energy use model is integrated with the transportation energy model to provide a more comprehensive model of urban energy use that includes both significant energy use contributors in cities. We tested several machine learning algorithms to identify the best performing model(s) that enable integrating the two energy end-uses and that capture the complex and non-linear relationships between urban key attributes and urban building and transportation energy dynamics across neighborhoods in the city simultaneously. For example, the ANN model is a robust technique that enables capturing non-linear patterns. However, it has limitations in terms of interpretability and quantifying the relative contribution of input variable on response variable, which makes ANN a black-box model [90]. There are methods that can be extended to ANN to add the explanatory capabilities of this algorithm and allow for quantifying the relative contributions of each variable in the model. Gevrey et al [91] provide a comprehensive review on explanatory capacities of ANN models. A well-known review article [91] discusses several methods for explaining the contribution of independent variables on dependent variable in an ANN model including: PaD (Partial Derivatives/Dependence), Weights, Profile, Classical stepwise, and Improved stepwise. We used the Weights method,

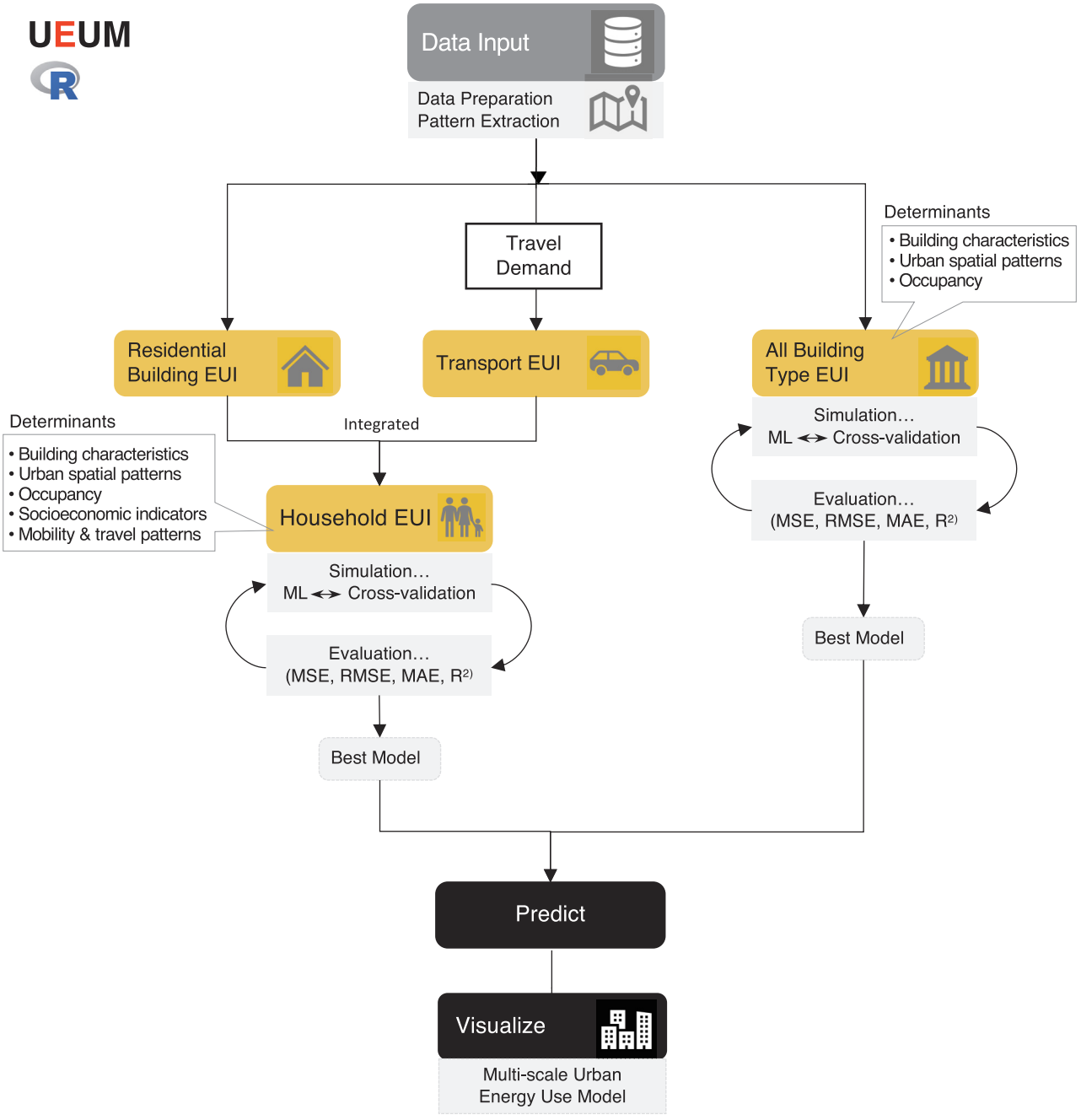

Fig. 7. The workflow of the data-driven urban energy use modeling.

which applies the connection weights. Garson algorithm [92] and Olden algorithm [93] are two popular algorithms for importance variable based on the Weights method. Fischer [94] suggests that Garson's model is preferable to the Olden's method. In this research, we employed the Garson algorithm for explaining the relative contribution of each variable on urban building and transportation energy use.

3. Results and discussion

This section presents the results for (1) the urban building energy use model and (2) the integrated building and transportation energy use model, separately.

3.1. Urban building energy use model

This section presents results of model effectiveness and energy prediction of the data-driven urban building energy use model for Chicago at multiple scales. The models, including the MLR, NLR, RDF, C &RT, k-NN, and ANNs algorithms, were tested across the merged dataset of 58,205 observations to identify the best model for predicting building EUIs for the 820,606 out-of-sample buildings in Chicago for which energy use data is not available. The models were executed for predicting the EUI of all building types, including both residential and non-residential buildings. The models incorporate three major energy use factors: building characteristics (building height, type, size, and age); urban attributes representing building location and sprawl dimensions; and occupancy characteristics (total number of occupants, household size, worker density, weekly working hours, and percentage

Table 9
Performance evaluation of the predictive models for urban building energy use in Chicago.

	MLR	NLR	RDF	C&RT	k-NN	ANN
R2	0.28	0.29	0.33	0.42	0.75	0.33
MAE	0.22	0.21	0.21	0.20	0.08	0.21
MSE	0.09	0.09	0.09	0.08	0.04	0.08
RMSE	0.31	0.30	0.29	0.28	0.19	0.29
MAPE (%)	5.35	5.14	5.14	4.87	1.83	5.16

of occupied units). Table 9 presents the predictive performance assessment of the models reporting on the MAE, MSE, RMSE, and MAPE performance metrics as defined based on measuring the errors between the predicted and actual values, as discussed in the methodology section. The lower the values of these metrics the better predictive performance of the model. The results suggest that k-NN and MLR represent the overall best and weakest predictive performance compared to the other models tested here, respectively. The C&RT, ANN, RDF, and NLR models were the next best performing models after k-NN, respectively. As shown in Table 9, k-NN model shows a MAPE of 1.83% while C&RT shows a MAPE of 4.87%, ANN, RDF, and NLR showed a MAPE of 5.14%, and MLR presented a MAPE of 5.35%. The results indicate that k-NN is able to decrease the error by 62%, relative to C&RT and 64%, relative to ANN, RDF, and NLR. The results of this study indicate that using k-NN model compared with MLR, which is the most commonly utilized method in previous studies for data-driven urban energy prediction, enhances the accuracy of the model, decreasing MAPE by 66%. For example, in our MLR model, the resulting R^2 value of 0.28 indicates that the model explains 28% of the variance in building operational EUI. Conversely, our results show that among the tested models, k-NN provides a significantly improved model with R^2 of 0.75, as calculated based on actual vs. predicted energy use values. In other words, k-NN is able to explain 75% of the variation in building energy use in the model.

The k-NN model has been rarely applied in previous studies for urban energy use prediction. The few previous studies that applied k-NN are mainly time-series models. For example, two prior studies [95,96] confirm the enhanced accuracy of k-NN using time-series data for electricity demand forecasting compared with other conventional statistical models. In a peer data-driven study [97], an energy model was developed for New York City using New York Energy Benchmarking. This research tests and compares the predictive performance of the linear regression with OLS, RDF, and SVM algorithms to predict annual building EUI for remaining residential and commercial buildings in New York City at building and zip code levels. This study reports error based on MAE ranged from 0.4 to 1.48 for different algorithms tested in their model. The results of this study suggest that SVM provides the lowest MAE for energy prediction within the sample and OLS provides better performance relative to SVM and RDF when generalizing to a city level energy prediction. In another previous study [98], MLP-ANN was used in improving the energy consumption benchmarking for 7700 schools in the UK and reported 20.6 to 22% error.

The peer simulation models were selected based on error rates reported by previous studies. Reinhart and Davila [9] provide a review on several urban building energy use modeling and they reported error range of 4% to 69% based on total building EUI as simulation output (4–18% at aggregate scale validation and 5–69% at building scale validation). In another study, Fonseca and Schlueter [44] report 6–66% error for building energy use modeling in district context using simulation methods hybrid with data-driven model using k-means algorithm. Nutkiewicz et al. [14] propose a framework (DUE-S) in which simulation is coupled with a Convolutional Neural Network (CNN) model for predicting urban scale electricity-use and reported 8.28–49.6% error. The framework was tested for 22 campus buildings in California. It should be noted that the error rates for the simulation-based models are reported based on monthly, daily, and hourly temporal granularity. Tables 10 and 11 present comparisons of the prediction performance of multiple peer data-driven and simulation-based urban building energy use modeling.

Table 10
Reported data-driven error range as an accuracy metric for multiple previous studies of urban building energy use.

Source	Model	Accuracy Metrics		
		MAPE (%)	RMSE	MAD/MAE
[97]	MLR (OLS)	–	–	0.41–1.48
	RDF	–	–	0.40–1.32
	SVM	–	–	0.40–1.26
[98]	ANN (MLP)	20.6–22%	–	–
[96]	k-NN	1.81–2.38%	–	–

Table 11
Reported simulation-based error range as accuracy metric for multiple previous studies of urban building energy use.

Source	Model	Reported Error Range %
[9]	Simulation	4–69%
[44]	Simulation + k-means	6–66%
[14]	Simulation + CNN	8.28–49.6%

In sum, the result of the data-driven UEUM building energy model was validated and compared against actual and peer data-driven and simulation-based models. The results of the UEUM building energy model, developed by this research, as summarized in Table 9, show an acceptable error rate compared to the previous studies. The results suggest k-NN as the best overall accurate and reliable model compared with other algorithm tested in this research is able to predict building energy consumption at urban scale. The results also indicate that applying disaggregated data, including the local features of urban context along with using more advanced data-driven machine learning models rather than the generalized linear methods can fill the prediction performance gap significantly.

Fig. 8 illustrates a map of energy use prediction results for all buildings in Chicago with individual building level resolution using k-NN. Fig. 8a illustrates the energy benchmarking data for 2717 buildings, which are the only available energy data at building level, and Fig. 8b depicts the predicted energy for the 820,606 buildings in the city of Chicago. The median predicted building site EUI for all building types was found to be 71.88 kBtu/ft^2 (226.75 kWh/m^2), including 68.15 kBtu/ft^2 (214.98 kWh/m^2) for residential and mixed-use buildings and 95.01 kBtu/ft^2 (299.73 kWh/m^2) for commercial buildings. The predicted site EUI for out-of-sample buildings was compared with the data in the sample, i.e., the dataset merging Chicago Energy Usage dataset and Energy Benchmarking (with 58,205 observations including buildings of all sizes) and the Energy Benchmarking dataset. The median building site EUI in the 58,205-building sample of all building types is 61.14 kBtu/ft^2 (192.87 kWh/m^2), including 60.24 kBtu/ft^2 (190.03 kWh/m^2) for residential and mixed-use buildings and 68.02 kBtu/ft^2 (214.57 kWh/m^2) for commercial buildings. According to the Chicago Energy Benchmarking dataset, which reports energy use for buildings greater than 50,000 ft^2 (4645 m^2), the median building site EUI for all building types is 79.40 kBtu/ft^2 (250.47 kWh/m^2), including 75.95 kBtu/ft^2 (239.59 kWh/m^2) for residential and mixed-use buildings and 83.65 kBtu/ft^2 (263.88 kWh/m^2) for commercial buildings.

The UEUM predictive model as a tool has two specific applications. The first application is to quantify energy use of buildings at an urban scale with a building level resolution to assist designers, planners, and policymakers in energy driven planning, design and optimization. The model provides essential building EUI data for buildings in the city in

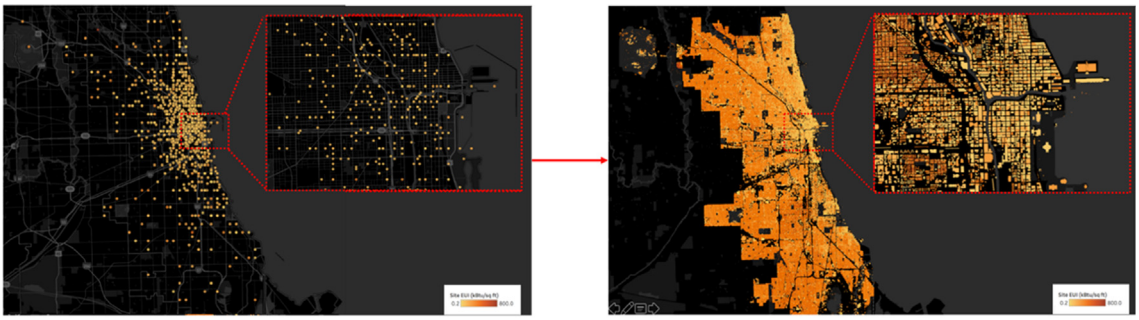

Fig. 8. A building scale energy use prediction and visualization for (a) 2700 buildings and (b) 820,606 buildings.

which their energy information is not available with an acceptable error rate. In the context that many cities across the world have planned to reduce their GHG emissions, the first step is to have a city-wide understanding of energy use patterns with a high level of resolution to address climate change targets and achieve more sustainable cities. For example, Chicago has targeted to reduce energy consumption by 80% by 2050 [42]. Buildings are the main target of reduction in Chicago because they account for 70% of emissions while the transportation sector accounts for 21% [42]. The Energy Benchmarking dataset is currently the only available disclosure dataset but covers only ~1% of buildings in Chicago. Without energy information, a gap remains between aggregated targets and energy reduction strategies. With a lack of building energy information, a data-driven urban scale energy use modeling can provide this energy information. The major contribution of this proposed predictive model is to provide a city-wide building energy consumption model. This model can help in providing how energy is used in city and inform on existing urban building energy profiles and can help in early-stage energy driven planning and design and the application and evaluation of energy efficiency policies.

As the second application, through aggregating across multiple scales, the model performs as a decision-making tool that can help in identifying patterns of energy demand and providing suitable strategies at multi-levels of individual building level, block, neighborhood, and the city scales. The model has the potential to aid energy-driven urban planning and design and help in evaluating a multi-scale energy performance analysis. For example, the multi-scale analysis can be used for informing retrofit targets at the individual building level. It also can capture inter-building effects on energy consumption at block level such as how changes in building height likely impacts energy use. At a larger scale of neighborhood and city, it can inform decision-making regarding how changes in urban spatial patterns such as urban density, zoning, land use, and other planning policies can affect building operational energy use. Fig. 9 illustrates an example visualization that helps communicate multi-scale analysis of urban energy use.

3.2. Integrated urban building and transportation energy use model

This section presents the results of the integrated UEUM model, which captures both building and transportation energy use as a unified system. The integrated model includes building data only for residential and mixed-use buildings and transportation energy use per household, including 46,843 observations. Non-residential buildings are excluded because transportation data are for households so the integrated model can be used to predict urban energy use on a per household basis.

Transportation EUI was modeled for different modes of travel including light vehicle, bus, subway, and rail commuter across various neighborhoods. The urban transportation EUI is estimated through incorporating factors such as mile traveled, fuel economy of different modes, and energy intensity factors per mode of travel. First, transportation EUI per household was estimated and then aggregated at the census tract level, and neighborhood level. In our analysis, the average values for Chicago neighborhoods were found to be 8209 miles per year, which is 19.6% lower than the national average. As the national average values of annual vehicle miles traveled is reported 10,200 mile/year [99]. We estimated the average value of transportation energy use in Chicago to be 43,269 kBtu (12680.89 kWh) per vehicle, which is 9.7% lower than the national average as reported as 47,960 kBtu (14055.68 kWh) per vehicle [99]. The lower estimation may be due to the exclusion of the suburbs in Chicago neighborhoods.

ANN showed the highest performance for predicting building and transportation EUI in the integrated model. Contrary to the single output model in the previous section, k-NN did not show the same performance for integrated model, particularly for transportation EUI. ANN is believed to be one of the most appropriate machine learning methods for capturing complex and non-linear relationships which has increasingly become popular in the field of urban energy studies [59,100–102]. According to the literature, there is still a lack of reliable approach in structuring the topology of the networks and training highly efficient artificial neural network model [103], and this area of research is still under investigation [104]. Thus, finding an efficient model of the artificial neural network is usually based on the rule of thumb approach. For example, with adding or dropping a neuron unit in a hidden layer of a network, the performance of the model can be significantly altered. In this research, an automation code was developed to capture the most efficient network topology from testing various number of hidden neuron, decaying weight, maximum number of iterations and different learning rates based on (1) minimum errors; and (2) minimum variation between train and test sets. This automation code was scripted in nested for-loop approach using R programming language software [69]. The network was trained using Broyden-Fletcher-Goldfarb-Shanno (BFGS) optimization algorithm [105] to increase the convergence speed and select the parameters calculated via sum of squared error (SSE) and activated via *tanh* function. The "nnet" [106] function in R was used to simulate the model. Through applying this approach, all error measurements (MAE, MSE, RMSE, and MAPE), and R-squared for both train and test sets simultaneously the execution time in finding the efficient model was significantly decreased.

Fig. 10 shows the architecture of the best ANN model obtained in this study. The optimum model comprises of 22 input units plus a bias unit, a single hidden layer with 20 hidden neurons plus a bias unit, and two output variables including building EUI and transportation EUI. The black and orange weights on the ANN diagram indicate the magnitude of distributed positive and negative weights, respectively.

Table 12 shows the results of the integrated ANN model, a MAPE of 4.1 for building and MAPE of 0.1 for transportation indicating the predictive power for integrated urban energy use modeling. The results show an R^2 value of 0.41 and 0.96 for building and transportation energy use variables, respectively, indicating that the model explains 41% and 96% of the variance in building EUI and transportation EUI per household. The results suggest considering the urban socio-spatial

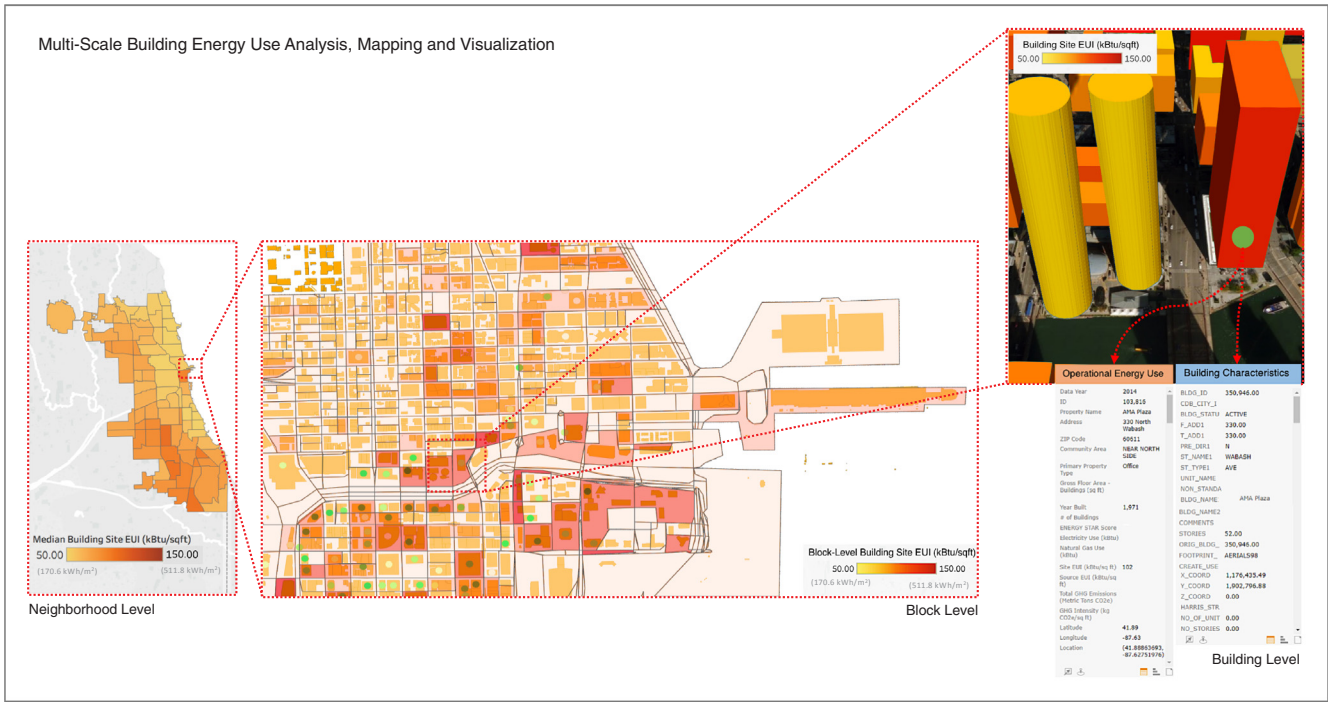

Fig. 9. A multi-scale building energy use analysis and visualization: building level, census block level, neighborhood level, and urban level.

patterns is essential in order to address the energy reduction goals. It should be noted that many other influential factors (e.g., occupant behavior factors) would be needed to incorporate to provide a more comprehensive explanation for variations of energy use in cities as relying on limited factors fails to capture all the variance.

As mentioned in the methodology section, ANN results are difficult to interpret, which makes the ANN to be considered as black box model. There are, however, robust methodologies for illuminating this black-box model [93]. In fact, there are explanatory methods that can be added to neural networks that allow for interpreting the complex relationships captured [91,93,94,107]. In this research, the Garson algorithm [92,94] was applied for explaining the relative contribution of each variable on urban building and transportation energy use using the ANN model. The Garson method presents the relative importance of each predictor by identifying all weighted connections between the nodes through partitioning hidden-output connection weights into input neuron and computing the absolute values of connection weights. The "NeuralNetTools" package [108] in R software was used for quantifying Garson's connection weights.

Fig. 11 shows the results from Garson algorithm, including the relative contributions of each variable on building (Fig. 11a) and transportation (Fig. 11b) energy use as scaled between 0 and 1. Building size (GFA), building height, and year built were found to be the most important variables influencing building EUI, suggesting that these variables are robust urban building energy predictors. After these three building characteristics, socioeconomic indicators, including income, unemployment, crowded housing, and total number of occupants, were found to have the most contribution to building EUI. Regarding transportation EUI, the essential variables impacting transportation EUI were found to be neighborhood features including transit-oriented, bikeability, and VMT, followed by socioeconomic factors, including household education level, poverty level, household dependency level, and income, respectively. Neighborhood walkability and distance to CBD were found to be the next most significant contributors to transportation energy use, followed by unemployment factor. Urban spatial patterns such as building height, as an urban intensity metric, and sprawl index are the next most important contributors to explain urban transportation energy use.

These results suggest that all the factors in the model impact both building and transportation but with different magnitudes. Also, these results demonstrate how socioeconomic factors play a crucial role in urban energy use models for both building and transportation EUI. Urban spatial pattern determinants such as distance to CBD and sprawl index were found to be predictors of both transportation EUI and building EUI. However, these factors have relatively lower impact compared with other predictors such as building characteristics, building height and size, and built year for building EUI prediction, and mobility indices such as transit-oriented and bikeability indices of neighborhoods for transportation EUI prediction, and occupancy and socioeconomic variables for predicting both building EUI and transportation EUI.

The impact of urban spatial patterns such as urban density on building and transportation energy use has gained significant attention in the literature, and there is a consensus that denser urban neighborhoods are associated with lower energy use per capita [27,30]. For example, Newman and Kenworthy [39] suggests that there is a strong negative relationship between urban population density and transportation energy use. However, it is crucial to examine the density variable with considering the importance of other factors such as mobility factors including transit-oriented, walkability, and bikeability as well as land use mix factors that affect transport energy use. By employing a multi-dimensional transport energy use analysis and including urban attributes such as transit-oriented, bikeability, walkability, and sprawl index in addition to building characteristics, such as building height, and socioeconomic factors, this research demonstrates the importance of the other variables for prediction of transportation energy use. The urban spatial patterns such as building height, sprawl dimensions, and distance to CBD affect transportation energy use due to their effects on travel mode and distances. Our results demonstrate that other factors such as mobility factors, including transit-oriented and bikeability features of neighborhoods, can have a significant impact on urban energy use. However, it should be noted that urban transportation energy analysis is associated with a high level of complexity and uncertainty, mostly due to the complex nature of urban systems and the factors involved. The results of this study are also case-specific to Chicago.

While previous studies acknowledge the importance of considering

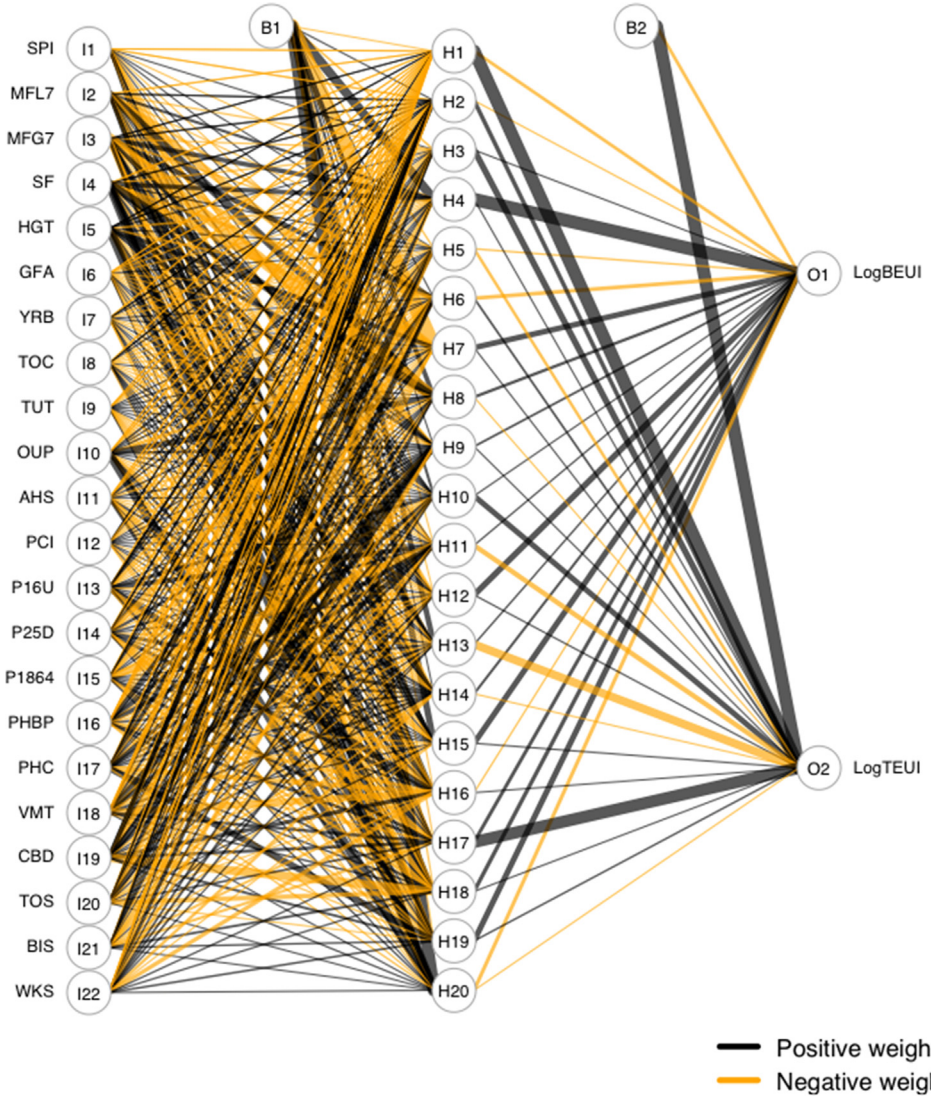

Fig. 10. Architecture of the urban transportation energy model using ANN algorithm. BEUI = Building Energy Use Intensity and TEUI = Transportation EUI.

Table 12
Performance evaluation of the integrated building and transpiration energy use model for Chicago.

Energy model	MSE	RMSE	MAE	MAPE (%)	R^2
Building	0.052	0.229	0.166	4.10	0.41
Transportation	0.002	0.04	0.014	0. 1	0.96

both building and transportation energy end-uses simultaneously due to their interrelationships as well as urban spatial patterns and building characteristics that affect both, few previous studies examined the impact of key urban attributes on both, simultaneously. Moreover, the extant literature tends to primarily focus on urban spatial characteristics [5,27,102]. Integrating transportation and building variables in one framework presents an opportunity to quantify the interdependency of the two components and the impact of the various socio-spatial patterns in both constituents of urban energy model. The results here suggest that all variables in the model are relevant in predicting both sectors, but their effect magnitudes vary. For example, building-related variables such as building height are among the influential factors for building energy prediction which also impact transportation energy with an indirect relationship. Building height also influences some urban spatial attributes such as vertical density which impact travel distances and modes. As another example, urban spatial factors such as distance to CBD or sprawl index influence both building and transportation energy use through their impacts on building thermal needs and mobility factors. Separate modeling and study of building and transportation lead to neglecting the trade-offs between the building energy and transportation energy use.

Finally, including urban spatial patterns along with socioeconomic and occupancy indicators can help more in-depth modeling of the integrated urban energy use. For future study, we aim to provide a quantitative analysis of the complex interplay between the socio-spatial patterns on two main components of urban energy use (building and transportation at the same time) by using Lek's profile and partial dependence (PaD) methods extended on ANN.

4. Limitations and future study

There are a number of limitations involved with this study. First, the accuracy of data-driven models depends upon availability of data as well as sufficient representative variables in the model. There are limitations regarding quality and quantity of the data used in this research. For example, only a limited number of datasets provide energy information at the building level in Chicago as well as other major cities. Chicago's *Energy Benchmarking* dataset, which has been released as a part of disclosure laws similarly adopted by several U.S. cities, does not

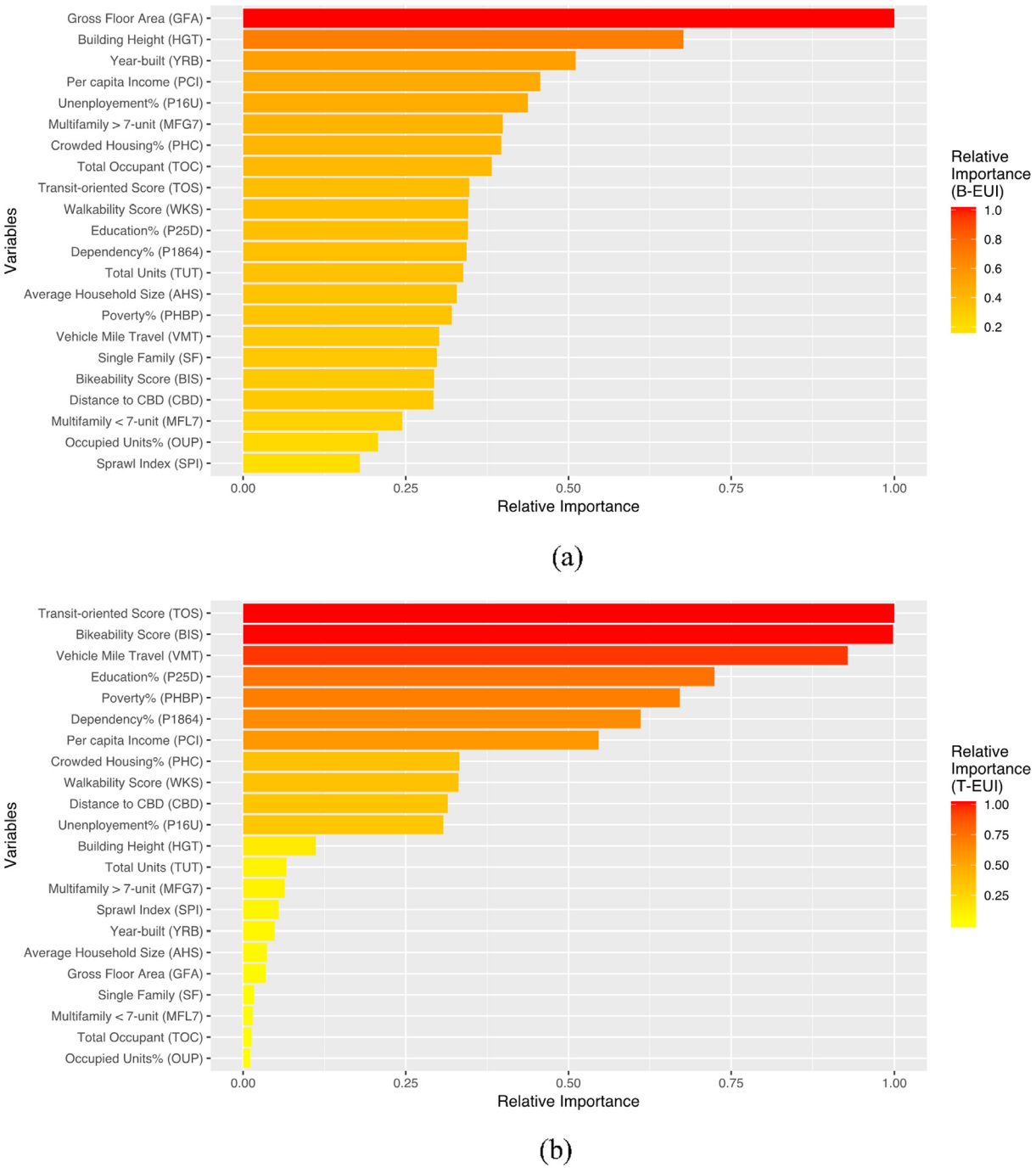

Fig. 11. Importance plot of variables in the integrated model; (a) building EUI and (b) transportation EUI.

provide energy information for all buildings. In addition, it includes only annual energy use data rather than more granular data [79], which leads to a lack of temporal data required for detailed energy analyses. While this study tried to address this issue by merging this dataset with the more comprehensive Chicago Energy Usage dataset, greater availability of datasets with higher quality data and larger number of observations can produce more reliable results. Other data used in this research to capture building characteristics and urban spatial patterns have limitations as well and often include substantial missing information. In addition, even though data-driven urban energy models can provide a realistic representation of energy consumption compared with simulation-based methods, they have limitations in terms of representation, modification and evaluation of building systems such as HVAC systems. Another limitation of this research is that, unlike building EUI data that are observed values, transportation EUI values were achieved through estimation, which leads to uncertainties in the transportation energy model.

Future work could improve the models used in this research by developing a hybrid urban energy modeling framework that integrates both engineering-based and data-driven energy prediction approaches and combines the strengthens of each model. For example, data-driven models can be coupled with simulation models to provide local data such as microclimate data or occupancy and human-related data, which are often oversimplified by current urban scale energy simulation models. Simulation models enable incorporating building construction systems, HVAC systems, and technology-related variables that data-driven models do not often capture. Also, simulation methods allow for increasing granularity of the model to a daily, hourly, and real-time

prediction.

The framework developed in this research does not account for embodied energy of buildings and road infrastructure. Future research therefore should also incorporate embodied energy of urban elements to present a more comprehensive definition and modeling of urban scale energy use. We acknowledge that the results of this study are case specific for Chicago and do not represent other cities. However, the UEUM framework has the potential to be applied to other cities. Also, in this article, we only report on relative importance of various urban socio-spatial factors in buildings and transportation energy use variations. We aim to provide more information on how these factors impact energy use in a future study.

5. Conclusion

This research developed a multi-scale data-driven urban energy use modeling framework that integrates building and transportation energy use at neighborhood and urban contexts. Our results show that applying advanced machine learning (ML) techniques, using disaggregated individual building level energy data, and considering the influential urban socio-spatial factors can improve the accuracy of an urban energy prediction model. This research tested several well-established ML algorithms, including Multiple Linear Regression (MLR), Nonlinear Regression (NLR), Classification and Regression Trees (C&RT), Random Decision Forest (RDF), k-Nearest Neighbors (k-NN), and Artificial Neural Networks (ANNs). Among these algorithms, k-NN and ANN trained by BFGS optimization algorithm provided the best overall performance in terms of accuracy. While k-NN algorithm was more accurate in urban building energy use modeling than in integrated building and transportation modeling, ANN algorithm had the best performance for integrated modeling compared to the other algorithms. This research also examined the relative role of key urban attributes on urban energy performance. Building characteristics such as building size and height, followed by occupancy and socioeconomic variables such as household income, education, and employment variables, have the greatest impact on building EUI. The most influential factors on transportation EUI were found to be mobility and travel patterns, including neighborhood transit-oriented score, daily Vehicle Miles Travel (VMT), and neighborhood bikeability and walkability indices, followed by socioeconomic factors and distance to Central Business District (CBD), which represents the location of neighborhood across the city. The results also indicate that while sprawl index, representing density, accessibility, and land use is one of the predictors of building and transportation EUI, its effects represent lower relative importance compared to the effects of other variables in the model, such as the effect of neighborhood's transit-oriented factor on transportation EUI or the effects of occupancy and socioeconomic factors on both building and transportation EUI. The outcomes of this study provide opportunities to model and analyze urban energy use dynamics across neighborhoods in a city with acceptable error rates. This is a fundamental step in identifying and addressing potential energy use challenges and improvement strategies in urban context. This work also provides insight into the trade-offs between different influential urban factors and urban energy use. However, it is important to note that more comprehensive understanding and modeling of cities as complex systems remain an open question and additional investigation and considering other influential factors are required. Overall, the results of this research can assist architects, engineers, urban designers and planners and policy-makers to understand interrelationships between urban socio-spatial patterns and urban energy consumption through a more comprehensive perspective that supports sustainable energy policies for existing and future cities. The proposed integrated urban energy framework can be used for interpreting the effects of each variable within an entire urban system and track subsequent variations in both building and transportation energy use, simultaneously.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

None.

Acknowledgements

The primary author, Narjes Abbasabadi, gratefully thanks and acknowledges the advice, insight, and support of Michelangelo Sabatino, her Ph.D. dissertation advisor, during the process of this research. The primary author also thanks Mahjoub Elnimeiri who provided insightful comments that assisted the research.

References

[1] World Population Prospects. United Nations; 2018. https://population.un.org/wpp/ [accessed June 16, 2019].
[2] IPCC. Climate change 2014: Synthesis report. Contribution of working groups I, II and III to the fifth assessment report of the intergovernmental panel on climate change. Geneva, Switzerland; 2014. https://doi.org/10.1017/CBO9781107415324.
[3] US EPA O. Inventory of U.S. Greenhouse Gas Emissions and Sinks. US EPA; 2017. https://www.epa.gov/ghgemissions/inventory-us-greenhouse-gas-emissions-and-sinks [accessed June 22, 2019].
[4] Sola A, Corchero C, Salom J, Sanmarti M. Simulation tools to build urban-scale energy models: a review. 3269–3269 Energies 2018;11. https://doi.org/10.3390/en11123269.
[5] Osório B, McCullen N, Walker I, Coley D. Integrating the energy costs of urban transport and buildings. Sustain Cities Soc 2017;32:669–81. https://doi.org/10.1016/j.scs.2017.04.020.
[6] Liu X, Ou J, Chen Y, Wang S, Li X, Jiao L, et al. Scenario simulation of urban energy-related CO2 emissions by coupling the socioeconomic factors and spatial structures. Appl Energy 2019;238:1163–78. https://doi.org/10.1016/j.apenergy.2019.01.173.
[7] de Wilde P. The gap between predicted and measured energy performance of buildings: a framework for investigation. Autom Constr 2014;41:40–9. https://doi.org/10.1016/j.autcon.2014.02.009.
[8] Srebric J, Heidarinejad M, Liu J. Building neighborhood emerging properties and their impacts on multi-scale modeling of building energy and airflows. Build Environ 2015;91:246–62. https://doi.org/10.1016/j.buildenv.2015.02.031.
[9] Reinhart CF, Cerezo Davila C. Urban building energy modeling – a review of a nascent field. Build Environ 2016;97:196–202. https://doi.org/10.1016/J.BUILDENV.2015.12.001.
[10] 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.
[11] 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.
[12] Chen Y, Hong T, Piette MA. Automatic generation and simulation of urban building energy models based on city datasets for city-scale building retrofit analysis. Appl Energy 2017;205:323–35. https://doi.org/10.1016/j.apenergy.2017.07.128.
[13] Li Q, Quan SJ, Augenbroe G, Pei P, Yang-Ju, Brown J. Building energy modelling at urban scale: integration of reduced order energy model with geographical information. Ibpsa; 2015. p. 190.
[14] Nutkiewicz A, Yang Z, Jain RK. Data-driven Urban Energy Simulation (DUE-S): a framework for integrating engineering simulation and machine learning methods in a multi-scale urban energy modeling workflow. Appl Energy 2018;225:1176–89. https://doi.org/10.1016/j.apenergy.2018.05.023.
[15] Hsu D. Identifying key variables and interactions in statistical models of building energy consumption using regularization. Energy 2015;83:144–55. https://doi.org/10.1016/J.ENERGY.2015.02.008.
[16] Paula Guillaumet M, Rosas-Casals M, Travesset-Baro O. Building archetypes in Urban Energy Models. A comparative case study of deterministic and statistical methods in Andorra. Conference: uSIM 2018 - urban energy simulation, at The University of Strathclyde Technology and Innovation Centre, 99 George St, Glasgow G1 1RD. 2018.
[17] Kontokosta C, Bonczak B, Duer-balkind M. Data IQ – A machine learning approach to anomaly detection for energy performance data quality and reliability. 2016 ACEEE summer study on energy efficiency in buildings; 2016.
[18] Keirstead J, Jennings M, Sivakumar A. A review of urban energy system models: approaches, challenges and opportunities. Renew Sustain Energy Rev 2012;16:3847–66. https://doi.org/10.1016/j.rser.2012.02.047.
[19] Mousavi MH, Ghavidel S. Structural time series model for energy demand in Iran's

transportation sector. Case Stud Transp Policy 2019;7:423–32. https://doi.org/10.1016/j.cstp.2019.02.004.
[20] Zhou W, Huang G, Cadenasso ML. Does spatial configuration matter? Understanding the effects of land cover pattern on land surface temperature in urban landscapes. Landsc Urban Plann 2011;102:54–63. https://doi.org/10.1016/J.LANDURBPLAN.2011.03.009.
[21] Jin H, Cui P, Wong N, Ignatius M, Jin H, Cui P, et al. Assessing the effects of urban morphology parameters on microclimate in Singapore to control the urban heat island effect. Sustainability 2018;10:206. https://doi.org/10.3390/su10010206.
[22] Gracik S, Heidarinejad M, Liu J, Srebric J. Effect of urban neighborhoods on the performance of building cooling systems. Build Environ 2015;90:15–29. https://doi.org/10.1016/j.buildenv.2015.02.037.
[23] Ahmed Memon R, Leung YCD, Liu C. A review on the generation, determination and mitigation of Urban Heat Island. J Environ Sci 2008;20:120–8. https://doi.org/10.1016/S1001-0742(08)60019-4.
[24] Jiang D, Jiang W, Liu H, Sun J. Systematic influence of different building spacing, height and layout on mean wind and turbulent characteristics within and over urban building arrays. Wind Struct 2008;11:275–89.
[25] Palme M, Ramírez J. A critical assessment and projection of urban vertical growth in Antofagasta, Chile. Sustainability 2013;5:2840–55. https://doi.org/10.3390/su5072840.
[26] Clark TA. Metropolitan density, energy efficiency and carbon emissions: multi-attribute tradeoffs and their policy implications. Energy Policy 2013;53:413–28. https://doi.org/10.1016/j.enpol.2012.11.006.
[27] Norman J, MacLean HL, Kennedy CA. Comparing high and low residential density: life-cycle analysis of energy use and greenhouse gas emissions. J Urban Plann Develop 2006:10–21. https://doi.org/10.1061/(ASCE)0733-9488(2006)132.
[28] Dall'O' G, Galante A, Torri M. A methodology for the energy performance classification of residential building stock on an urban scale. Energy Build 2012;48:211–9. https://doi.org/10.1016/j.enbuild.2012.01.034.
[29] Martins TA de L, Faraut S, Adolphe L. Influence of context-sensitive urban and architectural design factors on the energy demand of buildings in Toulouse, France. Energy Build 2019;190:262–78. https://doi.org/10.1016/j.enbuild.2019.02.019.
[30] Steemers K. Energy and the city: density, buildings and transport. Energy Build 2003;35:3–14. https://doi.org/10.1016/S0378-7788(02)00075-0.
[31] Jia M, Srinivasan RS, Raheem AA. From occupancy to occupant behavior: An analytical survey of data acquisition technologies, modeling methodologies and simulation coupling mechanisms for building energy efficiency. Renew Sustain Energy Rev 2017;68:525–40. https://doi.org/10.1016/j.rser.2016.10.011.
[32] Yu Z, Fung BCM, Haghighat F, Yoshino H, Morofsky E. A systematic procedure to study the influence of occupant behavior on building energy consumption. Energy Build 2011;43:1409–17. https://doi.org/10.1016/j.enbuild.2011.02.002.
[33] Kim Y-S, Heidarinejad M, Dahlhausen M, Srebric J. Building energy model calibration with schedules derived from electricity use data. Appl Energy 2017;190:997–1007. https://doi.org/10.1016/j.apenergy.2016.12.167.
[34] Happle G, Fonseca JA, Schlueter A. A review on occupant behavior in urban building energy models. Energy Build 2018;174:276–92. https://doi.org/10.1016/J.ENBUILD.2018.06.030.
[35] Wiedenhofer D, Lenzen M, Steinberger JK. Energy requirements of consumption: urban form, climatic and socio-economic factors, rebounds and their policy implications. Energy Policy 2013;63:696–707. https://doi.org/10.1016/j.enpol.2013.07.035.
[36] Yun GY, Steemers K. Behavioural, physical and socio-economic factors in household cooling energy consumption. Appl Energy 2011;88:2191–200. https://doi.org/10.1016/j.apenergy.2011.01.010.
[37] Büchs M, Schnepf SV. Who emits most? Associations between socio-economic factors and UK households' home energy, transport, indirect and total CO2 emissions. Ecol Econ 2013;90:114–23. https://doi.org/10.1016/j.ecolecon.2013.03.007.
[38] Dagoumas A. Modelling socio-economic and energy aspects of urban systems. Sustain Cities Soc 2014;13:192–206. https://doi.org/10.1016/j.scs.2013.11.003.
[39] Newman PWG, Kenworthy JR. Gasoline consumption and cities. J Am Plann Assoc 1989;55:24–37. https://doi.org/10.1080/01944368908975398.
[40] Fan C, Xiao F, Yan C, Liu C, Li Z, Wang J. A novel methodology to explain and evaluate data-driven building energy performance models based on interpretable machine learning. Appl Energy 2019;235:1551–60. https://doi.org/10.1016/j.apenergy.2018.11.081.
[41] Abbasabadi N, Ashayeri JKM. Urban energy use modeling methods and tools: A review and an outlook. Build Environ 2019;161:106270https://doi.org/10.1016/j.buildenv.2019.106270.
[42] City of Chicago Climate Action Plan. City of Chicago Climate Action Plan; 2019. http://www.chicagoclimateaction.org/ [accessed February 5, 2019].
[43] Pisello AL, Taylor JE, Xu X, Cotana F. Inter-building effect: Simulating the impact of a network of buildings on the accuracy of building energy performance predictions. Build Environ 2012;58:37–45. https://doi.org/10.1016/j.buildenv.2012.06.017.
[44] Fonseca JA, Schlueter A. Integrated model for characterization of spatiotemporal building energy consumption patterns in neighborhoods and city districts. Appl Energy 2015;142:247–65. https://doi.org/10.1016/j.apenergy.2014.12.068.
[45] Ramaswami A, Chavez A. What metrics best reflect the energy and carbon intensity of cities? Insights from theory and modeling of 20 US cities. Environ Res Lett 2013;8. https://doi.org/10.1088/1748-9326/8/3/035011.
[46] Costa A, Blanes LM, Donnelly C, Keane MM. Review of EU airport energy interests and priorities with respect to ICT, energy efficiency and enhanced building operation. International conference for enhanced building operations. 2012.
[47] Kim Y-S, Srebric J. Impact of occupancy rates on the building electricity consumption in commercial buildings. Energy Build 2017;138:591–600. https://doi.org/10.1016/J.ENBUILD.2016.12.056.
[48] Ewing R, Hamidi S. Measuring urban sprawl and validating sprawl measures Technical Report Prepared for the National Cancer Institute, National Institutes of Health, the Ford Foundation, and Smart Growth America Salt Lake City (UT, USA): University of Utah; 2014
[49] Rubin DB. Multiple Imputation after 18+ Years. J Am Stat Assoc 1996;91:473–89. https://doi.org/10.1080/01621459.1996.10476908.
[50] Rupert G, Miller J. Beyond ANOVA: basics of applied statistics. New York: Chapman & Hall/CRC, Taylor & . Francis Group; 1998.
[51] Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. J Am Stat Assoc 1952;47:583–621. https://doi.org/10.1080/01621459.1952.10483441.
[52] Mansfield ER, Helms BP. Detecting multicollinearity. Am Statist 1982;36:158–60. https://doi.org/10.1080/00031305.1982.10482818.
[53] Goswami S, Chakrabarti A. Feature selection: a practitioner view. IJ Inform Technol Comput Sci 2014;11:66–77.
[54] Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, Wu AY. An efficient k-Means clustering algorithm: analysis and implementation. IEEE Trans Patt Anal Mach Intell 2002;24:881. https://doi.org/10.1109/TPAMI.2002.1017616.
[55] Heidarinejad M, Dahlhausen M, McMahon S, Pyke C, Srebric J. Cluster analysis of simulated energy use for LEED certified U.S. office buildings. Energy Build 2014;85:86–97. https://doi.org/10.1016/j.enbuild.2014.09.017.
[56] Ahmad T, Chen H, Guo Y, Wang J. A comprehensive overview on the data driven and large scale based approaches for forecasting of building energy demand: a review. Energy Build 2018;165:301–20. https://doi.org/10.1016/j.enbuild.2018.01.017.
[57] Jovanović RŽ, Sretenović AA, Živković BD. Ensemble of various neural networks for prediction of heating energy consumption. Energy Build 2015;94:189–99. https://doi.org/10.1016/J.ENBUILD.2015.02.052.
[58] Amasyali K, El-Gohary NM. A review of data-driven building energy consumption prediction studies. Renew Sustain Energy Rev 2018;81:1192–205. https://doi.org/10.1016/j.rser.2017.04.095.
[59] Park SK, Moon HJ, Min KC, Hwang C, Kim S. Application of a multiple linear regression and an artificial neural network model for the heating performance analysis and hourly prediction of a large-scale ground source heat pump system. Energy Build 2018;165:206–15. https://doi.org/10.1016/j.enbuild.2018.01.029.
[60] Leach LF, Henson RK, Finch WH, Fraas JW, Newman I, Walker DA. Multiple linear regression viewpoints. AERA Special Interest Group on Multiple Linear Regression: General Linear Model through the University of Alabama at Birmingham, vol. 33; 2007.
[61] Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev 1958;65:386–408. https://doi.org/10.1037/h0042519.
[62] Deng H, Fannon D, Eckelman MJ. Predictive modeling for US commercial building energy use: a comparison of existing statistical and machine learning algorithms using CBECS microdata. Energy Build 2018;163:34–43. https://doi.org/10.1016/j.enbuild.2017.12.031.
[63] Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. Am Statist 1992;46:175–85. https://doi.org/10.1080/00031305.1992.10475879.
[64] Breiman L, editor. Classification and regression trees. Repr. Boca Raton: Chapman & Hall [u.a.]; 1998.
[65] Ho Tin Kam. Random decision forests. Proceedings of 3rd international conference on document analysis and recognition, vol. 1. Montreal (Que., Canada): IEEE Comput. Soc. Press; 1995. p. 278–82. https://doi.org/10.1109/ICDAR.1995.598994.
[66] Borra S, Di Ciaccio A. Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. Comput Stat Data Anal 2010;54:2976–89. https://doi.org/10.1016/j.csda.2010.03.004.
[67] Torabi Moghadam S, Toniolo J, Mutani G, Lombardi P. A GIS-statistical approach for assessing built environment energy use at urban scale. Sustain Cities Soc 2018;37:70–84. https://doi.org/10.1016/J.SCS.2017.10.002.
[68] Ahmad AS, Hassan MY, Abdullah MP, Rahman HA, Hussin F, Abdullah H, et al. A review on applications of ANN and SVM for building electrical energy consumption forecasting. Renew Sustain Energy Rev 2014;33:102–9. https://doi.org/10.1016/J.RSER.2014.01.069.
[69] R: The R Project for Statistical Computing. The R Project for Statistical Computing; 2019. https://www.r-project.org/ [accessed June 6, 2019].
[70] Building Footprints (current). Chicago Data Portal; 2015. https://data.cityofchicago.org/Buildings/Building-Footprints-current-/hz9b-7nh8?category=Buildings&view_name=Building-Footprints-current- [accessed February 5, 2019].
[71] PLUTO and MapPLUTO, Release 18v1. New York City, Department of City Planning; 2019. https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.page [accessed February 5, 2019].
[72] Boundaries - Zoning Districts (current). Chicago Data Portal; 2019. https://data.cityofchicago.org/Community-Economic-Development/Boundaries-Zoning-Districts-current-/7cve-jgbp [accessed February 5, 2019].
[73] Cook County Assessor Data. Cook County Government, Open Data; 2019. https://datacatalog.cookcountyil.gov/ [accessed May 6, 2019].
[74] Updated Urban Sprawl Data for the United States. National Cancer Institute, Geographic Information Systems & Science for Cancer Control; 2010. https://gis.cancer.gov/tools/urban-sprawl/ [accessed February 5, 2019].
[75] Ewing R, Meakins G, Hamidi S, Nelson AC. Relationship between urban sprawl and physical activity, obesity, and morbidity – update and refinement. Health Place 2014;26:118–26. https://doi.org/10.1016/j.healthplace.2013.12.008.

[76] Boundaries - Census Tracts - 2010. Chicago Data Portal; 2010. https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Census-Tracts-2010/5jrd-6zik [accessed February 5, 2019].
[77] Chicago neighborhoods on Walk Score. Walk Score; 2019. https://www.walkscore.com/IL/Chicago [accessed June 3, 2019].
[78] Census Data - Selected socioeconomic indicators in Chicago, 2008 – 2012. Chicago Data Portal; 2019. https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2 [accessed March 20, 2019].
[79] City of Chicago. Chicago Energy Benchmarking - 2016 Data Reported in 2017, City of Chicago, Data Portal. City of Chicago, Environment & Sustainable Development; 2016. https://data.cityofchicago.org/Environment-Sustainable-Development/Chicago-Energy-Benchmarking-2016-Data-Reported-in-/fpwt-snya.
[80] City of Chicago. Energy Usage 2010, City of Chicago, Data Portal. City of Chicago, Environment & Sustainable Development; 2010. https://data.cityofchicago.org/Environment-Sustainable-Development/Energy-Usage-2010/8yq3-m6wp [accessed February 5, 2019].
[81] Household Travel Survey. Chicago Metropolitan Agency for Planning (CMAP); 2016. http://www.cmap.illinois.gov/data/transportation/travel-survey [accessed February 5, 2019].
[82] Fuel Economy. US of Energy; 2019. https://www.fueleconomy.gov/ [accessed February 5, 2019].
[83] National Transportation Statistics. United States Department of Transportation, Bureau of Transportation Statistics; 2019. https://www.bts.gov/topics/national-transportation-statistics [accessed February 5, 2019].
[84] Kutner MH, Chris Nachtsheim, John Neter. Applied linear regression models. 5th ed. McGraw-Hill/Irwin; 2004.
[85] Cook D. Detection of influatial observation in linear regression. SAGE Handb Regress Anal Caus Infer 1977;19:15–8. https://doi.org/10.4135/9781446288146.
[86] Belsley DA, Edwin Kuh, Welsch RE. Regression diagnostics: identifying influential data and sources of collinearity. Wiley; 1980.
[87] Wilcoxon F. Individual comparisons by ranking methods. Biometr Bull 1945;1:80. https://doi.org/10.2307/3001968.
[88] Dupuis DJ, Victoria-Feser M-P. Robust VIF regression with application to variable selection in large data sets. Ann Appl Stat 2013;7:319–41. https://doi.org/10.1214/12-AOAS584.
[89] Lindsey M, Schofer JL, Durango-Cohen P, Gray KA. The effect of residential location on vehicle miles of travel, energy consumption and greenhouse gas emissions: Chicago case study. Transp Res Part D: Transp Environ 2011;16:1–9. https://doi.org/10.1016/J.TRD.2010.08.004.
[90] Olden JD, Jackson DA. Illuminating the “black box”: a randomization approach for understanding variable contributions in artificial neural networks. Ecol Model 2002;154:135–50. https://doi.org/10.1016/S0304-3800(02)00064-9.
[91] Gevrey M, Dimopoulos I, Lek S. Review and comparison of methods to study the contribution of variables in artificial neural network models. Ecol Model 2003;160:249–64. https://doi.org/10.1016/S0304-3800(02)00257-0.
[92] Garson GD. Interpreting neural-network connection weights. AI Expert 1991;6:46–51.
[93] Olden JD, Joy MK, Death RG. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecol Model 2004;178:389–97. https://doi.org/10.1016/j.ecolmodel.2004.03.013.
[94] Fischer A. How to determine the unique contributions of input-variables to the nonlinear regression function of a multilayer perceptron. Ecol Model 2015;309–310:60–3. https://doi.org/10.1016/j.ecolmodel.2015.04.015.
[95] Johannesen NJ, Kolhe M, Goodwin M. Relative evaluation of regression tools for urban area electrical energy demand forecasting. J Cleaner Prod 2019;218:555–64. https://doi.org/10.1016/j.jclepro.2019.01.108.
[96] Al-Qahtani FH, Crone SF. Multivariate k-nearest neighbour regression for time series data — a novel algorithm for forecasting UK electricity demand. 2013 international joint conference on neural networks (IJCNN 2013 - Dallas) IEEE; 2013. p. 1–8. https://doi.org/10.1109/IJCNN.2013.6706742.
[97] Kontokosta CE, Tull C. A data-driven predictive model of city-scale energy use in buildings. Appl Energy 2017;197:303–17. https://doi.org/10.1016/j.apenergy.2017.04.005.
[98] Hong S-M, Paterson G, Mumovic D, Steadman P. Improved benchmarking comparability for energy consumption in schools. Building Research & Information 2014;42:47–61. https://doi.org/10.1080/09613218.2013.814746.
[99] Davis SC, Williams SE, Boundy RG. Transportation energy data book. U.S. Department of Energy, Oak Ridge National Laboratory; 2018.
[100] Murat YS, Ceylan H. Use of artificial neural networks for transport energy demand modeling. Energy Policy 2006;34:3165–72. https://doi.org/10.1016/j.enpol.2005.02.010.
[101] Azadeh A, Ghaderi SF, Sohrabkhani S. Annual electricity consumption forecasting by neural network in high energy consuming industrial sectors. Energy Convers Manage 2008;49:2272–8. https://doi.org/10.1016/J.ENCONMAN.2008.01.035.
[102] Silva MC, Horta IM, Leal V, Oliveira V. A spatially-explicit methodological framework based on neural networks to assess the effect of urban form on energy demand. Appl Energy 2017;202:386–98. https://doi.org/10.1016/j.apenergy.2017.05.113.
[103] Thomas AJ, Petridis M, Walters SD, Gheytassi SM, Morgan RE. On predicting the optimal number of hidden nodes. 2015 international conference on computational science and computational intelligence (CSCI), Las Vegas, NV, USA IEEE; 2015. p. 565–70. https://doi.org/10.1109/CSCI.2015.33.
[104] Sambatti SBM, Anochi JA, Luz EF avero P, Shiguemori EH, Carvalho AR, Velho HF de C. Automatic configuration for neural network applied to atmospheric temperature profile identification. EngOpt 2012, Rio de Janeiro, Brazil; 2012.
[105] Battiti R, Masulli F. BFGS optimization for faster and automated supervised learning. International neural network conference: July 9–13, 1990 Palais Des Congres — Paris — France 60: DordrechtSpringer Netherlands; 1990. p. 757. https://doi.org/10.1007/978-94-009-0643-3_68.
[106] Venables WN, Ripley BD. Modern applied statistics with S ISBN 0-387-95457-0 4th ed. New York: Springer; 2002.
[107] Gevrey M, Dimopoulos I, Lek S. Two-way interaction of input variables in the sensitivity analysis of neural network models. Ecol Model 2006;195:43–50. https://doi.org/10.1016/j.ecolmodel.2005.11.008.
[108] Beck MW. NeuralNetTools: Visualization and Analysis Tools for Neural Networks. J Stat Software 2018;85:1–20. https://doi.org/10.18637/jss.v085.i11.

further reading

[109] Papadopoulos S, Azar E, Woon W-L, Kontokosta CE. Evaluation of tree-based ensemble learning algorithms for building energy performance estimation. J Build Perform Simul 2018;11:322–32. https://doi.org/10.1080/19401493.2017.1354919.