

# **Urban Building Energy Prediction at Community Scale: A Case Study Using Data-Driven Methods in Jianhu City, China**

A Capstone Presented to the Faculty of Architecture, Planning and Preservation  
COLUMBIA UNIVERSITY

In Partial Fulfillment  
of the Requirements for the Degree  
Master of Science in Urban Planning

by  
Qi Lin

March 2021

Advisor: Malo Hutson  
Reader: Boyeong Hong

## **ABSTRACT**

Predictive models for urban building energy use have been the focus of much research in recent years, especially using data-driven techniques. However, these models still need to address recognized challenges, such as employing sufficient energy use data in spatial and temporal scales and accounting for interbuilding effects. In this regard, several typical data-driven predictive models for urban building energy use were proposed in this capstone to reduce the large data requirements and improve the prediction accuracy. Using a dataset of four years of electricity consumption by public buildings in Jianhu City, a county-level city in Jiangsu Province, China, and data on the corresponding building morphological parameters, this project compares the predictive performance of these models under different algorithms. The results suggest that a building network based on building morphological similarity can improve the overall performance of energy consumption prediction models for individual buildings in an urban context. This building network can also obtain relatively reliable energy consumption prediction results in the absence of historical energy consumption data of the target building. The project also reveals that the data-driven models can accurately predict total building consumption in a region when historical energy consumption of some buildings is not available. This study provides more comprehensive references and improved accuracy and robustness of urban building energy demand prediction, resulting in potential solutions reduced data requirements of urban energy models.

# ACKNOWLEDGEMENT

First, I would like to deeply thank Professor Malo Hutson for being not only an experienced advisor but also my career mentor, who has always given me invaluable advice. He provided me with maximum freedom and support in all aspects of my project. I received guidance on proposal submissions, literature reviews, and teamwork, as well as paper composition and presentation.

I would like to thank Professor Hong Zhang and Professor Wei Wang at Southeast University, China. They provided me with valuable data necessary for the smooth conduct of research and indispensable suggestions on how to improve my study. This project could not have been completed without their support.

I am also very grateful to Professor Boyeong Hong and her willingness to be my reader. Her patience and expertise were extremely helpful. She provided me with meaningful strategies to advance this research project, from the simplest code-debugging to the selection of machine learning models and exploration of research ideas.

I would like to thank Professor Anthony Vanky for enlightening me on the consideration of cities from a data-driven perspective. I also appreciate his continuous support and encouragement.

I also wish to thank Professor Weiping Wu, the director of the M.S. Urban Planning program. Her constant dedication to the development of the Urban Planning program is much appreciated, and she is always striving for more opportunities for students. Her rigorous attitude and persistence in academics have always inspired me.

I would also like to show appreciation to the managers and staff of the Urban Planning program, especially Emily Junker, Leigh Brown and Michael Montilla, who made every effort to help students complete their studies and graduate successfully.

Last but not least, I would like to express my gratitude to my family and friends from the bottom of my heart.

# TABLE OF CONTENTS

<b>ABSTRACT</b> .....	<b>ii</b>
<b>ACKNOWLEDGEMENT</b> .....	<b>iii</b>
<b>CHAPTER 1: INTRODUCTION</b> .....	<b>1</b>
1.1 Client organization information.....	1
1.2 Purpose of the project.....	1
1.3 Research goals.....	2
1.4 Background.....	2
1.4.1 Building energy consumption in urban contexts.....	2
1.4.2 Energy consumption prediction methods for urban building.....	3
1.4.3 Review of data-driven studies for urban building energy modeling (UBEM).....	4
<b>CHAPTER 2: METHODOLOGY</b> .....	<b>8</b>
2.1 Data description & data processing.....	8
2.1.1 Building energy data.....	8
2.1.2 Building design parameter data.....	13
2.2 Data-driven predictive techniques.....	19
2.2.1 K-nearest neighbors.....	19
2.2.2 Support vector regression.....	20
2.2.3 Long-short memory network algorithm.....	20
2.3 Prediction models.....	22
2.3.1 Time-series predictive model for individual building energy use.....	22
2.3.2 Building-network predictive model for individual buildings based on the similarity of building energy consumption curves.....	23
2.3.3 Building-network predictive model for individual buildings based on the similarity of building morphological data.....	25
2.3.4 Time-series total energy predictive model for all buildings.....	26

2.4 Model parameter tuning and evaluation .....	27
<b>CHAPTER 3: RESULTS AND DISCUSSION .....</b>	<b>29</b>
3.1 Energy prediction results for Model 1 .....	29
3.2 Energy prediction results for Models 2 and 3.....	29
3.3 Energy prediction results for Model 4.....	32
3.4 Discussion.....	37
<b>CHAPTER 4: LIMITATIONS AND FUTURE WORK.....</b>	<b>39</b>
<b>CHAPTER 5: CONCLUSIONS .....</b>	<b>41</b>
<b>BIBLIOGRAPHY .....</b>	<b>43</b>

# LIST OF TABLES

Table 2.1 Building energy consumption.....	8
Table 2.2 Monthly EUI (kWh/m <sup>2</sup> ) of all buildings.....	9
Table 2.3 Building design parameters .....	15
Table 2.4 Building design parameters statistics summary.....	16
Table 3.1 Model 1 results.....	29
Table 3.2 Model 2 results.....	30
Table 3.3 Model 3 results.....	30
Table 3.4 R <sup>2</sup> results of Model 4 .....	35
Table 3.5 RMSE (kWh/m <sup>2</sup> ) results of Model 4 .....	36

# LIST OF FIGURES

Figure 2.1 Monthly EUI (kWh/m <sup>2</sup> ) distribution of all buildings .....	10
Figure 2.2 Monthly EUI (kWh/m <sup>2</sup> ) distribution of each building .....	10
Figure 2.3 Normalized monthly EUI trend of each building .....	12
Figure 2.4 Normalized monthly EUI distribution of each month .....	12
Figure 2.5 Normalized monthly EUI distribution of each building .....	13
Figure 2.6 Location and layout of Jianhu City and the distribution of the available buildings in the original dataset .....	14
Figure 2.7 Basic geometry of buildings .....	17
Figure 2.8 Building shape factors .....	17
Figure 2.9 Obstruction height to canyon width .....	18
Figure 2.10 Building orientation .....	18
Figure 2.11 Building function .....	19
Figure 2.12 The schematic diagram of Model 1 .....	23
Figure 2.13 The schematic diagram of Model 2 .....	25
Figure 2.14 The schematic diagram of Model 3 .....	26
Figure 2.15 The schematic diagram of Model 4 .....	27
Figure 3.1 The first case of Model 2 and Model 3 compared with the results of Model 1 .....	31
Figure 3.2 The second case of Model 2 and Model 3 compared with the results of Model 1 .....	32
Figure 3.3 R <sup>2</sup> results of Model 4 .....	33
Figure 3.4 RMSE (kWh/m <sup>2</sup> ) results of Model 4 .....	34

# **CHAPTER 1: INTRODUCTION**

## **1.1 Client organization information**

The research of the Institute of Building Technology and Science at Southeast University, China is effectively integrated through the academic backgrounds of architectural design, urban design, building technology and urban data analysis, forming a technical community represented by multiple research directions. These include: 1) theory and method of industrialized building design systems based on component methods; 2) technology and method for active protection of architectural heritage; 3) research and application of environmentally sustainable, low-carbon and healthy building; and 4) design and planning of human-oriented and energy-saving smart cities.

## **1.2 Purpose of the project**

With economic development and accelerated urbanization, China's emerging cities are developing and building at an increasing rate. At the same time, policy makers and researchers are also becoming increasingly aware of the opportunities for implementing energy efficiency strategies in county-level cities. Building energy consumption accounts for a large proportion of overall urban energy consumption; therefore, effective building energy efficiency strategies can contribute significantly to achieving the ultimate goal of urban energy conservation. Furthermore, understanding the characteristics of building energy consumption is the key to developing a reasonable and effective building energy-saving strategy. Effective prediction of building energy consumption becomes an essential means of grasping the characteristics and trends of building energy consumption. In recent years, machine learning has become more widely used, and data-

driven prediction models have been a hot research topic in the study of urban building energy use. However, the traditional data-driven research on the energy use of urban buildings relies on a large volume of multi-dimensional historical energy consumption data and easily ignores inter-effects between buildings, both of which are key challenges to building energy prediction models at the neighborhood scale in county-level cities in China. In this regard, several typical data-driven models at an urban scale were proposed to find opportunities to reduce the volume of required data and improve prediction accuracy. These models were then to a county-level city in China, validating the models with monthly energy datasets and building morphological datasets.

### **1.3 Research goals**

Using the building morphological data and energy consumption data of a county-level city in China, this study seeks to:

1. Test the predictive accuracy of existing urban building energy models.
2. Propose and investigate a scenario that considers the building network in an urban context and reduces building dataset requirements.
3. Learn how to reduce the required number of urban buildings to predict the energy use of a neighborhood or predict the energy demand of one building when its energy use information is unavailable.

### **1.4 Background**

#### **1.4.1 Building energy consumption in urban contexts**

Buildings consume up to 75% of total primary energy use in cities (IMT, n.d.). They account for about 28% of total energy-related carbon dioxide emissions, two-thirds of which come from burgeoning energy consumption (IEA, 2019); in particular, electricity is the dominant driver. The latest Electric Power Monthly data, reported by the U.S. Department of Energy (DOE) in May 2020, show that commercial and residential buildings in the U.S. account for 73.5% of electricity consumption (*Electric Power Monthly with Data for January 2018*, 2018). China's building energy consumption increased by nearly 45% in the 20 years after 1990, from  $30 \times 10^{18}$  joules in 1990 to  $43 \times 10^{18}$  joules in 2010 (Berardi, 2017). At the same time, cities around the world began setting targets to reduce greenhouse gas emissions for a low environmental impact and sustainable response to climate change (Sokol et al., 2017). With its rapid economic development and rapid urbanization, China has become the world's largest CO<sub>2</sub> emitter (Jefferson, 2015). Between 2006 and 2010, the government of China set a target of 20% energy savings; building energy efficiency is one of the priority projects, requiring strict implementation of building energy efficiency codes for new buildings. Specific measures include banning the use of energy-inefficient building materials and promoting the use of renewable energy systems in new buildings, as well as the development of green and efficient buildings (Kong et al., 2012; Wei et al., 2018).

#### **1.4.2 Energy consumption prediction methods for urban building**

For energy consumption in cities, in addition to top-down policies, bottom-up energy studies are also essential for energy planning, management, and conservation. In terms of building energy consumption research, the prediction of energy use in buildings can effectively help architects, urban planners and policy makers to better understand the characteristics of building energy

consumption in an urban context and the factors influencing it, so as to achieve the sustainable goal of energy saving and emissions reduction.

The bottom-up model localizes the energy use study and considers larger scale attributes (such as community or city) at the micro scale. The model extrapolates individual energy end-use estimates to cities, regions, and even countries. Bottom-up approaches fall into two broad categories: simulation-based engineering models and data-driven models. Simulation-based models are easily adaptable and can be combined with, for example, building information modeling (BIM) to help designers evaluate different energy efficiency strategies at an early stage of the design process. The main limitation of simulation-based models is the oversimplification of the urban environment and the influences of residential and urban microclimate effects (Abbasabadi & Ashayeri, 2019).

In recent years, in addition to engineering models for energy simulation, some researchers have recognized the importance of data-driven urban building energy predictive models (UBEPM) on a large scale with distributed building groups (Hsu, 2015; Kontokosta & Tull, 2017; Sun et al., 2018), which can potentially provide insights into large-scale building energy use patterns and opportunities to save energy (W. Li et al., 2017; Reinhart & Cerezo Davila, 2016). This development has become the center of attention in urban building energy use modeling.

### **1.4.3 Review of data-driven studies for urban building energy modeling (UBEM)**

Traditional energy models for single buildings focus on developing system-level reproductions of physical buildings (Kadir Amasyali & Nora M. El-Gohary, 2018; Zhao & Magoulès, 2012). In earlier studies, data-driven models were often applied to individual buildings. For example, Guo et al. used a machine learning-based model to predict the energy demand for building heating (Guo

et al., 2018). Wei et al. applied system recognition and neural networks to predict office buildings' energy use and occupancy (Wei et al., 2019). Fan et al. evaluated a short-term building energy prediction strategy based on deep recurrent neural networks (Fan et al., 2019). Wang et al. used long short-term memory (LSTM) to predict office buildings' internal heat gain (Z. Wang et al., 2019). In addition to energy consumption predictions for buildings, some researchers have also studied the prediction of coupled renewable energy systems in buildings, such as wind and solar. Raza et al. proposed an ensemble framework of five predictions for PV-integrated buildings and provided accurate seasonal monthly predictions for smart buildings with rooftop PV (Masaki & Zhang, Lijun, Xia, 2018). González-Aparicio and Zucker provided a regression method to predict wind power generation with a three-year dataset (González-Aparicio & Zucker, 2015).

As technological innovations and data sources have increased, data-driven research has focused more on large-scale building energy forecasting. For example, Heiple and Sailor estimated daily building energy consumption using annual building simulations of prototype buildings in Texas and matched them with prototype simulation outputs of existing buildings (Heiple & Sailor, 2008). Chen et al. developed an automated UBEM generation and simulation tool that considers neighborhood shadows for urban-scale building renovation analysis (Chen et al., 2017). Li et al. analyzed 51 high-performance commercial buildings in the United States, Europe, and Asia through portfolio analysis and case studies (C. Li et al., 2014). Kontokosta and Tull proposed a data-driven model for predicting city-scale buildings' energy consumption by analyzing 20,000 buildings and comparing three machine learning algorithms, including ordinary least-squares, support vector regression, and random forest (Kontokosta & Tull, 2017). Using six regression and machine learning algorithms for urban energy forecasting, Deng et al. reported a 10–15% reduction

in error compared to statistical models (Deng et al., 2018). Robinson et al. used a small number of building features in a machine learning algorithm for energy consumption prediction and validated the algorithm using the energy consumption dataset required by New York Local Law 84 (LL84) (Robinson et al., 2017). Fonseca and Schlueter proposed a spatiotemporal model for consumption patterns in neighborhoods and urban areas (Fonseca & Schlueter, 2015). The model allows the calculation of power and temperature requirements in the residential, commercial, and industrial sectors through spatial (building location) and temporal (hourly) dimensional analysis. Kalogirou et al. used electricity data from 225 buildings and applied the backpropagation algorithm to predict the required thermal load (Kalogirou et al., 1997). Hawkins et al. applied statistical and artificial neural network (ANN) methods to identify the determinants of energy consumption in UK university buildings, resulting in an average absolute prediction error of 34% and 25% for electricity and heating fuels, respectively. Kavgic used Monte Carlo methods to predict the space heat energy of Belgrade's housing stock (Miroslava Kavgic et al., 2015) as well as sensitivity analysis of uncertainty in a city-scale residential energy model (M. Kavgic et al., 2013).

However, data-driven models rely on large amounts of actual energy data, and the accuracy of the model is also dependent on the availability and the quality of data. The number of datasets at building and household levels is limited (Abbasabadi & Ashayeri, 2019); this is the main challenge to data-driven models. For the newly developed neighborhood in county-level cities of China, the historic building level energy data is even more difficult to obtain. Challenges also include the variations in building energy use on the urban scale, both spatially and temporally, which make the model more complex. Another obstacle is accurately predicting building energy use on the

urban scale with limited data sources and defining and integrating inter-network (or inter-effect) of building groups in urban context in UBEPM.

## CHAPTER 2: METHODOLOGY

### 2.1 Data description & data processing

#### 2.1.1 Building energy data

Electricity consumption data were obtained from the Department of Power Supply in Jianhu City. The data sets include yearly data from 2016 to 2018 for residential buildings and monthly data from 2015 to 2018 for public buildings. This project uses only historical electricity consumption data from public buildings. Energy use intensity (EUI), per area and per month, was selected to quantify the target variable of this study. EUI is usually calculated using the total energy usage and building floor area (S) as follows:

$$EUI_{building,S} = \frac{E_{building}}{\sum S} \quad (2.1)$$

where  $EUI_{building,S}$  is the EUI at the building level,  $E_{building}$  is the electricity usage, and S is the total building floor area.

Table 2.1 Building energy consumption

Indicator	Variables	Description	Measure	Source
Building Energy Consumption	EUI	The monthly electricity consumption of buildings from 2015-2018.	kWh/m <sup>2</sup>	Department of Power Supply in Jianhu City

Although the original data set contained 153 public buildings, only 71 buildings had four full years (2015–2018) of monthly electricity consumption data. Only these 71 buildings were retained for this study to ensure the longest possible time series data and the same period of energy consumption for all buildings used as model input data.

Table 2.2 shows the maximum, minimum, and mean monthly EUI of the 71 public buildings, and Figure 2.1 shows their distribution. These buildings' maximum monthly EUI, mean monthly EUI, and minimum monthly EUI are concentrated at 0–50 kWh/m<sup>2</sup>, 0–20 kWh/m<sup>2</sup>, and 0–5 kWh/m<sup>2</sup>, respectively. Figure 2.2 shows the monthly EUI distribution for each building from 2015 to 2018.

Table 2.2 Monthly EUI (kWh/m<sup>2</sup>) of all buildings

	Max EUI	Mean EUI	Min EUI
mean	13.235	5.727	2.154
median	5.313	2.450	0.760
std	31.688	12.971	5.991
min	0.451	0.220	0.017
max	239.933	86.539	39.782

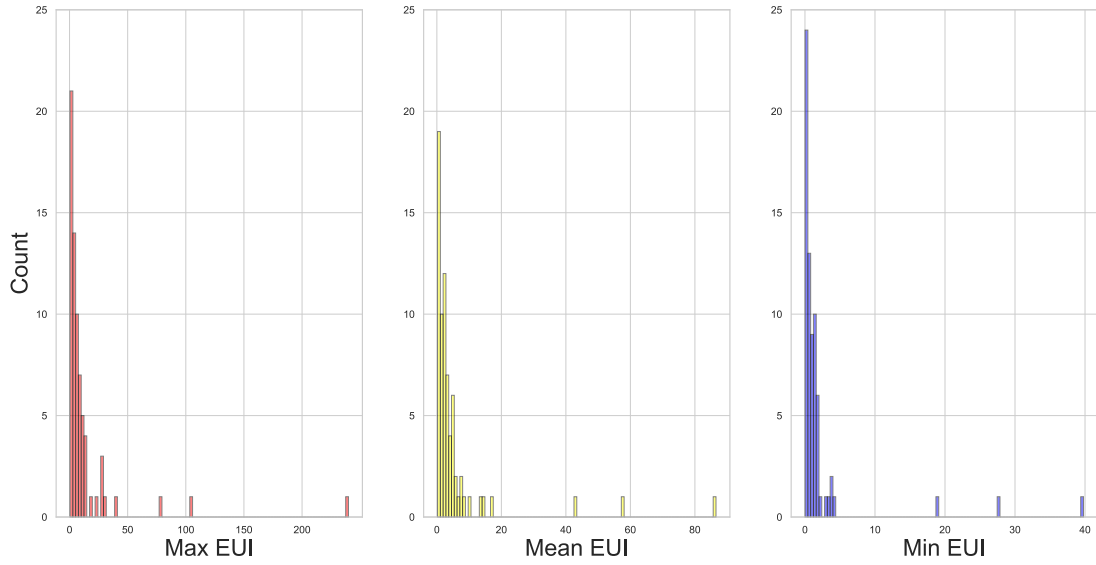


Figure 2.1 Monthly EUI (kWh/m<sup>2</sup>) distribution of all buildings

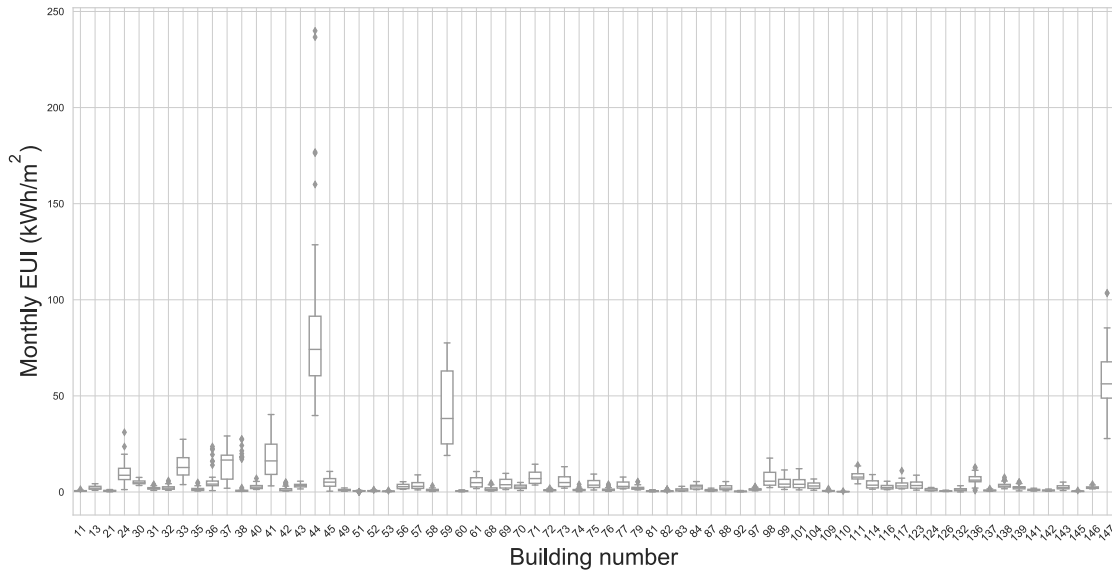


Figure 2.2 Monthly EUI (kWh/m<sup>2</sup>) distribution of each building

To see the electricity consumption trend more clearly for each month during the four-year period, all EUI data were normalized between 0 and 1 in terms of buildings using Equation 2.2 before plotting the change in electricity consumption for each month.

$$EUI_{i,j}^{normalized} = \frac{EUI_{i,j} - EUI_i^{min}}{EUI_i^{max} - EUI_i^{min}} \quad (2.2)$$

where  $EUI_{i,j}^{normalized}$  is the normalized EUI of the  $j$ th month of the  $i$ th building,  $EUI_{i,j}$  is the actual monthly EUI of the  $j$ th month of the  $i$ th building,  $x_i^{min}$  is the minimum value of monthly EUI in four years for the  $i$ th building, and  $x_i^{max}$  is the maximum value of monthly EUI in four years for the  $i$ th building.

Figure 2.3 reflects the trend of monthly electricity consumption for each building, and Figure 2.4 reflects the distribution of normalized EUI for the 71 buildings in each month. The monthly electricity consumption shows a seasonal pattern for nearly all buildings, but their maximum and minimum values occurred at various times. Moreover, each building's normalized EUI distribution varies greatly between months, with some months having a relatively concentrated distribution of normalized EUI, while others have a more dispersed distribution. Figure 2.5 reflects the distribution of normalized monthly EUI for each building over the four-year period, illustrating that the difference in and distribution of monthly EUI varies between buildings.

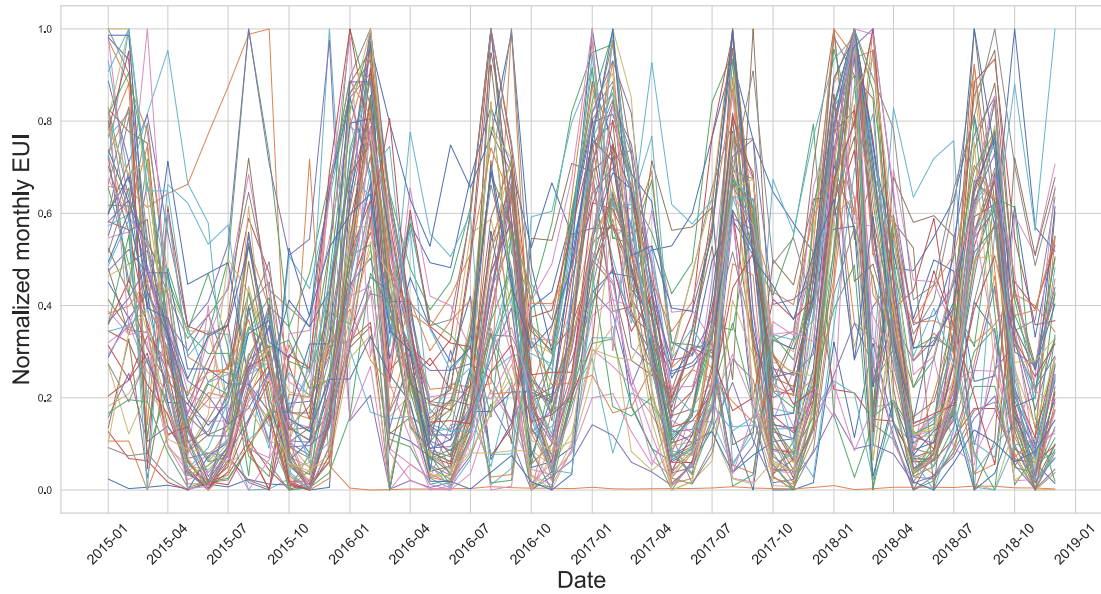


Figure 2.3 Normalized monthly EUI trend of each building

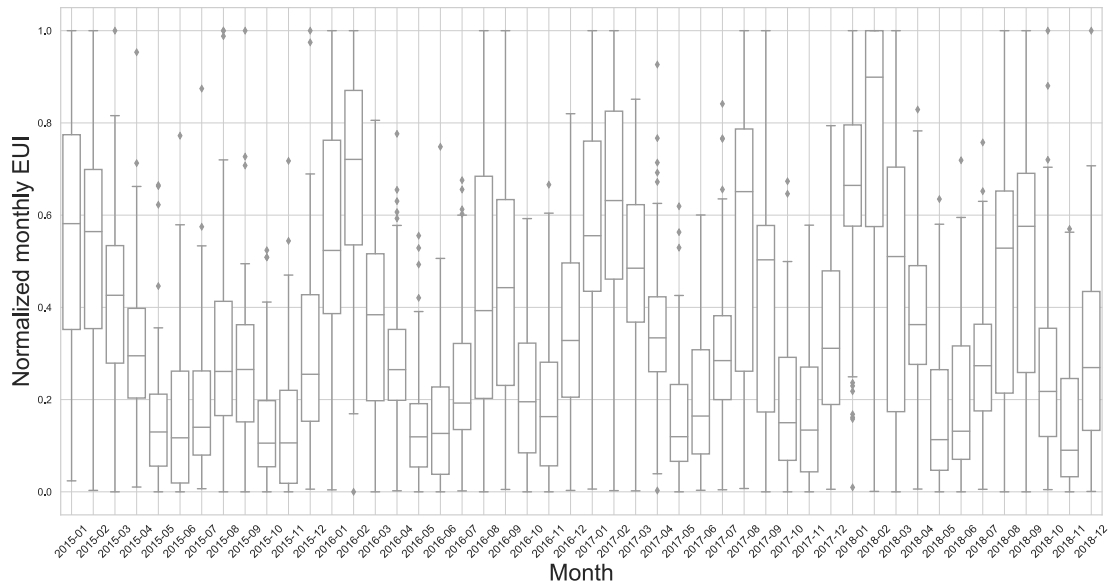


Figure 2.4 Normalized monthly EUI distribution of each month

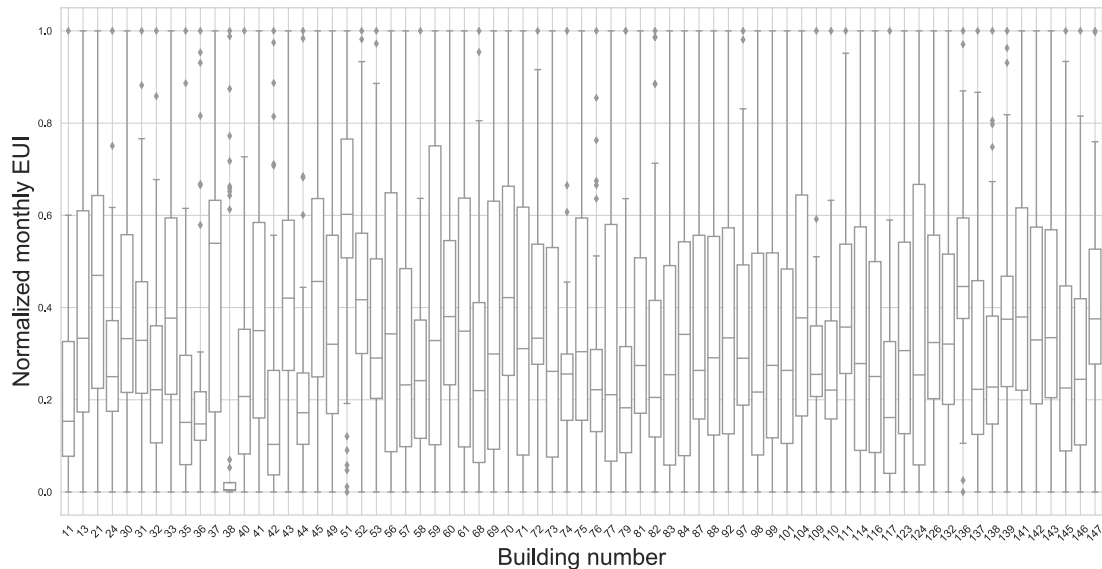


Figure 2.5 Normalized monthly EUI distribution of each building

### 2.1.2 Building design parameter data

The selected case city is Jianhu, a county-level city in Jiangsu Province, China, with an area of approximately 1160 km<sup>2</sup> and a population of around 0.8 million. Jianhu has a subtropical climate. Its elevation is about 2 meters above sea level, and its longitude and latitude are 33°16' to 33°41' and 119°33' to 120°05', respectively. The original dataset included 539 dwellings in 153 public buildings. This building information data were compiled and provided by the Institute of Building Technology and Science, Southeast University, China. The building parameters data set contains basic information and data on the inter-building effect (IBE), along with descriptions of building functions. The previous study by the client of this project explains the calculation process and description of the original building dataset and attributes (X. Li et al., 2020), and the location of

those buildings are shown in Figure 2.6. Only the 71 public buildings with completed four-year monthly historical electricity consumption data were used in this capstone project.

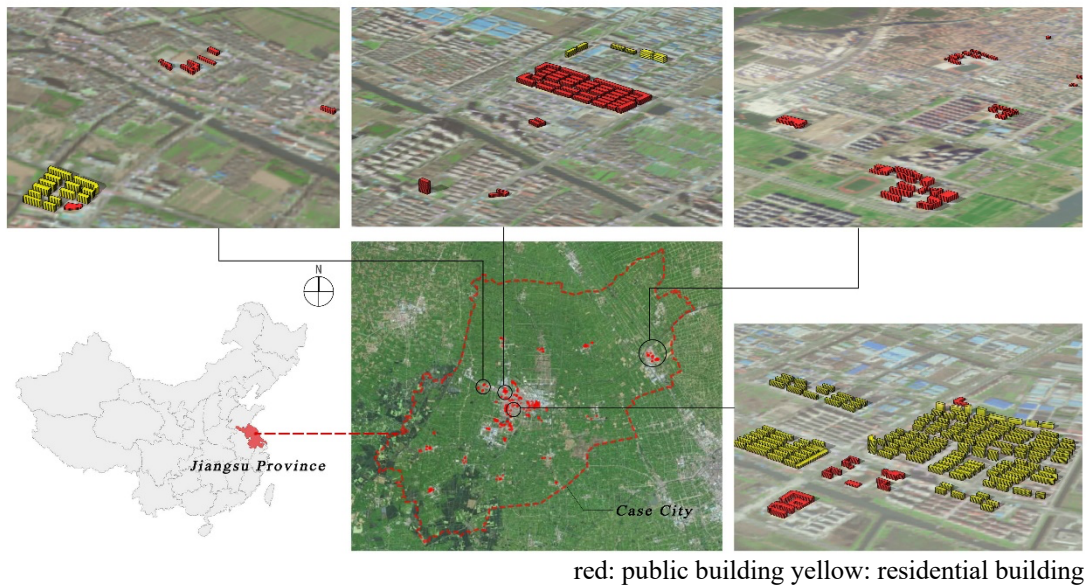


Figure 2.6 Location and layout of Jianhu City and the distribution of the available buildings in the original dataset

Table 2.3 Building design parameters

Indicator	Variables	Description	Measure	Source
Basic Building Morphological Information	floor_area	The total floor area of the building.	m <sup>2</sup>	Institute of Building Technology and Science, Southeast University
	floor_number	Number of floors in the building.	number	
	height	The height of the building.	m	
	short	The width of the building.	m	
	long	The length of the building.	m	
	perimeter	The perimeter of the building.	m	
	surface	The surface area of the building.	m <sup>2</sup>	
	volume	The volume of the building.	m <sup>3</sup>	
	Orientation	The orientation of the building.	degree	
Inter-building effect (IBE)	HW_South	The ratio of the obstruction height to canyon width of the south direction.	ratio	
	HW_West	The ratio of the obstruction height to canyon width of the west direction.	ratio	
	HW_North	The ratio of the obstruction height to canyon width of the north direction.	ratio	
	HW_East	The ratio of the obstruction height to canyon width of the east direction.	ratio	
	S_V	The shape coefficient of the building.	ratio	
	P_A	The perimeter to area ratio of the building.	ratio	
	BAR	The building aspect ratio of the building.	ratio	
Building Function	function	The function of the building.		

Table 2.4 contains statistical information on the buildings' design parameters of the 71 public buildings. To visualize the distribution of these data, some basic geometric parameters of the buildings and key inter-building effects are plotted. As shown in Figure 2.7, the buildings' width and height distributions are relatively concentrated, while the buildings' length distribution is relatively discrete. Figure 2.8 shows that the shape coefficient and perimeter-to-area ratio of these buildings are mostly distributed between 0.28–0.40 and 0.2–0.27, respectively. These indicators

will impact a building's indoor and outdoor heat exchange (X. Li et al., 2020). Figure 2.9 shows the height to canyon width of the buildings in four directions, which, to some extent, can reflect the differences in the buildings' urban environment. As seen in the figure, the differences between the four directions are minimal. Figure 2.10 shows the orientation of the buildings, most of which are facing due south or nearly due south.

Table 2.4 Building design parameters statistics summary

	mean	median	std	min	max
floor_number	4.127	4.000	2.918	1.000	18.000
Height (m)	12.353	11.475	8.742	3.000	54.000
floor_area (m <sup>2</sup> )	806.522	488.825	1142.954	158.270	7549.180
short (m)	9.739	9.689	4.647	1.790	24.840
long (m)	40.040	33.540	24.626	15.850	190.630
Perimeter (m)	137.429	102.110	95.325	55.950	493.920
surface (m <sup>2</sup> )	2446.024	1757.180	2150.222	327.590	11640.250
volume (m <sup>3</sup> )	9468.223	5865.953	11622.254	474.810	67942.660
orientation	12.659	0.000	37.269	-30.000	131.000
HW_South	0.615	0.031	1.469	0.000	9.000
HW_West	2.765	0.232	14.352	0.000	119.541
HW_North	1.558	0.173	5.299	0.000	38.679
HW_East	1.444	0.210	4.576	0.000	31.083
S_V	0.358	0.330	0.169	0.145	1.465
P_A	0.244	0.230	0.124	0.060	1.130
BAR	5.947	3.950	5.660	1.150	32.660

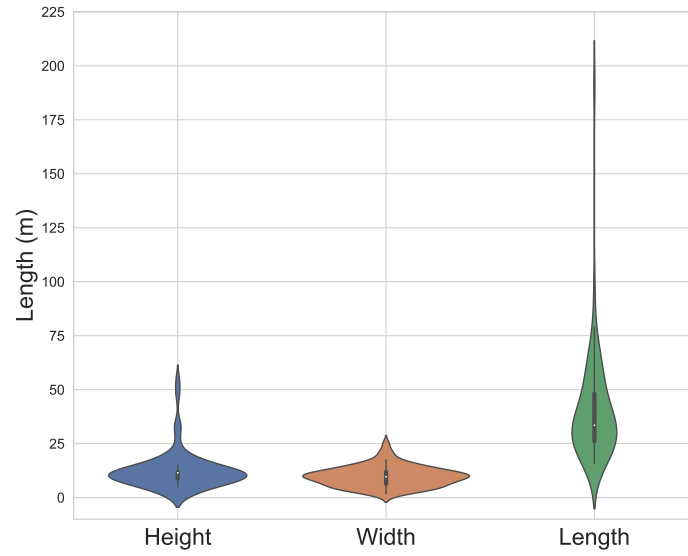


Figure 2.7 Basic geometry of buildings

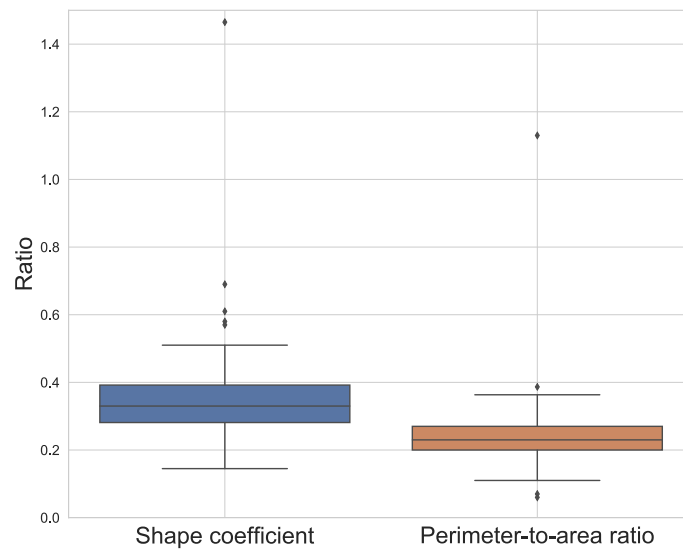


Figure 2.8 Building shape factors

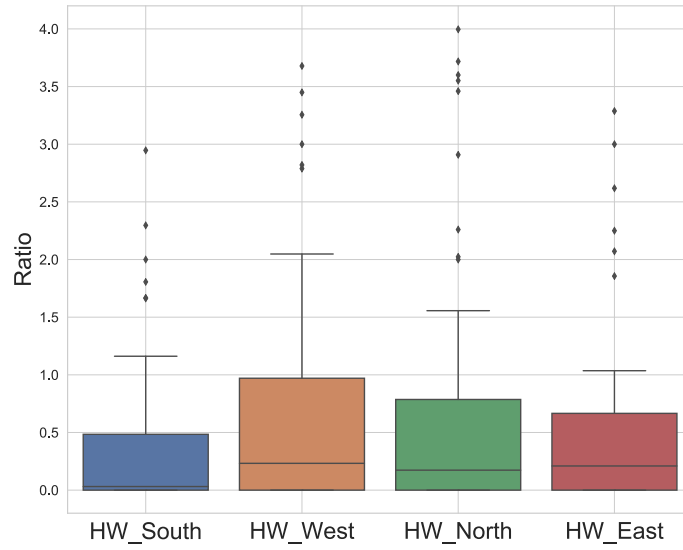


Figure 2.9 Obstruction height to canyon width

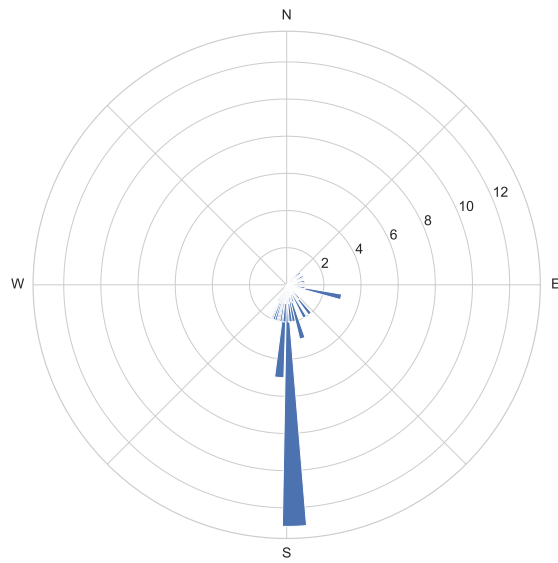


Figure 2.10 Building orientation

In terms of function, the 71 buildings are mostly used for public services (50.6%), followed by education (21.1%), industry (15.5%), hospitals (9.9%), and hotels (2.8%).

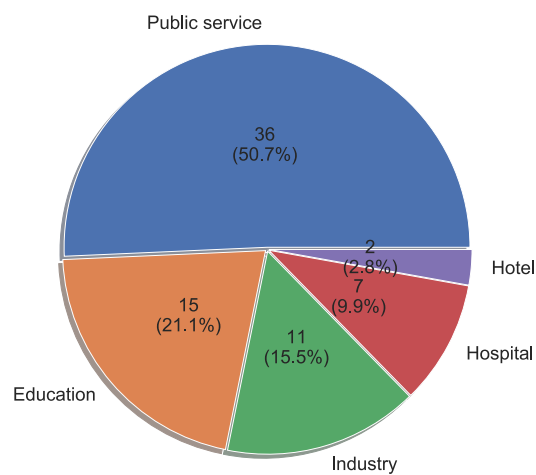


Figure 2.11 Building function

## 2.2 Data-driven predictive techniques

### 2.2.1 K-nearest neighbors

K-nearest neighbor regression (KNN) is a non-parametric method that makes predictions based on the target output of  $K$  nearest neighbors of a given query point. Specifically, given a data point, the distance between that point and all points in the training set is calculated. Then the closest  $K$  training data points are selected, and the predicted value is set to the average of the target output

values of these  $K$  points (Ahmed et al., 2010a). KNN is an effective non-parametric algorithm that is widely used in classification and regression. In addition to its simple and intuitive recovery from large-dimensional feature spaces and its tolerance of high-dimensional and incomplete data, KNN is often used in time series-related predictions (Ban et al., 2013). The `KNeighborsRegressor` in the `scikit-learn` package (Pedregosa et al., 2011) is used in this study to build the model. In the model, the “weight points” are calculated by the inverse of their distance, which means that the closer query points have greater influence than the distant ones.

### **2.2.2 Support vector regression**

Support vector machines (SVMs) are used in many machine learning tasks, such as pattern recognition, object classification, and regression analysis, as well as in time series forecasting (Sapankevych & Sankar, 2009). Support vector regression (SVR) is a method of “training” an SVM using observation data to estimate functions. The forecasting of financial data time series and power loads are the most common practical applications for SVR. SVR is a successful approach based on the use of a high-dimensional space created by converting the initial variables and adjusting the resulting complexity using a penalty term applied to the error function (Ahmed et al., 2010b). Through minimizing hinge losses, an SVM ignores training data that are very close to the actual results but models a boundary that includes as many samples as possible to optimize the model reliability. In this study, the SVR function of the `scikit-learn` package (Pedregosa et al., 2011) in Python is used to build a regression model.

### **2.2.3 Long-short memory network algorithm**

The algorithm of Long Short-Term Memory (LSTM) networks is an improved version of recurrent neural networks (RNN) (Hochreiter & Schmidhuber, 1997), making it easier to remember past data. LSTM is very suitable for classifying, processing, and predicting time series, given a time lag of unknown duration. LSTM network algorithm is usually composed of an input layer, a hidden layer, an output layer, a context layer, and a forget layer. In this study, as the building energy usage data are typical temporal datasets (time-series data), LSTM is extremely suitable for prediction. The input and output layers are assumed to be  $X$  and  $Y$ , respectively.  $S$  is the state layer, and there is also one context layer ( $C$ ) to store feedback signals for the state layer in the next interval and one forget layer ( $F$ ) to lead the information, which should be formed based on the current input and previous output:

$$y_t = f(W * x_t + U * h_{t-1} + b) \quad (2.3)$$

where  $x_t \in X$ ,  $y_t \in Y$ , and  $y$  donates the output of one layer and includes  $f_t, i_t, z_t, o_t$ , and  $f$  is the activation function for each layer.  $W$  donates the weights of each layer as  $W_f, W_i, W_z, W_o$ , and  $U$  donates the weights for the last state as  $U_f, U_i, U_z, U_o$ .  $b$  donates the bias of each layer as  $b_f, b_i, b_z, b_o$ . The subscripts (f, i, z, o) represent the forget layer, the output and the state of the input layer for next hidden layer, and the output of the hidden layer. The subscript  $t$  indexes the time step. The output of the context layer can be a function (Equation 2.4) of the forget layer, the context layer at the previous timestep, and the input of the hidden layer. The output of the hidden layer  $h_t$  can be updated as Equation 2.5.

$$c_t = f_t * c_{t-1} + i_t * z_t \quad (2.4)$$

$$h_t = o_t * \tanh c_t \quad (2.5)$$

In this study, the LSTM function was created through Python Tensorflow package.

## 2.3 Prediction models

The urban building energy dataset is usually spatial and temporal. Besides the time-series energy dataset, urban buildings inherently have their own morphological characteristics and inter-relationships between buildings in the urban context, e.g., building shading from neighbors. Therefore, the dataset for urban buildings could be the assembly of energy use, their characteristics and relationship. This study proposed four typical data-driven predictive models for urban building energy prediction. In all the models, the SVR, KNN, and LSTM predictive techniques were applied to determine which is more suitable.

### 2.3.1 Time-series predictive model for individual building energy use

Model 1 is the most traditional and fundamental of the time-series prediction models for building energy consumption. In this study, Model 1 is used to verify the feasibility and accuracy of the traditional building energy forecasting model in the context of county-level cities in China. It is suitable for predicting the energy consumption of a single building over time in scenarios where historical energy consumption data of the building is available. Furthermore, Model 1 is also used

as the reference model for the next two models (Model 2 and Model 3) created with urban building networks.

Model 1 is the most common time-series predictive energy model; as such, this study format inputs historical energy data of 71 public buildings and runs the model for predicting individual building energy use. Shown in Figure 2.12, a five-month time window was set up to create a series of moving energy inputs; therefore, the input can be formatted as Equation 2.6. The previous 42-month EUI data was used for model training and the six-month EUI data for model validation. The EUI values of the first 42 months were used as the training set, and the EUI value of the next six months was used to verify the accuracy of the model.

$$x_{t,i} \leftarrow (x_{1,i}, x_{2,i}, x_{3,i}, \dots, x_{t-1,i}), i = 1, 2, 3, \dots, N \quad (2.6)$$

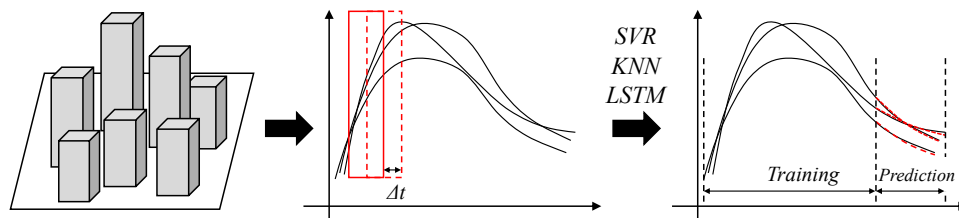


Figure 2.12 The schematic diagram of Model 1

### 2.3.2 Building-network predictive model for individual buildings based on the similarity of building energy consumption curves

Model 2 has similar data requirements to Model 1. The difference is that Model 2 introduces building-network, and the prediction of a single building is based on all the buildings in the network where it is located. In Model 2, Equation 2.7 was used to define and abstract the building network by calculating the similarity of the energy use between the buildings, and the threshold was set to  $\geq 0.60$ . Since the concern in Model 2 is the comparison of energy consumption variations, cosine similarity is useful in that it bases its calculation on the angle rather than the difference in the actual values (Euclidean distance) of the data objects (Prabhakaran, 2018). Therefore, two building types are defined: target building and similar building. Similar buildings that have more than 60% similarity to target buildings are applied to predict each building's energy use. If the number of similar buildings exceeded five, only the five most similar buildings were selected. All the buildings were then calculated in a loop by applying model 1 for time-series energy prediction (Equation 2.8).

Model 2 was also used as two validations: 1) whether the prediction accuracy of a single building can be improved by building-network, which is created by building energy consumption data, given the same data as in Model 1; and 2) to verify whether the prediction of a building can be made from other buildings in the network if the historical energy consumption data for that building is missing.

$$R_{i,j} = \text{Similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n \sum_{j=1}^n A_i B_j}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{j=1}^n B_j^2}} \quad (2.7)$$

$$x_{t,i} \leftarrow (x_{1,j}, x_{2,j}, x_{3,j}, \dots, x_{t-1,j}) \quad j \in [0, n] \quad (2.8)$$

$$n \leq N$$

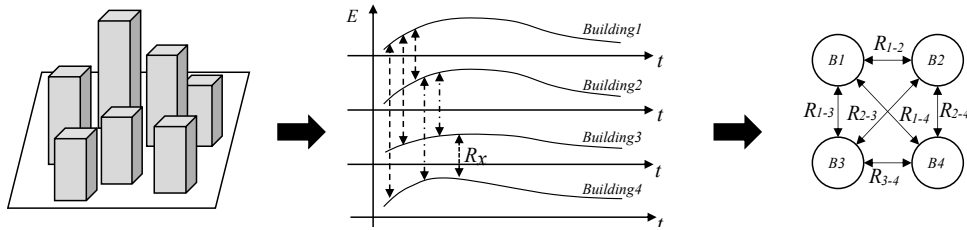


Figure 2.13 The schematic diagram of Model 2

### 2.3.3 Building-network predictive model for individual buildings based on the similarity of building morphological data

Model 3 and Model 2 have the same idea—predicting the energy consumption of individual buildings through building-network. The difference is that the building-network in Model 3 is established based on the morphological data of the building. Model 3 was used as two validations: 1) whether the prediction accuracy of a single building can be improved by building-network, which is created by building morphological data; and 2) to verify whether the prediction of a building can be made from other buildings in the network if the historical energy consumption data for that building is missing, which is similar to Model 2.

The building network in Model 3 is abstracted by calculating the similarity of the morphological data between the buildings. The morphological data consists of the basic form design parameter of each building and inter-effect between buildings. The Euclidean distance method was applied to calculate the similarity between buildings. The threshold was set to be lower than or equal to 0.23. If the number of selected buildings exceeded five, only the five most similar buildings were selected. The next predictive procedure is the same as in Model 2.

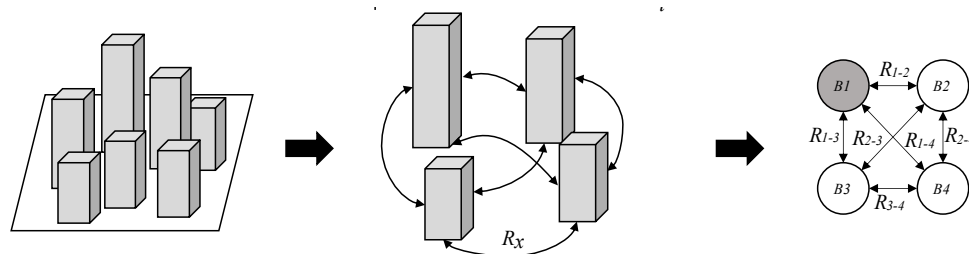


Figure 2.14 The schematic diagram of Model 3

### 2.3.4 Time-series total energy predictive model for all buildings

Model 4 is proposed to predict the total energy use of all buildings in the neighborhood. Models 1–3 are concerned with the prediction of energy consumption for a single building. Model 4 focuses on the overall building energy consumption trends and changes within a particular area. First, the monthly EUI data from 71 buildings was used to predict total energy use, which is the sum of all buildings' energy use. The result of this prediction is used as the reference for the total energy use of all the buildings.

In the following steps, an interval of 5% is set up to reduce the number of buildings required—that is, 95%, 90%, 85%, 80%, 75%, 70%, 65%, and 60% of the buildings are randomly selected to

predict the total energy consumption of 71 buildings. To randomly select the buildings, each case was run randomly 100 times to eliminate random errors. The prediction procedure and technique are the same as Model 1. This model is intended to provide a potential way to figure out how many buildings' data are required for total energy prediction for buildings and to find the opportunity to reduce the required number of buildings.

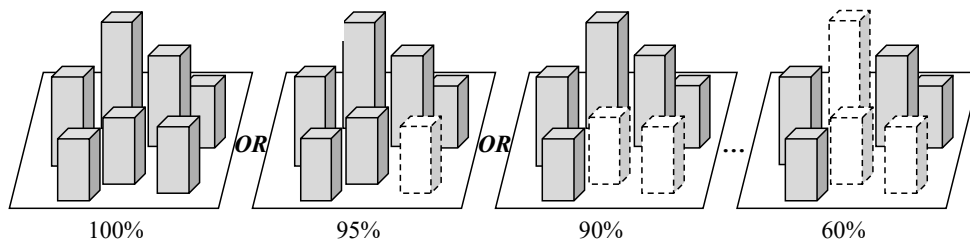


Figure 2.15 The schematic diagram of Model 4

## 2.4 Model parameter tuning and evaluation

In the model evaluation, the performance of the mean square error was used to adjust the parameters for the three algorithms to determine the best fit. For the SVR algorithm, a kernel function was adjusted in a group of linear, rbf, sigmoid. Gamma was looped from 0.1 to 5.0 with a step of 0.1, and C was set to  $1e-3$ ,  $1e-2$ ,  $1e-1$ , 1, 2, 4, 10, 100, and 1000. For LSTM, the activation function is 'tanh' in the package and optimizer is the Adam function. The learning rate was set as 0.01 and 0.001, the batch size was 16, 32, and 72, and epochs were 10, 50, 100, 500, and 1000. The general evaluation metrics included the root mean square error (RMSE) and R-squared ( $R^2$ ). They are formatted as follows:

$$RMSE(EUI) := \sqrt{\sum_{n=1}^N (EUI_i^{actual} - EUI_i^{predicted})^2 / N} \quad (2.9)$$

$$R^2 = 1 - \frac{\sum_{n=1}^N (EUI_i^{actual} - EUI_i^{predicted})^2}{\sum_{n=1}^N (EUI_i^{actual} - \overline{EUI^{actual}})^2} \quad (2.10)$$

## CHAPTER 3: RESULTS AND DISCUSSION

### 3.1 Energy prediction results for Model 1

The performance of the three algorithms on the testing dataset was used to evaluate the accuracy of the model. The time-series energy model was the first to be analyzed for individual buildings with three predictive techniques – SVR, KNN, and LSTM. Table 3.1 shows the results of each technique with Model 1. In terms of  $R^2$ , among the 71 buildings, all three techniques performed sufficiently. LSTM was the most accurate data-driven technique for an individual building, considering both  $R^2$  and RMSE. The results of Model 1 were also used as the baseline model for Models 2 and 3.

Table 3.1 Model 1 results

	KNN	SVR	LSTM
$R^2$	0.871	0.909	0.927
RMSE (kWh/m <sup>2</sup> )	4.198	3.520	3.162

### 3.2 Energy prediction results for Models 2 and 3

In this study, Models 2 and 3 were proposed to create a building network in the urban context to enhance energy prediction. To predict the energy use of the target building, two cases were tested with each model. The first case (Self) integrated the historical energy data of similar buildings in networks and the target building as input. The second case (Noself) involved only the historical energy data of similar buildings. Tables 3.2 and 3.3 show the performance of the three predictive

techniques for Models 2 and 3 on the testing dataset, respectively. Both the building networks incorporating energy consumption similarity (Model 2) and morphology similarity type (Model 3) have higher prediction accuracy in the first case (self) than in the second case (noself). As expected, the availability of historical energy consumption data for the target building likely improves the prediction accuracy. The different algorithms perform differently in the two models. KNN is more suitable for Model 2, while LSTM is more suitable for Model 3, and SVR has little difference in performance in the two models.

Table 3.2 Model 2 results

	KNN		SVR		LSTM	
	self	noself	self	noself	self	noself
R <sup>2</sup>	0.910	0.857	0.916	0.878	0.914	0.863
RMSE (kWh/m <sup>2</sup> )	3.503	4.418	3.388	4.085	3.426	4.325

Table 3.3 Model 3 results

	KNN		SVR		LSTM	
	self	noself	self	noself	self	noself
R <sup>2</sup>	0.850	0.849	0.912	0.879	0.941	0.919
RMSE (kWh/m <sup>2</sup> )	4.523	4.551	3.478	4.074	2.830	3.320

To determine whether the prediction accuracy can be improved or the target building's historical energy consumption is required, the two cases using Models 2 and 3 are compared with Model 1.

As shown in Figure 3.1, for the first case compared to Model 1 performance, the KNN algorithm has more improved accuracy with Model 2, while the LSTM algorithm has a larger accuracy

improvement with Model 3. Taken together, the energy consumption of a single building by establishing a building network can be predicted more accurately by some algorithms, with the LSTM algorithm yielding the best result in Model 3, whose  $R^2$  and RMSE are 0.941 and 2.830 kWh/m<sup>2</sup>, respectively.

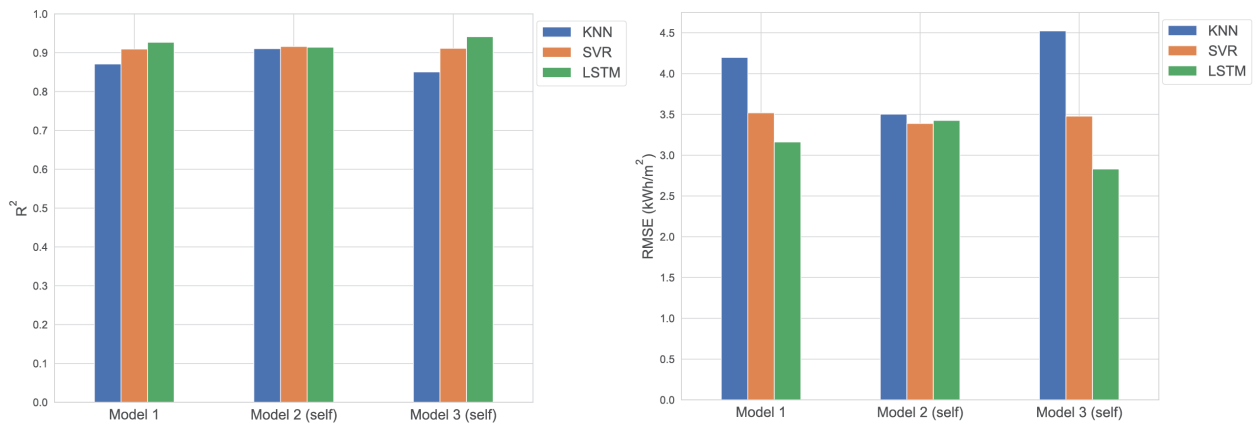


Figure 3.1 The first case of Model 2 and Model 3 compared with the results of Model 1

The second case is the prediction without historical energy consumption data of the target building. Figure 3.2 indicates that the accuracies of the KNN and SVR algorithms decrease more significantly, while the LSTM is not significantly impacted. The LSTM algorithm achieves the best prediction accuracy in this case, which is better than that of the KNN and SVR algorithms in Model 1 and comparable to the performance of the LSTM algorithm itself in Model 1.

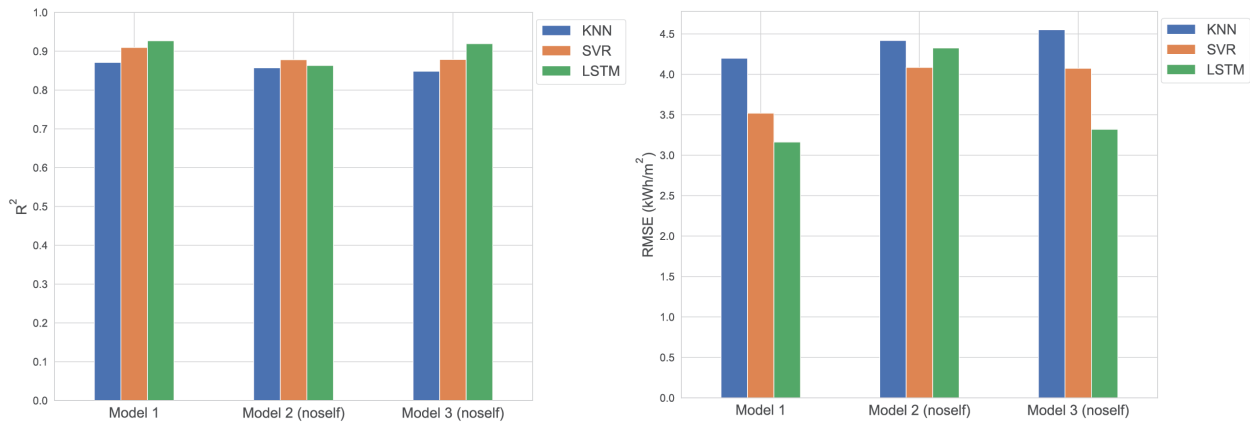


Figure 3.2 The second case of Model 2 and Model 3 compared with the results of Model 1

### 3.3 Energy prediction results for Model 4

Model 4 was used to investigate the difference in performance when predicting the total energy consumption of all buildings when varying the number of buildings (percent of the total) that provide historical energy consumption data as input. Since the buildings were randomly selected, the model was simulated 100 times for each case to reduce random errors, except for the 100% case, which was run only once.

Figures 3.3 and 3.4 show the distribution of  $R^2$  and RMSE, respectively, for the three algorithms on the testing dataset at different percentages. The performance of the SVR algorithm is clearly worse than that of KNN and LSTM in all cases.

However, the performances of KNN and LSTM have their merits. The figures together with Table 3.4 and Table 3.5 indicate that the LSTM algorithm yields more accurate results than the KNN algorithm in the best performance in each case. Still, the KNN algorithm performs better overall with respect to  $R^2$ , with the median of  $R^2$  higher than those of the LSTM algorithm in various cases

with missing some building historical energy consumption data. Regarding RMSE, LSTM obtains higher accuracy than KNN in the best performance in each case, which is the same as in  $R^2$  results, while KNN yields a lower median RMSE than LSTM in various cases with missing historical building energy consumption data. In the case where historical energy consumption data of all buildings are used as input (i.e., 100%), LSTM outperforms KNN.

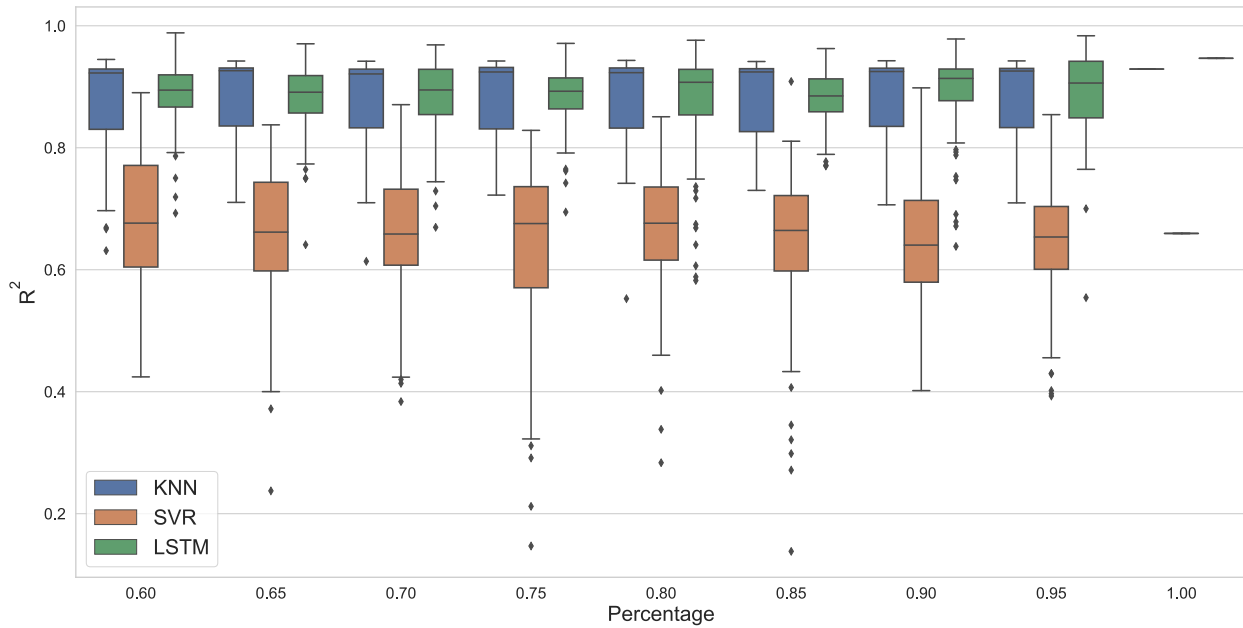


Figure 3.3  $R^2$  results of Model 4

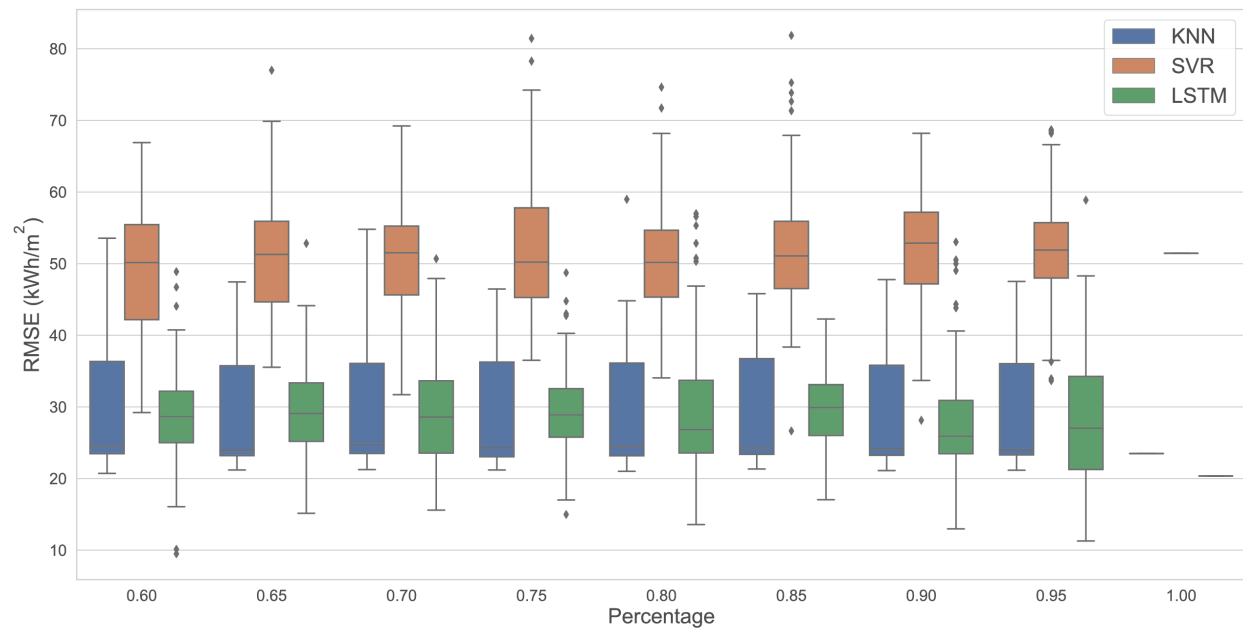


Figure 3.4 RMSE (kWh/m<sup>2</sup>) results of Model 4

Table 3.4 R<sup>2</sup> results of Model 4

Percentage		KNN	LSTM	SVR
0.6	max	0.945	0.988	0.890
	mean	0.877	0.888	0.682
	min	0.631	0.693	0.424
	std	0.071	0.050	0.114
0.65	max	0.942	0.970	0.838
	mean	0.894	0.884	0.657
	min	0.710	0.641	0.237
	std	0.055	0.054	0.116
0.7	max	0.942	0.969	0.871
	mean	0.876	0.884	0.662
	min	0.614	0.670	0.384
	std	0.066	0.058	0.111
0.75	max	0.942	0.971	0.829
	mean	0.888	0.883	0.635
	min	0.722	0.695	0.147
	std	0.057	0.049	0.140
0.8	max	0.943	0.976	0.851
	mean	0.884	0.875	0.665
	min	0.553	0.582	0.284
	std	0.063	0.085	0.107
0.85	max	0.941	0.963	0.909
	mean	0.882	0.883	0.640
	min	0.730	0.770	0.138
	std	0.060	0.041	0.122
0.9	max	0.943	0.978	0.898
	mean	0.893	0.894	0.649
	min	0.706	0.638	0.402
	std	0.055	0.065	0.090
0.95	max	0.942	0.984	0.854
	mean	0.892	0.889	0.647
	min	0.710	0.554	0.393
	std	0.055	0.070	0.100
1		0.929	0.947	0.660

Table 3.5 RMSE (kWh/m<sup>2</sup>) results of Model 4

		KNN	LSTM	SVR
0.60	max	53.545	48.880	66.904
	mean	29.873	28.743	48.872
	min	20.716	9.498	29.208
	std	8.202	6.609	9.226
0.65	max	47.450	52.830	77.002
	mean	27.915	29.268	50.911
	min	21.195	15.155	35.538
	std	6.849	6.799	8.586
0.70	max	54.801	50.685	69.218
	mean	30.089	29.243	50.522
	min	21.251	15.601	31.712
	std	7.827	7.083	8.590
0.75	max	46.461	48.733	81.438
	mean	28.611	29.484	52.377
	min	21.192	14.980	36.513
	std	7.152	6.091	9.658
0.80	max	58.980	56.986	74.637
	mean	29.113	29.657	50.448
	min	21.010	13.575	34.058
	std	7.398	9.491	7.963
0.85	max	45.809	42.270	81.856
	mean	29.329	29.728	52.220
	min	21.342	17.047	26.644
	std	7.369	5.379	8.449
0.90	max	47.775	53.031	68.198
	mean	28.063	27.639	51.732
	min	21.119	12.973	28.132
	std	6.818	7.829	7.051
0.95	max	47.516	58.872	68.703
	mean	28.204	28.017	51.832
	min	21.161	11.284	33.658
	std	6.860	8.860	7.528
1.00		23.487	20.357	51.452

### 3.4 Discussion

The evaluation of the results determines the feasibility and accuracy of the model. The first cases of Model 2 and Model 3 are of the building network created to improve the prediction accuracy when the historical energy consumption data of individual buildings are known. A comparison with the best results of Model 1 shows that building networks based on building morphological features can achieve better results with the LSTM algorithm, which outperforms all the algorithms in Model 1, which use the historical energy consumption data of only the target building for prediction. Therefore, in an urban environment, if the historical energy consumption data of the target building and other nearby buildings for the same time span are available, the energy consumption prediction accuracy of the target building can be effectively improved using the LSTM algorithm with a building network based on similar building morphological properties.

Model 3 exhibits higher effectiveness of the building network based on similar building energy consumption compared to Model 2. In the Self case, although the performance of Model 2 using KNN and SVR is improved over that of Model 1, the best outcome of the LSTM algorithm does not exceed that of Model 1. Additionally, in practical applications, acquiring historical building energy consumption data is more difficult than acquiring morphological building parameters. Specifically, the dataset needs to account for the method, frequency, and time of collection. The means of collection is relatively fewer, and the dataset requires regular updates. On the other hand, building design parameters are relatively easy to obtain, for example, from the architectural firm responsible for the design, the construction company, or the building management. The increasingly abundant satellite maps, open maps, and city models can also be sources of data. More

importantly, the morphological parameters of a building are do not change over time, except for major reconstruction. In summary, building networks created from building design or morphological parameters are more suitable for energy consumption prediction of individual buildings at the urban scale.

The results of Model 4 show that inputting the historical energy consumption of 60% to 95% of the total number of buildings achieves similar accuracy when predicting the total building energy consumption using the data of all buildings as input. In addition, the distribution of all the predictions at each percentage chosen at random shows little difference in performance between them. Thus, this model confirms that the prediction of total building energy consumption in a given area can be achieved effectively even without complete energy consumption data for all buildings. Furthermore, the LSTM is the optimal algorithm when complete data are available, and the KNN algorithm performs more consistently when complete data are not available.

## **CHAPTER 4: LIMITATIONS AND FUTURE WORK**

This study has several limitations. First, only three algorithms—KNN, SVR, and LSTM—were used, while no ensemble algorithm (e.g., Random Forest) was tested. Secondly, the acquired data were limited. Although complete design parameters for 71 public buildings and four years of monthly energy consumption data were available, the number of observations was small for a time series-based study, making it difficult to optimize the prediction accuracy of the regression model. Thirdly, the data contained historical energy consumption data and building morphological parameters, but no data related to human activities or other urban environmental context parameters, such as historical weather parameters, were applied. As stated in most energy studies, the uncertainty of occupancy significantly impacts the urban building energy model (W. Wang et al., 2019).

The prediction of energy consumption in a single building in an urban context demonstrates that building networks created using the similarity of building morphological parameters can improve the prediction accuracy when complete historical data are available. These models achieve prediction accuracies close to those that would have been achieved without considering building networks in the absence of historical data for the building being predicted. However, this performance improvement is only for the data-driven building energy prediction model and is comparable to results of other energy prediction methods, such as engineering simulation building energy models. In addition, the models were designed and tuned with respect to the overall performance of all buildings at the urban scale and are not optimized for a single building. Therefore, these models are not necessarily suitable for the study of the energy consumption of

individual buildings. When it is necessary to accurately predict the energy consumption of a specific building in a specific month, the error of the model may be very large.

In this study, only public buildings with complete monthly electricity consumption data were retained to ensure the longest possible time-series data. In future studies, the further examination of non-time series–related models using data from other buildings in the original dataset is planned, along with attempts at a more accurate prediction of building energy consumption in the absence of complete historical energy consumption data. Also, additional variables that potentially affect the fluctuation of building energy consumption may be incorporated in the model, and more algorithms will be tested to improve the overall model reliability. In addition, historical energy consumption data for longer periods and shorter temporal granularity are needed to improve the predictive power of the time series model. In this project, the prediction targets are actual values. However, in some cases at the city scale, for regional energy policy–making and energy management, the classification approach is more intuitive than the regression approach and can be used to effectively respond and adjust to trends in building energy consumption in the region.

## CHAPTER 5: CONCLUSIONS

The project proposes three models for predicting the energy consumption of individual buildings in an urban environment and one model for predicting the total building energy consumption in a specific area, using three algorithms: KNN, SVR, and LSTM.

In the model for single building energy consumption prediction, a building network based on energy consumption similarity and a building network based on morphology similarity are proposed. The performances of the models using these two building networks are compared with that of the traditional models that only consider the predicted building's historical energy consumption data. The results of the study show that a building network based on building morphological similarity can improve the overall performance of prediction models for the energy consumption of individual buildings in an urban context. Also, with this building network, the performance is close to that of a prediction model that only considers the predicted building's historical energy consumption data in the absence of the predicted building's historical energy consumption data.

In the model for total building energy consumption prediction, when predicting the total energy consumption of all buildings in a region with incomplete historical energy consumption data, accurate prediction results with data-driven prediction models are still possible.

The community and city-scale building energy prediction models explored in this capstone also have potential for practical applications. In small- and medium-sized cities where numerous commercial complexes and business parks are emerging, the understanding and accurate prediction of public building energy consumption in newly developed areas can help to formulate

sound urban energy efficiency and energy management policies. With the prediction models in this study, policymakers and researchers can use historical energy consumption data of buildings in similar building clusters or development areas in the city combined with morphological parameters of new buildings to predict the changes of overall building energy consumption in the area. These models can also be used as references for occupants and managers of new public buildings in the community to predict changes in building energy consumption, using historical energy consumption data from other buildings with similar building forms. This research project also recommends that city policymakers develop a reasonable policy to collect and disclose historical energy consumption data and appropriate design parameters of public buildings in cities. The result will be privacy protection, security, and an established comprehensive database, which can serve to urge public building managers to save energy and provide necessary references for research and related policy formulation.

In summary, the project provides a new perspective to associate urban morphology with building energy consumption data and improve the robustness and accuracy of the data-driven urban building energy model. The project also yields potential solutions to reduce the requirements for large volumes of data in urban energy models. Contributions to urban energy planning are made by revealing the energy demand of urban scales, especially for emerging cities or small- and medium-sized cities in developing countries undergoing urbanization.

## BIBLIOGRAPHY

- Abbasabadi, N., & Ashayeri, M. (2019). Urban energy use modeling methods and tools: A review and an outlook. *Building and Environment*, *161*, 106270. <https://doi.org/10.1016/j.buildenv.2019.106270>
- Ahmed, N. K., Atiya, A. F., Gayar, N. E., & El-Shishiny, H. (2010a). An Empirical Comparison of Machine Learning Models for Time Series Forecasting. *Econometric Reviews*, *29*(5–6), 594–621. <https://doi.org/10.1080/07474938.2010.481556>
- Ahmed, N. K., Atiya, A. F., Gayar, N. E., & El-Shishiny, H. (2010b). An Empirical Comparison of Machine Learning Models for Time Series Forecasting. *Econometric Reviews*, *29*(5–6), 594–621. <https://doi.org/10.1080/07474938.2010.481556>
- Ban, T., Zhang, R., Pang, S., Sarrafzadeh, A., & Inoue, D. (2013). Referential kNN Regression for Financial Time Series Forecasting. In M. Lee, A. Hirose, Z.-G. Hou, & R. M. Kil (Eds.), *Neural Information Processing* (pp. 601–608). Springer Berlin Heidelberg.
- Berardi, U. (2017). A cross-country comparison of the building energy consumptions and their trends. *Resources, Conservation and Recycling*, *123*, 230–241. <https://doi.org/10.1016/j.resconrec.2016.03.014>
- Chen, Y., Hong, T., & Piette, M. A. (2017). Automatic generation and simulation of urban building energy models based on city datasets for city-scale building retrofit analysis. *Applied Energy*, *205*, 323–335.

- Deng, H., Fannon, D., & Eckelman, M. J. (2018). Predictive modeling for US commercial building energy use: A comparison of existing statistical and machine learning algorithms using CBECS microdata. *Energy and Buildings*, *163*, 34–43. <https://doi.org/10.1016/j.enbuild.2017.12.031>
- Fan, C., Wang, J., Gang, W., & Li, S. (2019). Assessment of deep recurrent neural network-based strategies for short-term building energy predictions. *Applied Energy*, *236*(December 2018), 700–710. <https://doi.org/10.1016/j.apenergy.2018.12.004>
- Fonseca, J. A., & Schlueter, A. (2015). Integrated model for characterization of spatiotemporal building energy consumption patterns in neighborhoods and city districts. *Applied Energy*, *142*, 247–265. <https://doi.org/10.1016/j.apenergy.2014.12.068>
- González-Aparicio, I., & Zucker, A. (2015). Impact of wind power uncertainty forecasting on the market integration of wind energy in Spain. *Applied Energy*, *159*, 334–349. <https://doi.org/10.1016/j.apenergy.2015.08.104>
- Guo, Y., Wang, J., Chen, H., Li, G., Liu, J., Xu, C., Huang, R., & Huang, Y. (2018). Machine learning-based thermal response time ahead energy demand prediction for building heating systems. *Applied Energy*, *221*, 16–27. <https://doi.org/10.1016/J.APENERGY.2018.03.125>
- Heiple, S., & Sailor, D. J. (2008). Using building energy simulation and geospatial modeling techniques to determine high resolution building sector energy consumption profiles. *Energy and Buildings*, *40*(8), 1426–1436. <https://doi.org/10.1016/J.ENBUILD.2008.01.005>

- Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Hsu, D. (2015). Comparison of integrated clustering methods for accurate and stable prediction of building energy consumption data. *Applied Energy*, 160, 153–163. <https://doi.org/10.1016/j.apenergy.2015.08.126>
- IEA. (2019). Perspectives for a Clean Energy Transition. The Critical Role of Buildings. In *Energy Transition Progress and Outlook to 2020*. <https://doi.org/10.1017/CBO9781107415324.004>
- IMT, A. J. P. of N. +. (n.d.). *City Energy Project | A Joint Project of NRDC + IMT*. <http://www.cityenergyproject.org/>
- Jefferson, M. (2015). *IPCC fifth assessment synthesis report: “Climate change 2014: Longer report”*: Critical analysis. Elsevier.
- Kadir Amasyali, & Nora M. El-Gohary. (2018). A review of data-driven building energy consumption prediction studies. *Renewable and Sustainable Energy Reviews*, 81, 1192–1205. <https://doi.org/10.1016/J.RSER.2017.04.095>
- Kalogirou, S., Neocleous, C., & Schizas, C. (1997). Building Heating Load Estimation Using Artificial Neural Networks. *Proceedings of the 17th International Conference on Parallel Architectures and Compilation Techniques*, 1–8.

- Kavgic, M., Mumovic, D., Summerfield, A., Stevanovic, Z., & Ecim-Djuric, O. (2013). Uncertainty and modeling energy consumption: Sensitivity analysis for a city-scale domestic energy model. *Energy and Buildings*, *60*, 1–11. <https://doi.org/10.1016/J.ENBUILD.2013.01.005>
- Kavgic, Miroslava, Summerfield, A., Mumovic, D., & Stevanovic, Z. (2015). Application of a Monte Carlo model to predict space heating energy use of Belgrade's housing stock. *Journal of Building Performance Simulation*, *8*(6), 375–390. <https://doi.org/10.1080/19401493.2014.961031>
- Kong, X., Lu, S., & Wu, Y. (2012). A review of building energy efficiency in China during “Eleventh Five-Year Plan” period. *Energy Policy*, *41*, 624–635. <https://doi.org/10.1016/j.enpol.2011.11.024>
- Kontokosta, C. E., & Tull, C. (2017). A data-driven predictive model of city-scale energy use in buildings. *Applied Energy*, *197*, 303–317. <https://doi.org/10.1016/j.apenergy.2017.04.005>
- Li, C., Hong, T., & Yan, D. (2014). An insight into actual energy use and its drivers in high-performance buildings. *Applied Energy*, *131*, 394–410. <https://doi.org/10.1016/j.apenergy.2014.06.032>
- Li, W., Zhou, Y., Cetin, K., Eom, J., Wang, Y., Chen, G., & Zhang, X. (2017). Modeling urban building energy use: A review of modeling approaches and procedures. *Energy*, *141*, 2445–2457. <https://doi.org/10.1016/j.energy.2017.11.071>

- Li, X., Ying, Y., Xu, X., Wang, Y., Hussain, S. A., Hong, T., & Wang, W. (2020). Identifying key determinants for building energy analysis from urban building datasets. In *Building and Environment* (Vol. 181). <https://doi.org/10.1016/j.buildenv.2020.107114>
- Masaki, M., & Zhang, Lijun, Xia, X. (2018). A hierarchical predictive control strategy for renewable grid integrated hybrid energy storage systems. *Applied Energy*, *submitted*(August 2018), 393–402.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.
- Prabhakaran, S. (2018, October). *Cosine Similarity - Understanding the math and how it works? (With python)*. <https://www.machinelearningplus.com/nlp/cosine-similarity/>
- Reinhart, C. F., & Cerezo Davila, C. (2016). Urban building energy modeling—A review of a nascent field. *Building and Environment*, *97*, 196–202. <https://doi.org/10.1016/j.buildenv.2015.12.001>
- Robinson, C., Dilkina, B., Hubbs, J., Zhang, W., Guhathakurta, S., Brown, M. A., & Pendyala, R. M. (2017). Machine learning approaches for estimating commercial building energy consumption. *Applied Energy*, *208*(September), 889–904. <https://doi.org/10.1016/j.apenergy.2017.09.060>

- Sapankevych, N. I., & Sankar, R. (2009). Time Series Prediction Using Support Vector Machines: A Survey. *IEEE Computational Intelligence Magazine*, 4(2), 24–38. <https://doi.org/10.1109/MCI.2009.932254>
- Sokol, J., Cerezo Davila, C., & Reinhart, C. F. (2017). Validation of a Bayesian-based method for defining residential archetypes in urban building energy models. *Energy and Buildings*, 134, 11–24. <https://doi.org/10.1016/j.enbuild.2016.10.050>
- Sun, Y., Huang, G., Xu, X., & Lai, A. C.-K. (2018). Building-group-level performance evaluations of net zero energy buildings with non-collaborative controls. *Applied Energy*, 212, 565–576. <https://doi.org/10.1016/J.APENERGY.2017.11.076>
- Electric Power Monthly with data for January 2018*, (2018) (testimony of U.S. Energy Information Administration (EIA)).
- Wang, W., Hong, T., Li, N., Wang, R. Q., & Chen, J. (2019). Linking energy-cyber-physical systems with occupancy prediction and interpretation through WiFi probe-based ensemble classification. *Applied Energy*, 236, 55–69. <https://doi.org/10.1016/J.APENERGY.2018.11.079>
- Wang, Z., Hong, T., & Piette, M. A. (2019). Data fusion in predicting internal heat gains for office buildings through a deep learning approach. *Applied Energy*, 240, 386–398. <https://doi.org/10.1016/J.APENERGY.2019.02.066>

- Wei, Y., Xia, L., Pan, S., Wu, J., Zhang, X., Han, M., Zhang, W., Xie, J., & Li, Q. (2019). Prediction of occupancy level and energy consumption in office building using blind system identification and neural networks. *Applied Energy*, *240*, 276–294. <https://doi.org/10.1016/J.APENERGY.2019.02.056>
- Wei, Y., Zhang, X., Shi, Y., Xia, L., Pan, S., Wu, J., Han, M., & Zhao, X. (2018). A review of data-driven approaches for prediction and classification of building energy consumption. *Renewable and Sustainable Energy Reviews*, *82*, 1027–1047. <https://doi.org/10.1016/j.rser.2017.09.108>
- Zhao, H. X., & Magoulès, F. (2012). A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews*, *16*(6), 3586–3592. <https://doi.org/10.1016/j.rser.2012.02.049>