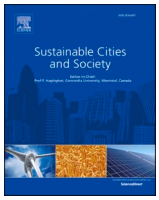
Sustainable Cities and Society 106 (2024) 105425

Contents lists available at ScienceDirect

Sustainable Cities and Society

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



# Developing urban building energy models for shanghai city with multi-source open data

Chengcheng Song [a], Zhang Deng [c], Wenxian Zhao [a], Yue Yuan [a], Mengyue Liu [a], Shen Xu [d], Yixing Chen [a,b,*]

[a] College of Civil Engineering, Hunan University, Changsha 410082, China
[b] Key Laboratory of Building Safety and Energy Efficiency of Ministry of Education, Hunan University, Changsha 410082, China
[c] School of Civil Engineering, Hunan University of Science and Technology, Xiangtan, 411201, China
[d] Huazhong University of Science and Technology, Wuhan, China

ARTICLE INFO

*Keywords:*
Urban building energy modeling
Open data
Gis
Prototypes modeling

ABSTRACT

Urban building energy modeling is crucial for guiding carbon reduction policies, but acquiring reliable data at the urban scale remains challenging. This study develops a model for Shanghai City, China, by integrating multi-source open data. Eight data sources were collected, including maps, satellite imagery, and GIS data, covering 609,763 building footprints and 539,119 buildings (1.57 billion $m^2$). Spatial analysis, supervised learning, and unsupervised machine learning methods were used to categorize buildings into 63 prototypes, and classification accuracy reached 95 %. Historical satellite data and community boundaries determined the year built for over 95 % of buildings. Prototypes were modeled in AutoBPS using local energy saving standards and simulated in EnergyPlus to derive energy use intensities aligning with government ranges. This work demonstrates a practical data fusion approach to develop large-scale, reliable urban building energy models. Integrating heterogeneous open data sources expands the coverage of open data and improves accuracy. The framework and insights provide a valuable foundation to leverage open data for advancing city-scale energy modeling and sustainability planning.

*Abbreviations*

| | |
|---|---|
| POIs | Points of Interest |
| EUI | Energy Use Intensity |
| UBEM | Urban Building Energy Modeling |
| HVAC | Heating, Ventilation, and Air Conditioning |
| AOIs | Areas of interest |
| GIS | Geographic Information System |
| DOE | U.S. Department of Energy |
| OSM | OpenStreetMap |

## 1. Introduction

China accounts for 22 % of global energy consumption. In recognition of this, China has committed to achieving carbon peaking in 2030 and carbon neutrality in 2060. Simultaneously, the urbanization rate in China has experienced a substantial increase from 37.7 % in 2001 to 65.22 % in 2022. This trend has led to a notable growth in the total urban building floor area, reaching 66 billion $m^2$ in 2020. This data includes 29.2 billion $m^2$ for urban residential buildings and 14 billion $m^2$ for public and commercial buildings (Hu et al., 2022). Consequently, it is crucial to model urban building energy at the micro-level to successfully achieve carbon peaking and carbon neutrality goals and develop guiding strategies for urban decision-makers.

Urban building energy modeling (UBEM) refers to the computational representation and simulation of the performance characteristics of a collective of buildings within an urban context. UBEM could provide quantitative insights to inform urban building design and energy policymaking (Hong et al., 2020a). Two paradigms are predominantly utilized in UBEM: the top-down and bottom-up paradigms (Swan & Ugursal, 2009). The top-down paradigm is a macro-level approach, commencing from a holistic perspective and progressively refining to micro-level elements (Kavgic et al., 2010), primarily focusing on the correlation between energy use and macroeconomic variables. However, it often lacks a detailed exploration of technology choices and spatiotemporal characteristics (Reinhart & Cerezo Davila, 2016).

* Corresponding author.
*E-mail address:* yixingchen@hnu.edu.cn (Y. Chen).

https://doi.org/10.1016/j.scs.2024.105425
Received 29 October 2023; Received in revised form 28 February 2024; Accepted 8 April 2024
Available online 9 April 2024

Conversely, the bottom-up paradigm operates at a micro-level, initiating with individual building characteristics and aggregating these to the city level, primarily employed in the study of individual behaviors, energy consumption analysis (Buckley et al., 2021), and optimization scheme formulation. The bottom-up physics-based urban building energy model is based on actual physical parameters and scenarios to simulate individual building's energy consumption (Ang et al., 2020). Despite its strength in providing high-precision solutions for specific issues, it requires substantial data and computational resources, potentially limiting its practicality for extensive system studies (Wang et al., 2022a).

Addressing data and computational resource challenges, UBEM often employs simplified methods by categorizing buildings into multiple prototypes. This paradigm process involves two steps: classification and characterization (Davila et al., 2016). Classification encompasses building type (Deng et al., 2021), construction year, internal structure, geometry (area, floor), heating, ventilation, and air conditioning (HVAC) systems (Pasichnyi et al., 2019a), and other relevant factors. Characterization then defines categories using local regulations or prior research. Despite potential errors of up to 99 % (Reinhart & Cerezo Davila, 2016) in individual buildings due to occupants' behavior variations and the microclimate, errors average at an urban scale. Validation studies show error ranges between 1 % and 15 % in large-scale modeling (Davila et al., 2016). Existing well-established prototype repositories include the U.S. Department of Energy (DOE)'s Prototype Building Models and TABULA's building typologies.

Nonetheless, UBEM demands extensive data, including building characteristics for modeling and aggregated urban energy consumption for calibration/validation. Developed regions like the United States and Singapore have ample city-level energy data, fostering UBEM research, which provides these cities with more guidance in energy policies (Jin et al., 2023). In contrast, many other countries, including China, face challenges in obtaining building information (Hu et al., 2022). Although some Chinese cities (e.g., Shanghai, Beijing, and Shenzhen) started gathering urban building data and promoting online energy monitoring, official building characteristics data remains unavailable. However, these cities still lack an open data portal specifically dedicated to building-related information.

Many studies have made their effort to expand urban building characteristics databases by obtaining relevant information from other approaches. Deng et al. (2021) harnessed Geographic Information Systems (GIS) data to assess 68,966 buildings in Changsha, China, identifying 59,332 building types with 86 % accuracy and using community boundaries to ascertain construction years. Their further research (Deng et al., 2022) classified these buildings into 66 prototype categories/ages, with urban energy consumption calculated by integrating prototype area and Energy Use Intensity (EUI). Chen et al. (2020) employed GIS and Natural Language Processing to automatically reclassify Beijing's Points of Interest (POIs), determining building types based on POIs relationships with footprints, achieving 89 % accuracy. Lu et al. (2014) examined structural, spectral, shape, and spatial characteristics of the buildings in satellite images, using Support Vector Machines and LiDAR data to categorize buildings into single-family, multiple-family, and non-residential, with a 70 % overall accuracy rate. Du et al. (2015) employed GIS data and Very High-Resolution remote sensing imagery, extracting spectral, texture, and geometric information of buildings. By integrating these features through an enhanced random forest algorithm, they categorized 6084 Beijing-based edifices into seven groups, achieving an overall accuracy rate of 79.54 %.

After the classification, there are two main diagrams to calculate the city's energy consumption in UBEM: archetype by aggregation and building-by-building. The aggregation method involves calculating the EUI of representative urban prototype buildings, determining the total area of these prototypes and subsequently integrating the combined area with their respective EUIs to estimate the city's energy consumption (Chen et al., 2019; Deng et al., 2022; Pasichnyi et al., 2019b). This paradigm calculates "averaged" building parameters (including geometry, building types, etc.) for different types of buildings based on statistics. The ultimate goal is to compute "averaged" building energy consumption for the entire large-scale area. And the climate files used represent the "average climate" parameters of the area. Hence, errors resulting from individual environmental and energy-use characteristics are "averaged out," and resulting in an acceptable error, Deng et al. (2022) validated their energy consumption results for 68,966 buildings in Changsha, China, against statistical yearbook data, ending with an 8.6 % error. Österbring et al. (2016), in their validation for 433 buildings in Gothenburg, Sweden, achieved a 3 % error. Dall'O' et al. (2012) reported a 10 % error in their validation for 6688 buildings in Lombardy, Italy. However, this paradigm fails to reach high accuracy on a small scale, as each building operates in a different climatic environment. And the occupant behavior patterns of buildings also vary, which can lead to significant errors when UBEM models are considered at the scale of individual buildings, sometimes even up to 1000 % (Wang et al., 2018). For smaller computational scales, it is feasible to calculate the shading and energy consumption of each building individually. Davila et al. (2016) assigned 52 prototype buildings to 83,541 structures in Boston using official GIS datasets and building archetypes. Subsequently, they employed Rhino to compute the potential shading surfaces for each building and utilized EnergyPlus, the most widely used building energy consumption modeling engine at present, for individual simulations. While this approach captures the interactive relationships between buildings (e.g., long-wave radiation heat exchange, shadows), it demands significant computational resources.

To automate UBEM analysis, researchers have devoted to developing UBEM tools: umi (Reinhart et al., 2013) is a Rhinoceros-based urban modeling tool that allows for comprehensive operational energy, daylighting, and walkability assessments of entire neighborhoods, utilizing simulation engines like EnergyPlus and Radiance/Daysim, along with Grasshopper and Python scripts. CitySim (Robinson et al., 2009) focuses on calculating heating and cooling demands; SimStadt, primarily employed for rapidly generating evaluation scenarios to assess city-scale heating requirements; City Energy Analyst (Fonseca et al., 2016), a Python-driven tool with an intuitive graphical interface, streamlines analysis of building heating and cooling loads for district energy planning. Concurrently, TEASER(Remmen et al., 2018), another Python-based application, seeks to unite UBEM and Urban System Energy Modeling, enabling a detailed representation of urban built environments and fostering an extensive understanding of city-scale energy systems; CityBES (Hong et al., 2016), a web-based platform for simulating large-scale building energy performance, aids energy benchmarking, urban planning, retrofit analysis, building management, photovoltaic potential evaluation, and urban microclimate visualization. AutoBPS (Deng et al., 2023), a recent UBEM tool, utilizes GeoJSON input and EnergyPlus as its engine, providing a comprehensive UBEM for residential and commercial buildings. AutoBPS supports diverse energy characteristic exploration, mixed-use area scenarios, urban energy demand, retrofit, and PV analyses. Although EnergyPlus was originally for individual building simulations, its physics-based features allow it to model urban scenarios effectively. It integrates well with Python, Grasshopper, and other tools, making it suitable for complex UBEM tasks. Luo et al. (2020) demonstrated its utility in a Chicago district study, where the district-level cooling energy demand increased by 1.39 % and heating demand decreased by 0.45 % when considering building surroundings. EnergyPlus is used in various UBEM tools such as CityBES, developed by the Lawrence Berkeley National Laboratory (Hong et al., 2016). URBANopt by the National Renewable Energy Laboratory (El Kontar et al., 2020), UMI by the Massachusetts Institute of Technology (Reinhart et al., 2013), and CESAR by the Swiss Federal Institute of Technology Zurich (Wang et al., 2018), all of which utilize EnergyPlus at their core.

As previously mentioned, UBEM is very important to the development and planning of urban areas. However, acquiring extensive data

and modeling a whole city is not easy. Prior studies have investigated prototype buildings in Shanghai, encompassing residential (Peng et al., 2021), office buildings (Hong et al., 2020b), commercial buildings, shopping malls, and hotel buildings, urban morphology (Wang et al., 2022b), and retrofit payback (Yu et al., 2021). However, to the best of the author's knowledge, extant research on Shanghai's buildings predominantly concentrates on specific building types or regions. An all-inclusive urban energy consumption model is still needed to evaluate the energy consumption of Shanghai's entire building stock.

Moreover, while publicly available GIS data offers potential for architectural modeling, it has limitations: reliance on single-source data, insufficient coverage of building data in China by well-known GIS platform OpenStreetMap (OSM), and possible redundancy and conflicting information from different sources. This study aims to overcome these limitations by integrating multi-source GIS data, enhancing the model's reliability and robustness. Furthermore, this study categorizes 609,763 building footprints and 539,119 buildings in Shanghai City into 21 building types and three construction periods, resulting in a total of 63 distinct prototypes. The construction year of these buildings was determined using historical satellite imagery coupled with deep learning methods. Building upon this, Shanghai's urban building energy consumption model was developed using AutoBPS. Subsequently, a comprehensive analysis of Shanghai's annual energy consumption outcomes has been carried out, and this study has compared the integrated data with publicly available government data to verify the reliability of this approach.

## 2. Methodology

### 2.1. The workflow of this study

Fig. 1 illustrates the workflow adopted in this study. Initially, relevant GIS data are gathered, encompassing building footprints, building heights, POIs, areas of interest (AOIs), as well as historical satellite images of Shanghai. Additionally, both the mandatory standards of Shanghai, China, and the ASHRAE Standards 90.1 are collected. Subsequently, the building attributes data are integrated to classify and identify the buildings in Shanghai by employing spatial analysis, clustering, supervised analysis, and unsupervised analysis methods. After that, the construction years of these buildings are determined with the historical satellite images. In the third step, Shanghai's prototype buildings are modeled using AutoBPS, incorporating the mandatory standards, ASHRAE Standards 90.1, and pertinent literature references. Next, EnergyPlus is utilized to compute the EUIs of these prototype buildings, enabling a comprehensive assessment of Shanghai's urban building energy consumption.

### 2.2. Introduction of the case study buildings information

Shanghai, one of the most developed cities in China, has a permanent population of 24.8 million. Located in the eastern region of China, Shanghai lies on the west coast of the Pacific Ocean, along the eastern edge of the Asian continent, between 120°52′ E and 122°12′ E longitude and 30°40′ N and 31°53′ N latitude. With an average elevation of 2.19 m, Shanghai covers a total area of 6340.5 square kilometers and is divided into 16 districts.

According to the classification of GB 50176-2016, Shanghai falls under the "3A" category, characterized as a "hot summer and cold winter". Fig. 2 depicts the map of Shanghai, including the kernel density of POIs, which can be regarded as an indicator of the area's prosperity. The most bustling areas in Shanghai encompass Huangpu, Xuhui, Changning, Jing'an, Putuo, Hongkou, Yangpu, and certain parts of Pudong. The following are Minhang, Baoshan, Jiading, and Songjiang. The peripheral regions include Qingpu, Fengxian, and Chongming.

### 2.3. Multi-source data fusion

#### 2.3.1. Data source introduction

UBEM primarily focus on buildings, yet acquiring specific building data is more challenging compared to other GIS data. For effective UBEM modeling, data from various sources must be further processed. A minimum GIS data resolution of 10 m, covering Shanghai, is required for building analysis. This article utilizes data from eight sources mentioned in Table 1. The information contained in the data mainly includes Footprints, POI, AOI, Height, History Satellite Image, Year, Satellite Image, Land Type, and Energy Use Intensity (EUI). As for the sources, Baidu Map, Amap, and Google are among the data providers. Baidu and AMAP data are accessed via their APIs, but limitations lead to gaps in the data. AI Earth, an Alibaba affiliate, offers up-to-date building coverage data at a 10-meter resolution. The Global Human Settlement Layer (GHSL), from the European Union Research Center, offers free tools and data for assessing human presence worldwide, including primary categorization of buildings into residential and non-residential. Additionally, Anjuke and Lianjia, Chinese real estate platforms, offer extensive residential community data. Some data sources may contain repeated information, while others can complement one another.

This study primarily employs data in the Shapefile (.shp) format, a universal data configuration established by Environmental Systems Research Institute (ESRI). Additionally, GeoJSON is another commonly used format. Beyond the open data for Shanghai and China, open data in

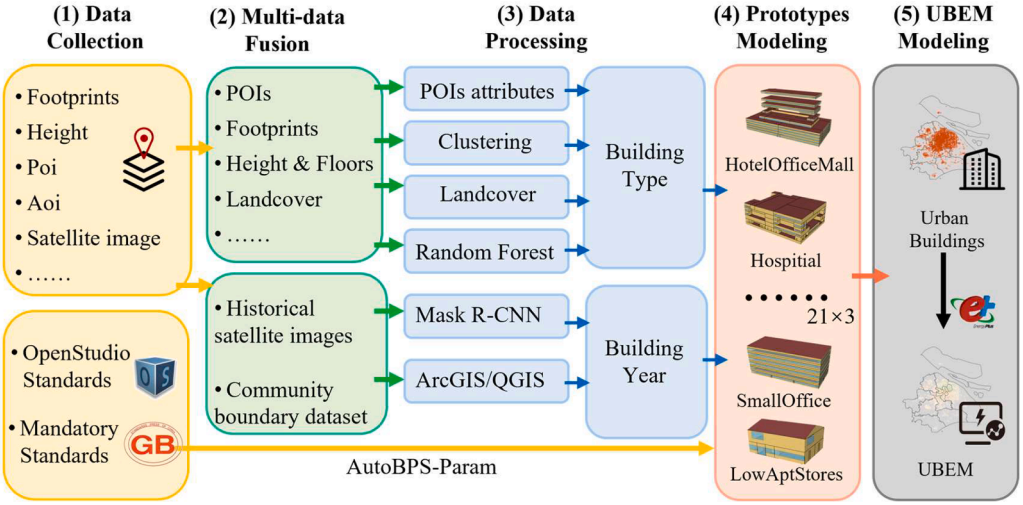



**Fig. 1.** Workflow of this study.

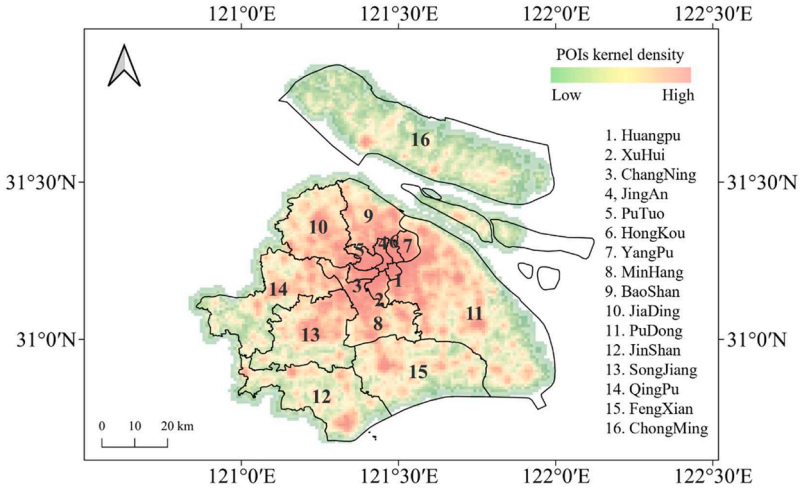



**Fig. 2.** Case study area: Shanghai.

**Table 1**
Data sources and information used in this study.

| Information | Data Source | Format | Cover Area | Update |
|---|---|---|---|---|
| Footprint | Baidu | Shapefile | China | Day |
| | Amap | Shapefile | China | Day |
| | OpenStreetMap | Shapefile | Worldwide | UN |
| AOIs | OpenStreetMap | Shapefile | Worldwide | UN |
| | Baidu | Shapefile | China | Day |
| | Anjuke, Lianjia | csv | China | Day |
| POIs | Baidu | Shapefile | China | Day |
| | Amap | Shapefile | China | Day |
| Land Cover | AI earth | tiff | Worldwide | Months |
| | GHSL | tiff | Worldwide | Year |

**Table 2**
Examples of multi-source data in typical cities.

| City | Data | Sources |
|---|---|---|
| New York | Footprints | OpenstreetMap, Open Buildings, NYC Opendata |
| | POIs | OpenstreetMap, NYC Opendata |
| | AOIs | OpenstreetMap, NYC Opendata |
| | Land cover | OpenstreetMap, AI Earth, GHSL |
| Singapore | Footprints | Singapore's open data, GlobalMLBuildingFootprints, Open Buildings, OpenstreetMap |
| | POIs | Singapore's open data, GlobalMLBuildingFootprints, Open Buildings, OpenstreetMap |
| | AOIs | OpenstreetMap |
| | Land cover | OpenstreetMap, AI Earth, GHSL |

similar formats is extensively available worldwide, with most formats being convertible to the Shapefile format.

Globally, OSM is recognized as the most comprehensive open data source. However, its dependence on user-contributed updates and maintenance results in variable quality and timeliness across different countries. Hence, researchers with access to high-quality local data are advised to integrate it with OSM data to enhance its quality. The practice is exemplified by local city databases (https://data.cityofnewyork.us/) and Singapore's open data (https://beta.data.gov.sg/) platforms serve as regional open data repositories, while GlobalMLBuildingFootprints (https://github.com/microsoft/GlobalMLBuildingFootprints) and Open Buildings (https://sites.research.google/open-buildings/) provide data across multiple regions worldwide.

Similar studies utilizing such data have been conducted in Ireland (Ali et al., 2020), Canada (HosseiniHaghighi et al., 2022), Beijing, China (W. Chen et al., 2020), Changsha, China (Deng et al., 2022), and Tokyo, Japan (Perwez et al., 2023), often relying solely on a single data source. Additional open data resources are available at https://3d.bk.tudelft.nl/opendata/opencities/ and related review(Jin et al., 2023). It should be noted that, besides the open data listed in the table, file formats used by paid APIs, such as those offered by Google and Microsoft, are also convertible to the formats used in this study, indicating that the methods described herein are fully scalable (Table 2).

## 3. Data fusion

A total of 1491671 POIs were collected, comprising 780,347 from Baidu and 711,324 from Amap. A deduplication process was implemented to mitigate the impact of duplicate POIs on building categorization. The fusion of the two sources followed the approach shown in Fig. 3(a). Initially, all Amap POIs within the tolerance range of the target Baidu source were identified. Then, a natural language processing approach was used to remove semantically redundant POIs, utilizing the matching algorithm (Wu et al., 2022) developed explicitly for similar POI matching. The resulting non-redundant POIs totaled 1065,840, with 780,347 from Baidu and 285,493 from Amap. The distribution of these POIs throughout Shanghai is illustrated in Fig. 3.

Regarding building footprints, it is crucial to recognize that the footprints do not represent the entire building, but rather their individual components. The two sources were mutually supplemented to preserve as much information as possible. For duplicate buildings, the source with more components was chosen. As shown in Fig. 3, the Baidu dataset, having more components, constituted a larger portion of the final dataset with 593,754 footprints, encompassing 512,359 buildings accounting for 95 % of the total. The Amap source contributed 26,760 footprints, which accounted for 5 % of the total. The final building database comprises 609,763 footprints and 539,119 buildings, covering 1567,921,014 $m^2$.

Building height and number of floors are important for statistical analysis of prototype buildings. Baidu source lacks floor information, while the Amap source lacks height data. Dividing Baidu's height by Amap's floors resulted in an unreasonable floor height distribution. Regulations stipulate residential buildings have a minimum height of 2.2 m, preferably 2.7 or 2.8 m. Shanghai government calculates building areas with heights between 2.2 m and 4.5 m. Therefore, it is necessary to calibrate this part of the data. According to the existing calculation standards, there are 275,929 footprints with heights below 2.2 m or above 4.5 m, accounting for 45.25 % of the data. Retaining the original Amap data, 333,834 footprints are available for analysis.

Considering the ease and accuracy of obtaining building height, this

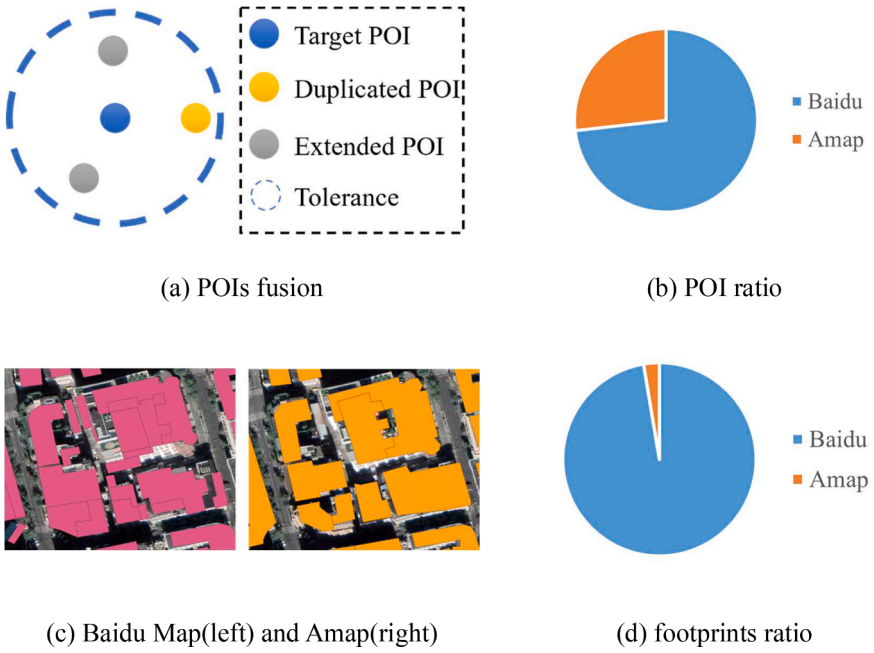



**Fig. 3.** Multi-source GIS data.

study recalibrates floor numbers based on actual building height. For buildings that do not comply with the standards, the number of floors is determined by dividing the building height by the established floor height norm: 2.8 m for residential and 3.5 m for non-residential buildings.

Compliant buildings retain Amap floor data. Just dividing height by the standard floor height leads to information loss and significant deviations in floor numbers. Several areas (Jing'an, Putuo, Xuhui, Yangpu, Changning) closely matching government land cover data were chosen for validation. We calibrated the building floors based on the total area data published officially, with Table 3 displaying the calibration results considering different data sources. In this process, the Amap data is multiplied by the built-in floor count, whereas the Baidu data is calculated by dividing the height by the standard floor height. Results align closely with government reports, affirming the method's relative accuracy compared to other approaches. The calibrated distribution of floor heights is shown in Fig. 4, indicating that the existing method yields relatively more reasonable floor heights than other data sources.

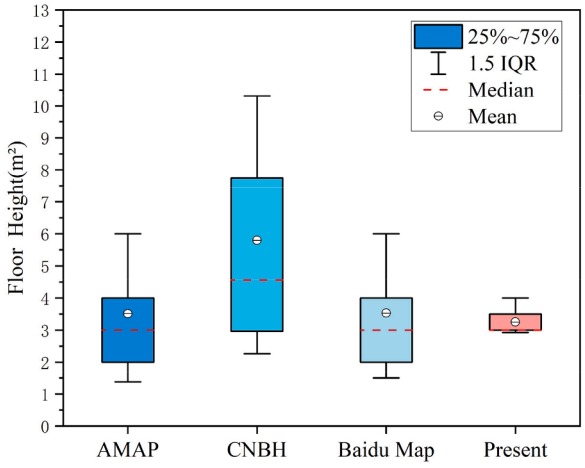



**Fig. 4.** Building Floor Height Distribution from Different Sources.

### 3.1. Building type identification

This study employs various data forms, including POIs, AOIs, and building footprints. POIs, obtained from web map services such as Google Maps, Baidu Maps, and Amap, are GIS points characterized by attributes like name, entity category, and sub-category. Some POIs explicitly describe building attributes, while others describe other information within the building. Based on their ability to reflect the building's function, POIs are classified into main categories and sub-categories. Main-category POIs serve as a benchmark for evaluating the building's main function, while sub-category POIs act as supplementary indicators. The main attributes of POIs include hotels, shopping malls, cultural art galleries, government institutions, residential buildings, hospitals, schools, commercial office buildings, and factories. For buildings lacking a main category, clustering is employed to ascertain their category. Buildings lacking any geographical information can be supplemented through supervised learning and land data. Building

**Table 3**
Comparison between government report data and present data.

| | Land Cover (10,000 m$^2$) | | Total Area (10,000 m$^2$) | | | |
|---|---|---|---|---|---|---|
| Distinct | Government | Present | Government | Amap | Baidu | Present |
| Jing'an | 975 | 954.37 | 6179 | 6633.46 | 6308.82 | 6097.93* |
| Putuo | 1218 | 1205.19 | 6424 | 7629.53 | 6173.55 | 6609.13* |
| Xuhui | 1148 | 1085.25 | 6633 | 6385.49* | 5807.6 | 6053.37 |
| Yangpu | 1319 | 1264.39 | 6357 | 7766.14 | 5817.84 | 6357.08* |
| Changning | 718 | 716.19 | 4355 | 4612.18 | 3868.74 | 4142.9* |

* Most close to the government report data.

footprints do not represent the buildings themselves. Single buildings with multiple height levels may have multiple footprints. AOIs, akin to POIs, are spatial boundaries carrying regional information. However, AOIs require assignments based on POIs. Consequently, their classification adheres to the same methodology as POIs, with the added benefit of encompassing multiple buildings.

### 3.1.1. Assign attributes to building footprints

Owing to the coordinate system settings in GIS and the matching discrepancies between different sources, numerous POIs that should be located within buildings may not be accurately positioned. To correctly associate these POIs with buildings, this study adopted the same approach, as shown in Fig. 5, to assign the POIs with a tolerance range. If a building is within the tolerance, the corresponding POI will be considered as within the building.

The study implemented a tolerance range analysis for buildings varying in size from 1 m to 10 m, examining 12,134 POIs to assign more attributes of POIs to buildings. The evaluation of the effectiveness of different tolerance settings was conducted using the Area Under Curve. This analysis anticipated a significantly higher incidence of POIs within buildings (9.37 times more than those outside), leading to the conclusion that the optimal tolerance setting was 5 m. Despite this, it is crucial to note that a considerable number of POIs still fell outside the building perimeters, which may be attributed to missing building footprints. Observations from Fig. 6 also revealed that the majority of POIs were located within 5 m of the buildings. Consequently, a tolerance of 5 m was established as the most suitable for the study.

When prioritizing buildings with more components, there is another issue to consider. Some POIs may be situated within one component and cannot be assigned to other components, even though they belong to the same building. The components are delineated based on the external height of the structure rather than its functional zoning. However, in most cases, the functional division of a building is determined by the floor on which it is located.

In this study, the adopted approach involved merging these components based on their adjacency relationships and assigning distinct building IDs to each. Subsequent analyses were conducted based on these newly defined buildings, and the results were assigned to the associated footprints. This methodology aimed to maximize the impact of these POIs and enable the accurate identification of mixed-use buildings. Before applying this approach, 85,373 footprints were associated with POI attributes; after adopting this approach, 175,016 footprints had POI attributes, marking a 105 % enhancement.

For the assignment of AOIs, buildings were considered to be within an AOI if they encompassed at least 80 % of its area. After the assignment of the AOIs attributes, 307,037 buildings had attributes.

### 3.1.2. Identify the building type

Additional geospatial information needs to be integrated to determine the categorical attributes of buildings in geographical space, as shown in Fig. 7.

*3.1.2.1. The building attributes approach.* The definition of building types is diverse. Table 4 displays three different classifications drawn from Shanghai government reports, research papers, and technical

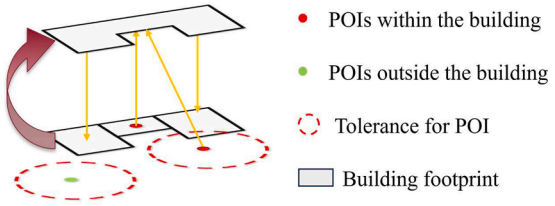



**Fig. 5.** Assign POI to the building footprints method.

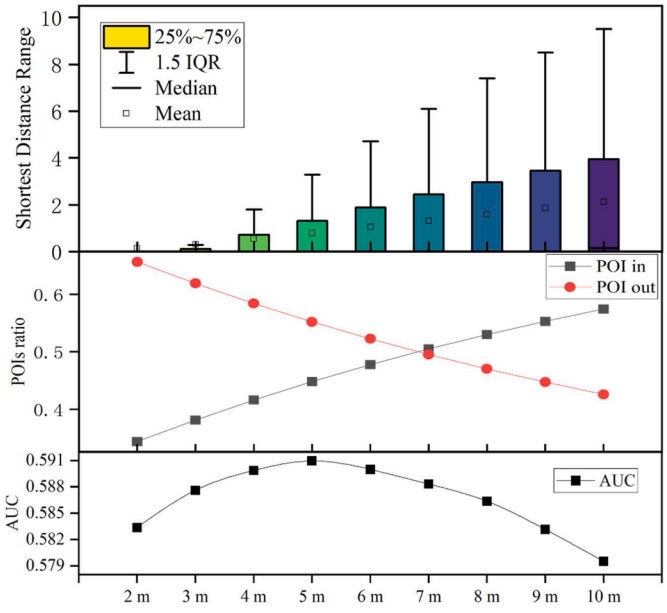



**Fig. 6.** Tolerance analysis.

reports from research institutions. These classifications led to the selection of Factories, Schools, Offices, Restaurants, Shopping Malls, Hospitals, Commercial Mixed-Use, Education, Retail, and Other Mixed-Use were ultimately selected as the primary building categories. Based on this selection, this study reclassified the POI categories. Reducing the original 227 POI categories by manually filtering out irrelevant ones for building type analysis, such as public toilets. The refined classification yielded seven major categories: Residential, School, Office Building, Hotel Building, Shopping Mall, Hospital, and Factory. These categories directly reflect the main functions of buildings. Additionally, 5 minor categories were identified, namely Retail, Restaurant, Office, Food and Beverage, Recreation.

After categorizing the attributes present in these buildings, further analysis was conducted by examining the quantities of each attribute. The following deduction rules were applied:

- Buildings containing only one main category POI were classified based on the description of the main category POI.
- Buildings containing two main category POIs accounting for more than 1 % were classified as corresponding mixed-use buildings.
- Buildings containing three or more main category POIs were classified as other mixed-use buildings.
- Buildings without main-category POIs but containing sub-category POIs were clustered separately for further observation.

With this approach, 255,513 buildings obtained their building types, accounting for 47.4 % of the total.

*3.1.2.2. The cluster approach.* After assessing the primary attributes, some buildings without any primary attribute POIs were identified, yet they possess numerous non-primary attribute POIs, such as Food and Beverage and Office. While these attributes do not directly indicate the building's function, they can provide insights for deducing it. Additionally, for buildings with primary attributes, clustering based on their non-primary attribute POIs enables the inference of secondary attributes, facilitating the identification of additional mixed-use buildings.

The number of clusters was determined using the "elbow method", and the clustering results were observed. Fig. 8 presents the proportion of POI attributes in buildings without primary attributes, divided into eight categories using the elbow method. These categories represent the typical POI distribution in eight types of buildings. Subsequent manual

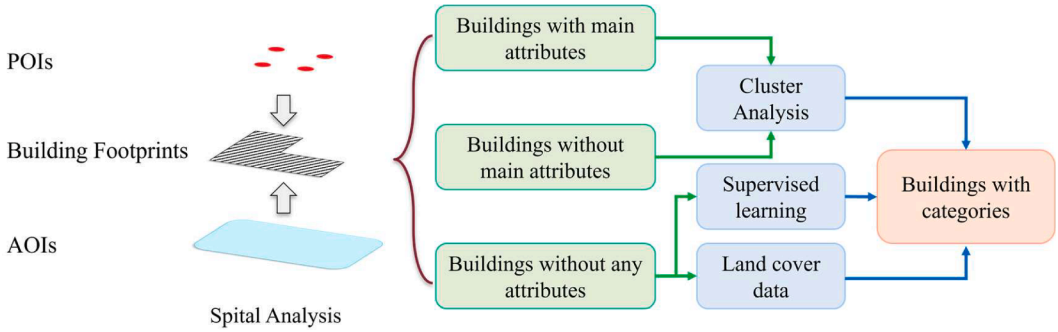



**Fig. 7.** Archetype identification of the buildings.

**Table 4**
Building types.

| Ref | Building Main Types | Source Type |
|---|---|---|
| (ShangHai Statistical Yearbook 2022, 2022) | Factory, Schools, Warehouses, Offices, Stores, Hospitals, Theatres | Government Report |
| (Shanghai Municipal Commission of Housing & Urban-Rural Development, 2021) | Government, Office, Restaurant, Shopping Mall, Health Care, Commercial Mixed-Use, Education, Culture, Sport, Tourist | Government Report |
| (Deng et al., 2022) | Residential, Office, Commercial Mixed-Use, Hotel, School, Shopping Mall, Hospital, Retail, Education, Tourist | Research Paper |
| (An et al., 2023) | Residential, Office, Education, Mercantile, Lodging, Health Care | Research Paper |
| (Prototype Building Models \| Building Energy Codes Program, 2021) | Residential, Office, Primary School, Retail, Restaurant, Hotel, Hospital, Shopping Mall | Technical Report |
| – | Residential, Factory, Office, Restaurant, Shopping Mall, Hospitals, Education, Retail, Mixed-Use, Tourist | This Paper |

categorization was based on these POI distributions: Clusters 1, 6, and 8, encompassing 13,277 buildings, were classified as Shopping Malls; Clusters 2, 4, and 7, totaling 17,903 buildings, as Shopping-Restaurant complexes; Cluster 3, with 10,420 buildings, as Office Buildings; and Cluster 5, consisting of 3019 buildings, as Other-Non-Residential.

For buildings already possessing primary attributes, if they also contain secondary attribute POIs, only these secondary attributes are clustered. The overall function of the building is then analyzed in conjunction with its primary attributes. Fig. 9 displays the classification of secondary attributes for buildings with the primary attribute of an office building. The elbow method identified four categories. Through manual assessment, Clusters 1 and 4 were designated as Office-Shopping, while Clusters 2 and 3 were classified as Office-Restaurant. This approach supplemented 51,517 additional building types, accounting for 9.5 % of the total.

*3.1.2.3. The landcover approach.* For the buildings that still lack attributes, land-cover data includes the classification of buildings. The land-cover data used in this study, sourced from the GHSL with a resolution of approximately 10 m, is adequate for discerning building types, as illustrated in Fig. 10. However, it only indicates whether a building is residential or not. In this research, if a building is covered by land-cover data and over 60 % of it is categorized as Residential, it is then classified as Residential. Similarly, if over 60 % is Non-Residential, it is classified as such. Buildings covered by less than 60 % of either category are considered outside the scope of this method. By using this method, 94,609 additional building types were assigned, accounting for 17.5 % of the total.

*3.1.2.4. The supervised learning approach.* As Fig. 11 illustrated, for the remaining buildings, considering the significant geometric differences between different building types, especially between residential and non-residential buildings, the attributes of the buildings were extracted by obtaining their bounding rectangles and related parameters, including Building footprint area, building footprint perimeter, number

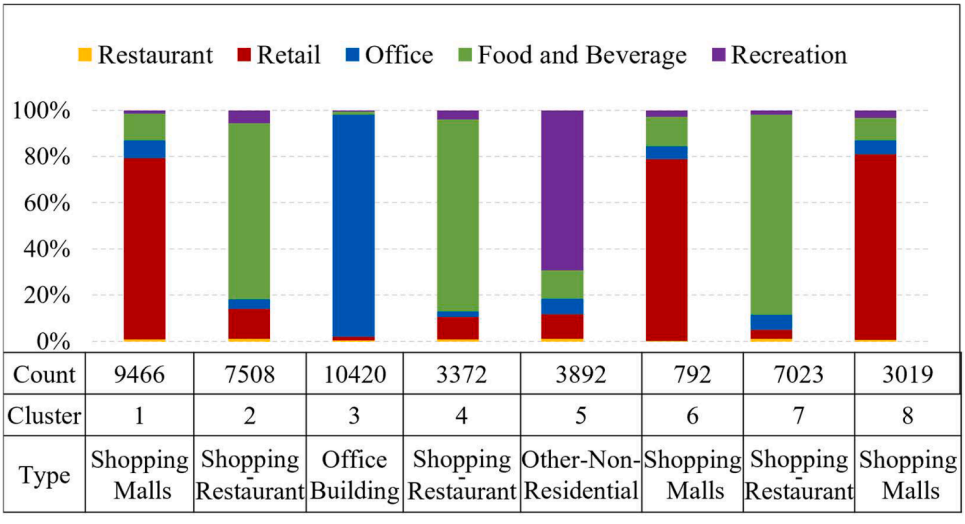
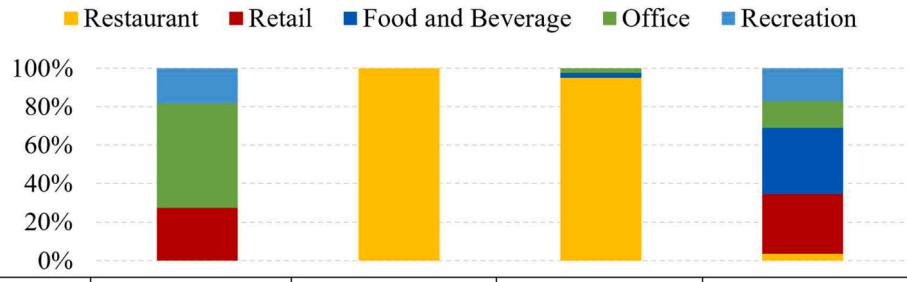



**Fig. 8.** Buildings with no main attribute clustering.




| Count | 9466 | 7508 | 10420 | 3372 |
|---|---|---|---|---|
| Cluster | 1 | 2 | 3 | 4 |
| Type | Office Shopping | Office Restaurant | Office Restaurant | Office Shopping |

**Fig. 9.** Office buildings clustering.

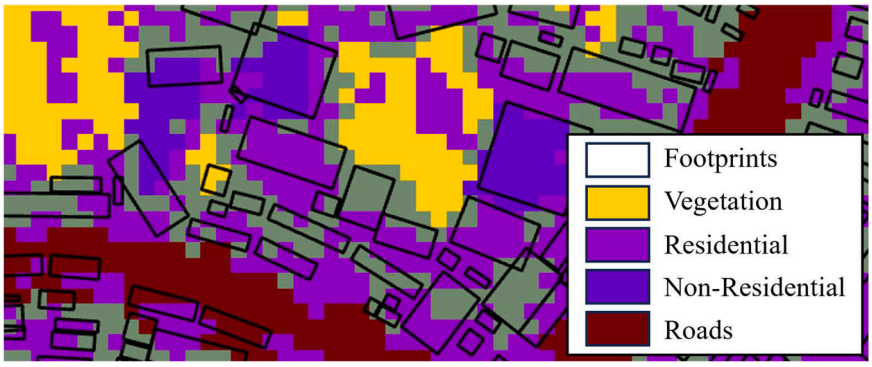



**Fig. 10.** The landcover data from GHSL.

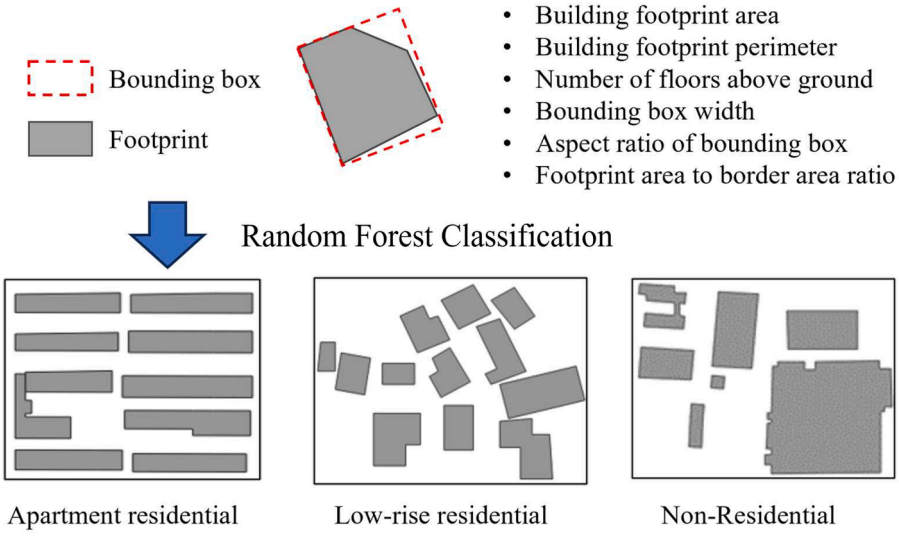



**Fig. 11.** The supervised learning approach.

of floors above ground, bounding box width, aspect ratio of bounding box, and footprint area to border area ratio. Subsequently, 3037 buildings were manually labeled with their corresponding types, and the remaining buildings were classified using the random forest algorithm.

### 3.2. *The year-built identification*

#### 3.2.1. *History satellite image approach*

With the development of computer vision, the detection of architectural changes using ultra-high-resolution satellite imagery (0.25 m/pixel) can be very accurate. This technique involves year-by-year

comparison of historical satellite imagery, allowing for accurate identification of each building's alterations and the determination of their year-built. However, the acquisition and processing of ultra-high-resolution satellite imagery is cost-prohibitive for expansive areas like Shanghai. The ultra-high-resolution imagery amounts to 2.52 terabytes for just one year of satellite image, while the quality of regular high-definition imagery (1.07 m/pixel) is relatively poorer. Employing ArcGIS Pro and mask R-CNN (He et al., 2017) for change detection in specific regions establishes a baseline era for all buildings. This method is particularly effective for non-residential buildings with larger footprints, achieving an accuracy of 91.1 %. In contrast, its efficacy diminishes for residential buildings, yielding only a 73.2 % accuracy rate. Therefore, another approach to obtaining the year built of buildings is by utilizing community boundaries.

#### 3.2.2. Community boundary datasets approach

Real estate websites offer extensive data about the locations of residential communities and their corresponding construction years. Additionally, the AOI data predominantly focuses on residential communities. Utilizing web scraping techniques to retrieve the construction years of these buildings and assigning them to the respective properties allows for effective augmentation of building age information. Through this approach, 228,622 buildings have been updated, accounting for 37.5 % of the total.

### 3.3. Prototype models development

This study presents a statistical analysis of the mean geometries of different building types, and corresponding prototype buildings were established. For the non-geometric parameters of these buildings, the parameters cited in authoritative sources will be used to augment the non-geometric information. These references encompass local government reports, technical reports, and relevant research papers.

The establishment and calibration of the prototype building were conducted using AutoBPS, which effectively determines specific parameter types in a building, including the building's envelope, HVAC system, solar hot water system, internal loads, etc., based on inputting the building's geometric parameters, building type, and construction year. AutoBPS can also model mixed-use buildings mentioned in this paper. For some building types, where other literature data or official data are available, the AutoBPS-Param module(Chen et al., 2023) was used for calibration. The specific modeling process, illustrated in Fig. 12: involves merging the geometric shapes of existing reference buildings in Shanghai and determining non-geometric parameters based on local mandatory standards. The Shanghai prototype buildings are then subsequently finalized using the AutoBPS-Param.

Regarding the reference for Geometry, including the detailed layout of the building, the main reference is the existing prototype, with most building types referring to the DOE Prototype Building Models (Prototype Building Models | Building Energy Codes Program, 2021). These models are widely employed in the development of residential and office building archetypes in the literature. In terms of the geometry of commercial mixed-use and school prototypes, this work refers to the mixed-use prototypes developed by Deng et al. (2022)

For the Non-Geomerty parameters, this study adheres to the mandatory national regulations used for different types of buildings, categorizing construction years into three distinct periods. Both residential and non-residential buildings are divided into three groups. The categorization of residential buildings follows the standards JGJ 134–2001 and JGJ 134–2010, which are pre-2001, 2002–2009, and post-2010, respectively. Commercial buildings are divided into pre-2005, 2005–2015, and post-2015 based on the standards GB 50189–2005 and GB 50189–2015. For mixed-use buildings, the specific parts corresponding to different types are referenced according to the respective regulations. Table 5 displays the summarized main parameters.

## 4. Results

### 4.1. Building dataset of Shanghai and validation

#### 4.1.1. The building dataset of Shanghai

The 539,119 buildings obtained in Shanghai were classified by applying a series of processing steps. Fig. 13 reveals that the area under consideration primarily comprises residential, non-residential, and mixed-use zones.

Residential areas constitute approximately 57.40 % of the total constructed space, notably with high-rise residential buildings comprising 28.45 % of this segment. Non-residential spaces encompass around 22.12 % of the total area, with diversified types including

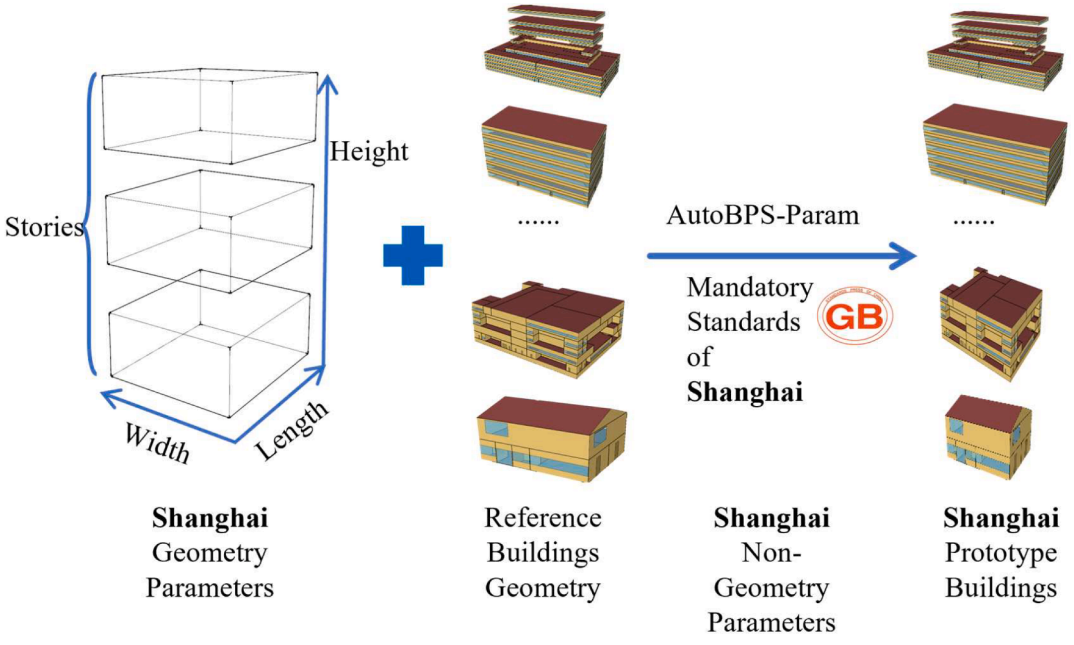



**Fig. 12.** Shanghai Prototype modeling with AutoBPS-Param.

**Table 5**
General EnergyPlus model settings of each prototype building parts.

| Parameters | Residential part | | | Commercial part | | |
|---|---|---|---|---|---|---|
| | Pre-2001 | 2002–2009 | Post-2010 | Pre-2005 | 2006–2014 | Post-2015 |
| Exterior wall U-value(W/(m$^2$*K)) | 1.96 | 1 | 0.8 | 2 | 1 | 0.6 |
| Roof U-value (W/(m$^2$*K)) | 1.66 | 0.8 | 0.5 | 1.5 | 0.7 | 0.4 |
| Window U-value (W/(m$^2$*K)) | 6.6 | 3.2 | 2.8 | 6.4 | 3 | 2.6 |
| Window SHGC | 0.85 | 0.48 | 0.34 | 0.69 | 0.43 | 0.35 |
| Lighting power density (W/m$^2$) | 7 | 7 | 6 | 15 | 11 | 9 |
| Equipment power density (W/m$^2$) | 4.3 | 4.3 | 4.3 | 20 | 20 | 15 |
| Occupancy (person/m$^2$) | | 0.05 | | 0.125 | 0.125 | 0.125 |
| Cooling/heating setpoints (°C) | | 26/18 | | 26/20 | 26/20 | 26/20 |
| Cooling/heating COP | 2.2/1 | 2.3/1.9 | 2.9/2.2 | 4.2 | 5.1 | 5.6 |

education facilities (4.63 %), office buildings (3.08 %), hotels (0.46 %), restaurants (0.06 %), shopping areas (2.25 %), hospitals (0.70 %), and factories (3.16 %). Other categories occupy the remaining 6.78 %. Fig. 14 shows the era distribution ranges for each type of building.

#### 4.1.2. Validation with manual identification building type

A total of 5872 buildings were selected and manually annotated based on satellite images, street view images, and POI data. The annotated results were then compared with the proposed method. To streamline the manual annotation process, this phase was confined to classifying buildings into eight categories: residential, hotel, shopping, education, factory, restaurant, hospital, and office.

In machine learning classification tasks, the confusion matrix and its related metrics – Precision, Recall, F1-score, and Support – are key for assessing model performance. The confusion matrix elements are True Positives (TP) - correct positive predictions; False Positives (FP) - incorrect positive predictions; True Negatives (TN) - correct negative predictions; and False Negatives (FN) - incorrect negative predictions. Precision (Precision = TP / (TP + FP)) measures the accuracy of positive predictions. Recall (Recall = TP / (TP + FN)) gauges the model's ability to identify actual positives. F1-score (F1 Score = 2 * (Precision * Recall) / (Precision + Recall)) balances precision and recall. Support indicates the occurrence count of each class in the dataset. Together, these metrics provide a comprehensive understanding of a classification model's performance. As shown in Table 6, The overall recall rate of the model is generally high, all exceeding 0.9. The precision and recall rates are high for easily identifiable buildings such as Primary School, Residential, Shopping Mall, and Mix-use buildings. However, the precision rate may sometimes be less accurate. Particularly noteworthy is the model's classification of Commercial Residential and Hotel as key POIs. However, manual calibration revealed that many hotels are not standalone buildings. While this is not apparent in the classification of large hotels, the classification of small hotels showed that they frequently coexist within residential or office buildings. This ambiguity complicates the criteria for manual classification, making it difficult to define such buildings. This is why Commercial Residential, Hotel, and Office have

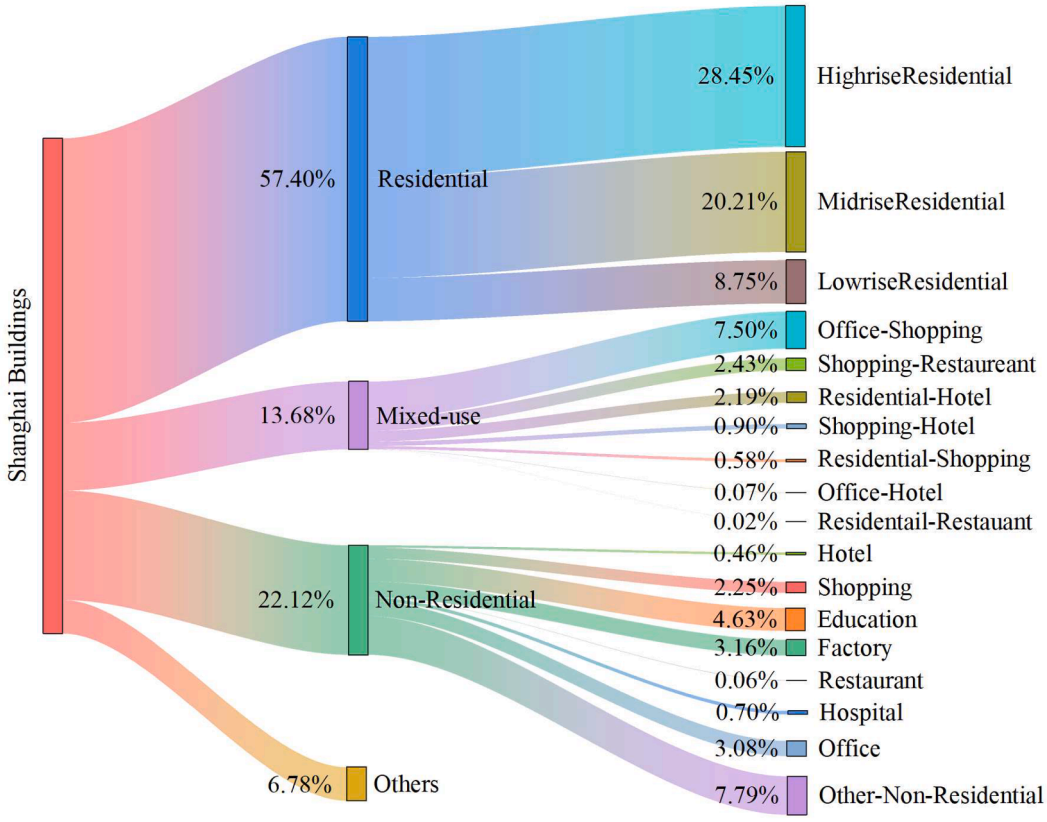



**Fig. 13.** The proportion of prototype buildings in Shanghai.

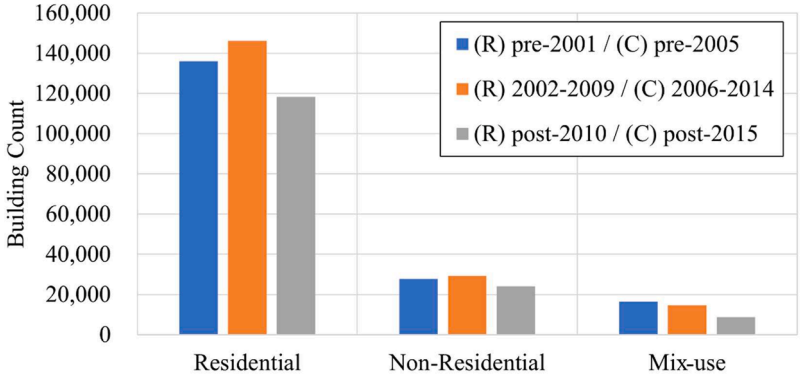



**Fig. 14.** The year distribution of the buildings: Each label corresponds to different eras for Residential (R) and Commercial (C) sectors. For example, the blue label represents pre-2001 for Residential buildings and pre-2005 for Commercial buildings.

**Table 6**
Classification model performance metrics table.

| Building type | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Commercial-Residential | 0.81 | 0.97 | 0.88 | 158 |
| Hospital | 0.95 | 0.97 | 0.79 | 59 |
| Hotel | 0.33 | 0.96 | 0.49 | 28 |
| Office | 0.53 | 0.97 | 0.69 | 37 |
| Office-Hotel | 0.84 | 0.94 | 0.89 | 79 |
| OfficeStores | 0.85 | 0.95 | 0.9 | 154 |
| School | 1 | 0.91 | 0.95 | 233 |
| Residential | 0.99 | 0.96 | 0.98 | 3926 |
| Shopping & Retail | 0.98 | 0.95 | 0.97 | 230 |
| Mix-use | 0.93 | 0.95 | 0.97 | 968 |
| Total | 0.95 | 0.96 | 0.97 | 5872 |

lower precision rates. In general, adopting a multimodal data approach achieved an accuracy rate of 95 %.

To further illustrate the differences between the results and actual situations of different types of buildings, the following formula is used to calculate the difference ratio:

$$D = \frac{P - P}{\frac{1}{2}(P + G)} \times 100\%$$

Where $P$ represents the area of building types calculated by the presenting method, $G$ represents the area of building types published by the Shanghai government (ShangHai Statistical Yearbook 2022, 2022), and $D$ represents the difference between them. The results are shown in Table 7.

Table 7 shows that the residential data exhibits the highest accuracy, a trend consistent across various areas. Schools rank second in accuracy, following residential buildings. The observed variance in hospitals accuracy might arise from including small hospitals within the count of large hospitals, potentially differing from the government's statistical methods. In the case of hotels and office buildings, most analyzed structures in this study fall under the category of mixed-use buildings, which are not accounted for in official building statistics, and the precise categorization of these buildings is challenging, potentially contributing to the observed inaccuracies. Moreover, the data indicates that Jing'an and Putuo, known for their greater affluence, exhibit higher accuracy in comparison to the Xuhui, Yangpu, and Changning districts.

**Table 7**
Differences with various types of buildings in each district.

| Building type | Jing'an | Putuo | Xuhui | Yangpu | Changning |
|---|---|---|---|---|---|
| Residential | −0.04 | −0.08 | 0.07 | −0.05 | 0.01 |
| School | −0.40 | 0.10 | −0.24 | −0.11 | −0.35 |
| Office | −0.02 | −0.19 | 0.25 | 0.18 | −0.83 |
| Shopping & Retail | −0.03 | 0.06 | −0.33 | −0.45 | −0.67 |
| Hospital | 0.47 | 0.21 | 0.43 | 0.07 | 0.48 |
| Hotel | −0.01 | −0.66 | −0.38 | −1.02 | −0.40 |

### 4.2. Prototype models of shanghai

In accordance with the process outlined in Section 2.6, prototype buildings specific to Shanghai were developed. The geometric parameters of these prototypes are derived from the average values of various types of buildings in Shanghai. Certain non-geometric parameters of the buildings were also adjusted in line with Shanghai's mandatory standards. Alterations in these parameters can lead to variations in the EUI of the prototype buildings. The prototype buildings used in this study and their corresponding average length, width, height, and number of floors are presented in Table 8.

Fig. 15 presents the energy density of 21 prototype buildings across

**Table 8**
Prototype attributes of Shanghai.

| Prototype | Length (m) | Width (m) | Height (m) | Floor | Building Area ($m^2$) |
|---|---|---|---|---|---|
| Low-Rise Residential | 22.39 | 11.22 | 5.89 | 2 | 503 |
| Mid-Rise Residential | 36.54 | 13.81 | 16.14 | 5 | 2018 |
| High-Rise Residential | 46.91 | 16.05 | 42.53 | 14 | 7530 |
| Primary/Secondary School | 38.17 | 18.58 | 12.77 | 4 | 2127 |
| Small Office | 39.93 | 19.24 | 16.09 | 5 | 5379 |
| Large Office | 70.84 | 40.52 | 64.39 | 21 | 57,411 |
| Small Hotel | 64.25 | 27.18 | 19.46 | 6 | 6985 |
| Large Hotel | 81.88 | 44.80 | 42.44 | 14 | 36,686 |
| Restaurant | 57.40 | 26.52 | 17.24 | 6 | 1522 |
| Retail stand alone | 42.30 | 18.16 | 12.07 | 3 | 863 |
| Shopping Mall | 55.45 | 24.29 | 12.15 | 4 | 26,620 |
| Hospital | 36.04 | 17.14 | 13.70 | 4 | 1853 |
| Low-rise Residential-Shopping | 29.21 | 11.07 | 5.90 | 2 | 646 |
| Mid-rise Residential-Shopping | 44.80 | 19.21 | 16.86 | 6 | 4302 |
| High-rise Residential-Shopping | 54.63 | 28.29 | 58.81 | 20 | 15,456 |
| Residential-Restaurant | 71.15 | 26.40 | 15.16 | 5 | 13,148 |
| Residential-Hotel | 58.70 | 20.01 | 22.05 | 7 | 8221 |
| Office-Hotel | 53.62 | 29.46 | 47.56 | 16 | 18,954 |
| Office-Shopping | 47.12 | 22.96 | 14.65 | 5 | 10,818 |
| Shopping-Restaurant | 39.58 | 17.28 | 11.64 | 4 | 29,621 |
| Shopping-Hotel | 37.88 | 19.10 | 16.62 | 6 | 2894 |

three distinct eras. To ensure consistency in the figure's representation, the Gas EUI is converted into kWh/m$^2$. The first row focuses on residential buildings, predominantly using gas for domestic hot water and cooking. Lower-level residential buildings, particularly those with attics, exhibit higher average gas consumption per unit area. Energy consumption is significantly higher in buildings that combine residential and hotel functions, a typical occurrence in Shanghai, in comparison to purely residential structures. The second and third rows primarily focus on commercial buildings. Hospitals exhibit the highest energy consumption, with a total EUI ranging from 494 kWh/m$^2$ to 688 kWh/m$^2$. Office buildings, characterized by a higher rate of electrification and lower gas usage, show an overall energy consumption ranging from 97 kWh/m$^2$ to 164 kWh/m$^2$. Large office buildings employ centralized heating, resulting in increased gas usage and an overall EUI range of 115 kWh/m$^2$ to 152 kWh/m$^2$. Additionally, most office buildings incorporate retail stores serving office workers, resulting in increased energy consumption compared to buildings exclusively used for office purposes, with a total consumption range of 162 kWh/m$^2$ to 248 kWh/m$^2$. Restaurants experience considerable gas usage, primarily for cooking purposes. The mainstream types of contemporary commercial buildings in Shanghai are mixed-use establishments integrating both restaurants and shopping facilities. These buildings exhibit relatively higher energy consumption, ranging from 173 kWh/m$^2$ to 336 kWh/m$^2$, in contrast to strictly shopping malls, with energy consumption ranging between 107 kWh/m$^2$ and 225 kWh/m$^2$. Tourism hotels, shopping malls, and healthcare buildings entail increased demands for indoor comfort and accommodate larger population densities. The continuous operation of heating and cooling systems throughout the year leads to an overall increase in energy consumption. Conversely, schools feature the lowest overall energy consumption due to the absence of the peak energy-consuming cooling and heating seasons during the summer break.

Fig. 16 compares the results of prototype buildings, measured data from the papers, and government data. It is worth noting that some of the composite buildings still lack relevant reference literature to provide measurement ranges. It can be observed that the majority of prototype buildings fall within the reference range.

However, the EUIs for certain building periods fell outside the measured values, which can be attributed to three main reasons. Firstly, the data sources for these measurements did not cover the three eras outlined in this study, resulting in discrepancies. Secondly, the complexity of the buildings in question could be a factor, particularly if the prototype buildings used in this study were not fully adapted to Shanghai's context. Future research should aim to develop prototypes that are more representative of Shanghai or reassess the representativeness of the measured buildings. Specifically, for hospitals built before 2005, their EUIs substantially exceeded the measurement range,

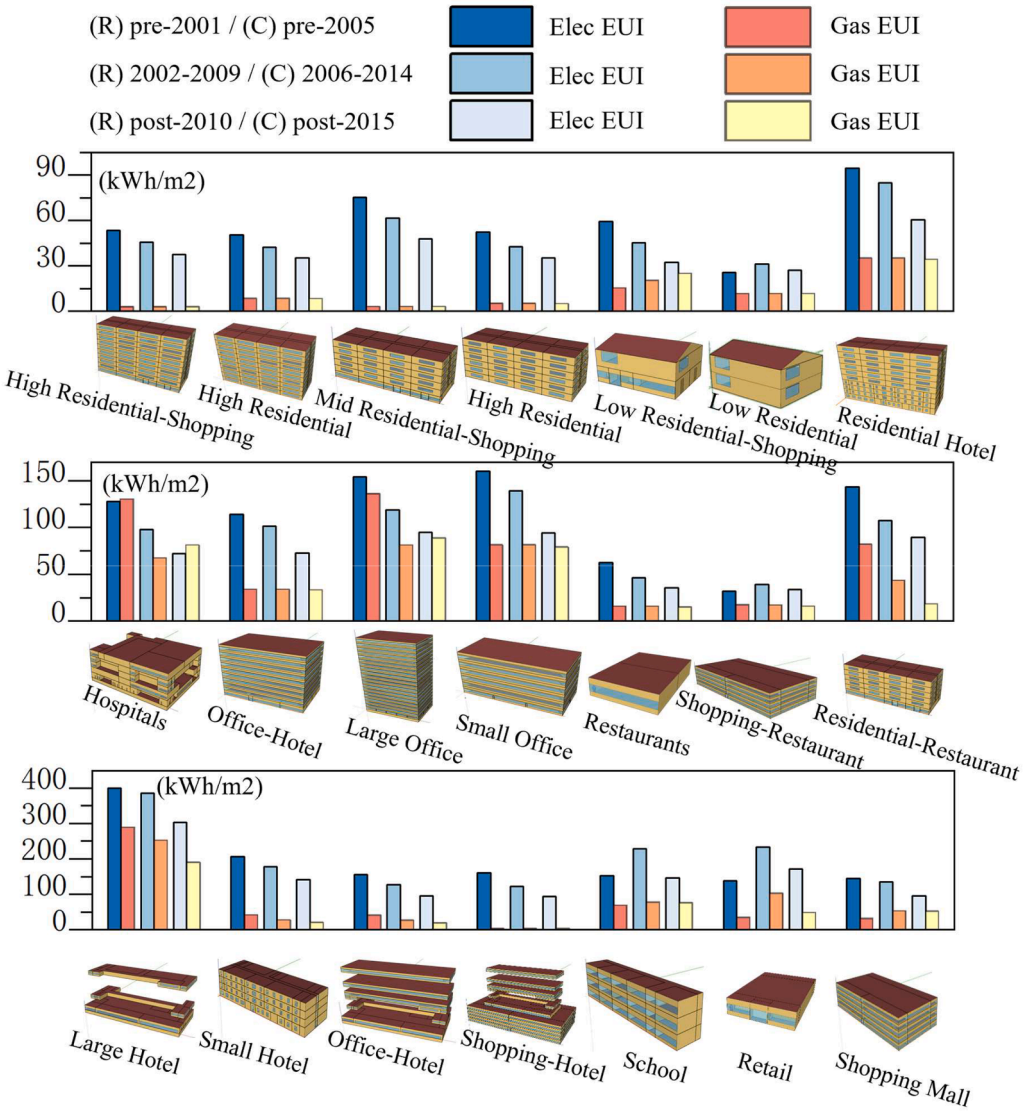



**Fig. 15.** The prototype buildings and their respective electrical and gas consumption densities, with gas density units converted to kWh/m$^2$.

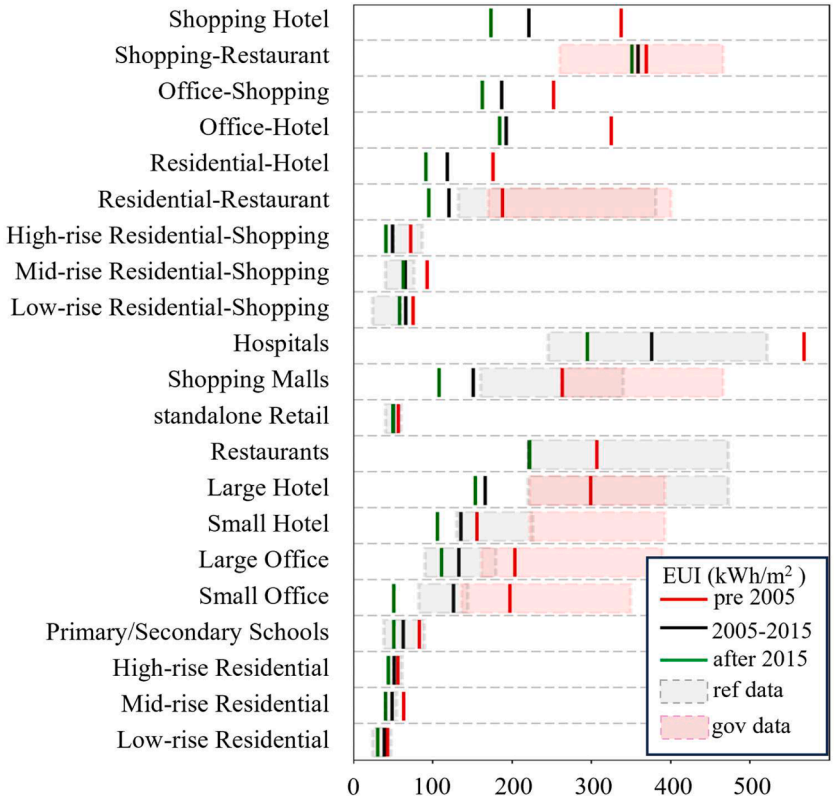



**Fig. 16.** Comparison between Simulation Results and References.

likely because most Shanghai hospitals were constructed post-2005 and have been updated to conform to newer standards. For shopping malls, the primary issue seems to stem from the second factor, leading to modeling results that are lower than the actual measurements. This may be due to the use of prototype buildings based on American standards by the DOE, which do not accurately reflect Shanghai's unique commercial environment. The discrepancy in EUIs between large and small hotels is primarily related to the first factor, with a greater emphasis on buildings established earlier. In mixed-use buildings, the majority of results were consistent with measurements, with the exception of Residential-Restaurant buildings, where lower results were observed, largely attributed to the buildings' complexity. On the other hand, when


**Fig. 17.** Detailed building energy consumption in Shanghai.

comparing the results of Shopping-Restaurant and Shopping Mall, it can be seen that Shopping-Restaurant aligns well with the middle region of the government measurement range. This indicates that in Shanghai, shopping malls are more akin to complexes like Shopping-Restaurant.

### 4.3. UBEM for Shanghai city

After completing the modeling of various prototype buildings in Shanghai, the energy consumption of each building in Shanghai was calculated based on the EUI, era, type, and total area of the buildings. Regarding the details of individual buildings, as shown in Fig. 17, in relatively complex commercial buildings, the method used in comparison to the untreated method performs well in multi-component buildings. It does not only cover the main or partial components but instead provides comprehensive coverage. In residential buildings, this method is effective in identifying mixed-use buildings, thereby further improving the granularity of simulation modeling in residential buildings. However, some buildings, such as factories and warehouses, still have unaccounted energy consumption. In a few cases, certain commercial buildings cannot be categorized, or their energy consumption cannot be determined. For these buildings, the method used is to assign their EUI based on the average of other similar commercial buildings. There is also a small number of residential buildings (less than 5 %) where the type couldn't be accurately determined.

To better present the energy consumption distribution across Shanghai, this study has divided the city into discrete grids with a unit size of 2 square kilometers. The energy consumption distribution in Shanghai's buildings is illustrated in Fig. 18.

Due to the difficulty in obtaining actual energy consumption from each building, a comparison can only be made with the data provided by the government. The simulation results indicate that the 539,119 buildings in Shanghai consumed a total of 82,471 GWh of electricity and 36,374 GWh of gas. The electricity consumption differs from the government's published value of 78,079 GWh by approximately 5 %. A relevant review (Reinhart & Cerezo Davila, 2016) indicates that this level of difference falls within a relatively good range for modeling on such a large scale.

Table 9 summarizes the total energy consumption of various types of buildings in Shanghai. It is evident that despite the lower energy consumption of residential buildings, their large total area results in the highest overall energy consumption. The second period, spanning from 2002 to 2009 (Residential) and 2006 to 2014 (Commercial), coincided with a period of rapid development in Shanghai. During this time, a significant number of buildings, particularly commercial ones, were constructed. Additionally, the proportion of mixed-use buildings should not be underestimated. These early constructions, due to the implementation of less rigorous efficiency standards, have substantial potential for optimization in both their maintenance structures and energy-using equipment.

**Table 9**
Electricity consumption of different building types in Shanghai (in GWh).

| Building type | First period | Second period | Third period | Total electricity |
|---|---|---|---|---|
| High-Rise Residential | 8207.52 | 6501.31 | 4248.06 | 18,956.89 |
| High-rise Residential-Shopping | 97.42 | 133.18 | 105.76 | 336.36 |
| Hospital | 963.25 | 2171.77 | 764.29 | 3899.31 |
| Large Hotel | 89.78 | 275.10 | 52.43 | 417.31 |
| Large Office | 1022.62 | 2798.85 | 973.98 | 4795.45 |
| Low-Rise Residential | 988.71 | 1333.45 | 779.06 | 3101.22 |
| Low-rise Residential-Shopping | 10.82 | 3.54 | 1.98 | 16.35 |
| Mid-Rise Residential | 3687.33 | 6347.93 | 2519.02 | 12,554.27 |
| Mid-rise Residential-Shopping | 22.62 | 29.60 | 9.45 | 61.68 |
| Office-Hotel | 14.89 | 74.31 | 40.74 | 129.93 |
| Office-Shopping | 5334.37 | 11,750.09 | 3236.67 | 20,321.12 |
| Primary/Secondary School | 482.85 | 936.25 | 351.45 | 1770.56 |
| Residential-Hotel | 695.92 | 1375.76 | 590.15 | 2661.83 |
| Residential-Restaurant | 2.28 | 23.99 | 6.12 | 32.39 |
| Restaurant | 23.95 | 123.48 | 30.52 | 177.95 |
| Retail stand alone | 44.12 | 267.41 | 53.95 | 365.49 |
| Shopping Mall | 769.86 | 1474.40 | 428.47 | 2672.73 |
| Shopping-Hotel | 400.14 | 1001.28 | 380.47 | 1781.89 |
| Shopping-Restaurant | 974.04 | 4770.40 | 1504.96 | 7249.40 |
| Small Hotel | 45.17 | 141.30 | 46.63 | 233.11 |
| Small Office | 248.61 | 547.35 | 139.80 | 935.75 |
| Total Electricity (GWh) | 24,126.26 | 42,080.75 | 16,263.96 | 82,470.97 |

## 5. Discussion

### 5.1. Data quality and UBEM accuracy

The quality of data sources significantly affects the accuracy of UBEM, as noted by Nouvel et al. (2017). Errors in these sources can propagate to UBEM results through a process known as 'forward transmission'. In this study, we examine the impact of different open

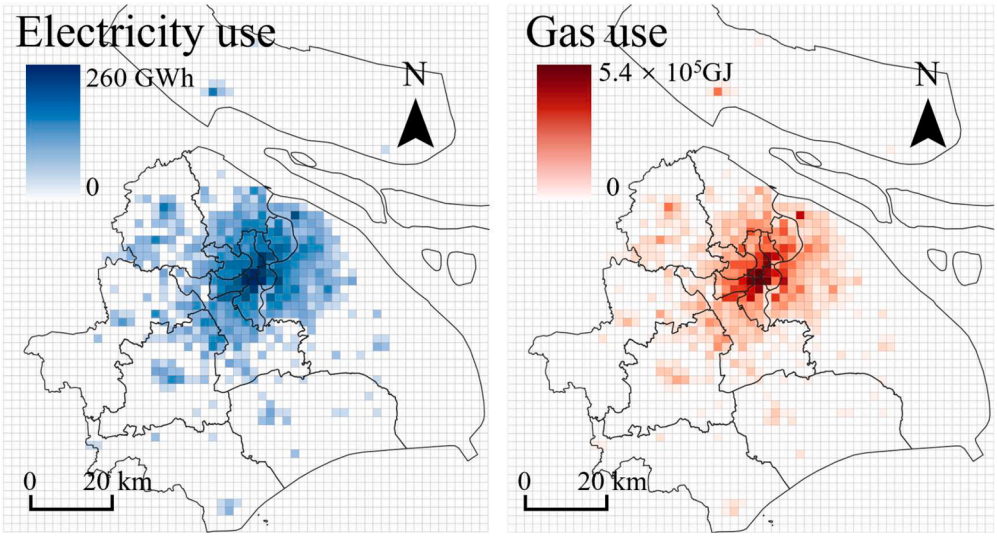



**Fig. 18.** The block distribution of annual building electricity use and gas use.

data sources on the accuracy of UBEM, focusing on three key aspects: completeness, accuracy, and timeliness.

Beginning with completeness, which primarily pertains to GIS-related issues, including POIs, AOIs, and building footprints, the research utilized OpenStreetMap. Despite its widespread use, its quality in China and developing regions is inadequate for a comprehensive UBEM due to missing building height/floor data and low coverage.

Due to the absence of real-world benchmarks, the accuracy of open data is challenging to validate accurately, especially at larger scales. The lack of validated data is a common issue in UBEM research, extensively discussed in numerous related papers (Hong et al., 2020a; Oraiopoulos & Howard, 2022; Reinhart & Cerezo Davila, 2016). Obtaining large-scale building energy consumption data is a widespread challenge in this domain. Oraiopoulos & Howard (2022) surveyed 535 UBEM papers, revealing that only 47 studies confirmed the potential accuracy of UBEMs, and these studies primarily emanated from regions with established building databases, such as Singapore and some cities in the United States. Open data, mainly from crowdsourcing, suffers from geographical accuracy inconsistencies, making manual area calibration unreliable for overall accuracy. Baidu and Amap, key for UBEM, differ in their building height data, using floor count and actual height. Prior research, typically using one source and estimating the other with a standard 3 m floor height, faces conflicts between sources as shown in Section 2.3.2, risking significant discrepancies and errors in UBEM results. This study leverages Amap and Baidu POIs for real-time tracking of building renovations, alongside ultra-high-resolution satellite imagery for temporal analysis. It highlights the challenge of inconsistent image capture times across cities, affecting building age accuracy and introducing biases. Recent imagery integration by providers mitigates this, though historical data still presents discrepancies.

### 5.2. Scalability of UBEM modeling with open data

UBEM has increasingly matured, with tools such as CityBES, AutoBPS, and UBEM.io capable of swiftly generating UBEMs based on existing data. Currently, the primary limitations of UBEM applications stem from data sourcing and processing. In existing research, these processes are often manually executed by experienced GIS experts, thus being perceived as time-consuming, labor-intensive, and costly (Davila et al., 2016). Therefore, this aspect is also widely recognized as one of the significant challenges in UBEM (Hong et al., 2020a). Discussions on data source processing are prevalent in other studies as well. Remmen et al. (2018)emphasize the importance of open data in urban energy modeling, particularly in dynamic simulations and large-scale urban scenarios, through their comparative analysis of data sources. This study focuses on simplifying and optimizing a universal methodology through the fusion of multi-source data and artificial intelligence techniques, leveraging accessible open GIS information and satellite imagery to categorize buildings by type and era. By employing a multi-source data fusion approach, this paper significantly enhances the robustness, reliability, coverage, and accuracy of the commonly used method. The data sources utilized in this study are standard in the remote sensing domain and are readily accessible as open data. By combining existing UBEM modeling tools, the complexity of developing UBEM can be significantly reduced, thereby enhancing the reliability of establishing UBEMs using open data. Furthermore, the modeling software AutoBPS utilized in this study fully encompasses prototype buildings and related standards from the United States, enabling the method to be automated and extended to any region within the country.

### 5.3. Limitations and future work

This research encounters several limitations in its aim to expand to more cities:

1) Calibration of floor counts is based on a simple metric of the difference between footprint area and total built area. Implementing more sophisticated methods would enhance accuracy.
2) This study introduces various methods for classifying building types, but validation is limited, not accounting for geographic variations.
3) The study mainly uses a prototypical building aggregation approach, which, while cost-effective, overlooks environmental interactions.
4) Due to the lack of detailed final energy consumption data in the publicly available information, this study was unable to conduct an accurate verification against actual results.

Given the diverse range of building types in urban contexts, the building models used in this study might not cover every possible scenario, including facilities like factories and laboratories, known for their considerable variability in energy use.Several aspects require further investigation in research:

1) Enhancing accuracy through expanded GIS methods, such as satellite imagery for building type deduction and street views for identification.Refining mixed-use building models by considering various mixed-use ratios.Further discussion on the thermal exchange between buildings is necessary.

In summary, open data fusion is a practical approach for developing reliable urban building energy models, with tailoring to specific cities and integrating more sources enhancing effectiveness. Its limitations suggest using advanced GIS, computer vision, and data mining for greater accuracy, and including mixed-use buildings and occupant factors improves simulations. This study offers a framework for open data in city energy planning and sustainability.

## 6. Conclusions

In conclusion, this study presents a methodology to developing an urban building energy model for Shanghai using multi-source open data. Through integrating and cross-validating data from various sources, including Baidu Map, Amap, Google Earth, AI Earth, GHSL, Anjuke, and Lianjia, this research significantly expanded the coverage and improved the reliability of open data for urban building modeling.

In total, 609,763 building footprints and 539,119 buildings in Shanghai were collected, covering an area of 1567,921,014 $m^2$. By employing spatial analysis, clustering, supervised learning, and unsupervised learning, the collected buildings were categorized into 21 prototypes, with over 93.22 % of buildings successfully classified. The overall distribution shows residential buildings account for 62.4 %, office buildings account for 11.3 %, hospitals account for 2.5 %, and the remaining 23.8 % are other types. Compared to manual sampling, the classification accuracy reached 95 %.

Furthermore, historical satellite imagery and community boundary data were combined to determine the construction year for over 95 % of buildings. Prototype models were then developed in AutoBPS for 63 main building types using mandatory local standards, literature references, and OpenStudio standards. These models represented the characteristics of real buildings in Shanghai. EnergyPlus simulations of these prototype buildings yielded energy use intensity values that aligned with ranges reported in government data.

By integrating the classified buildings with AutoBPS-Param, a comprehensive bottom-up urban building energy model was established for Shanghai. The results validate the reliability of the multi-source data fusion approach for developing UBEM. This study provides an effective workflow and valuable insights for leveraging open data to advance city-scale energy modeling and urban sustainability.

## CRediT authorship contribution statement

**Chengcheng Song:** Writing – review & editing, Writing – original

draft, Visualization, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Zhang Deng:** Writing – review & editing, Supervision. **Wenxian Zhao:** Data curation. **Yue Yuan:** Writing – review & editing. **Mengyue Liu:** Methodology. **Shen Xu:** Writing – review & editing, Funding acquisition. **Yixing Chen:** Writing – review & editing, Supervision, Funding acquisition.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## Acknowledgments

This work was funded by "A Project Supported by the Scientific Research Fund of Hunan Provincial Education Department, China (No. 23A0033)". Additional support was provided by "A Project Supported by the Scientific Research Fund of Hunan Provincial Education Department, China (No. 2023JGZD027)". Furthermore, this research was supported by the National Natural Science Foundation of China (No. 52378020).

## References

Ali, U., Shamsi, M. H., Bohacek, M., Purcell, K., Hoare, C., Mangina, E., & O'Donnell, J (2020). A data-driven approach for multi-scale GIS-based building energy modeling for analysis, planning and support decision making. *Applied Energy, 279*, Article 115834. https://doi.org/10.1016/j.apenergy.2020.115834

An, J., Wu, Y., Gui, C., & Yan, D. (2023). Chinese prototype building models for simulating the energy performance of the nationwide building stock. *Building Simulation*. https://doi.org/10.1007/s12273-023-1058-5

Ang, Y. Q., Berzolla, Z. M., & Reinhart, C. F. (2020). From concept to application: A review of use cases in urban building energy modeling. *Applied Energy, 279*. https://doi.org/10.1016/j.apenergy.2020.115738. Scopus.

Buckley, N., Mills, G., Reinhart, C., & Berzolla, Z. M. (2021). Using urban building energy modelling (UBEM) to support the new European Union's Green Deal: Case study of Dublin Ireland. *Energy and Buildings, 247*. https://doi.org/10.1016/j.enbuild.2021.111115. Scopus.

Cerezo Davila, C., Reinhart, C. F., & Bemis, J. L (2016). Modeling Boston: A workflow for the efficient generation and maintenance of urban building energy models from existing geospatial datasets. *Energy, 117*, 237–250. https://doi.org/10.1016/j.energy.2016.10.057

Chen, W., Zhou, Y., Wu, Q., Chen, G., & Yu, B. (2020). Urban building type mapping using geospatial data: A case study of Beijing, China. *Remote Sensing, 12*(17), 2805. https://doi.org/10.3390/rs12172805

Chen, Y., Hong, T., Luo, X., & Hooper, B. (2019). Development of city buildings dataset for urban building energy modeling. *Energy and Buildings, 183*, 252–265. https://doi.org/10.1016/j.enbuild.2018.11.008

Chen, Y., Wei, W., Song, C., Ren, Z., & Deng, Z. (2023). Rapid building energy modeling using prototype model and automatic model calibration for retrofit analysis with uncertainty. *Buildings, 13*(6), 1427. https://doi.org/10.3390/buildings13061427

Dall'O', G., Galante, A., & Torri, M (2012). A methodology for the energy performance classification of residential building stock on an urban scale. *Energy and Buildings, 48*, 211–219. https://doi.org/10.1016/j.enbuild.2012.01.034

Deng, Z., Chen, Y., Pan, X., Peng, Z., & Yang, J. (2021). Integrating GIS-based point of interest and community boundary datasets for urban building energy modeling. *Energies, 14*(4), 1049. https://doi.org/10.3390/en14041049

Deng, Z., Chen, Y., Yang, J., & Causone, F. (2023). AutoBPS: A tool for urban building energy modeling to support energy efficiency improvement at city-scale. *Energy and Buildings, 282*, Article 112794. https://doi.org/10.1016/j.enbuild.2023.112794

Deng, Z., Chen, Y., Yang, J., & Chen, Z. (2022). Archetype identification and urban building energy modeling for city-scale buildings based on GIS datasets. *Building Simulation, 15*(9), 1547–1559. https://doi.org/10.1007/s12273-021-0878-4

Du, S., Zhang, F., & Zhang, X. (2015). Semantic classification of urban buildings combining VHR image and GIS data: An improved random forest approach. *ISPRS Journal of Photogrammetry and Remote Sensing, 105*, 107–119. https://doi.org/10.1016/j.isprsjprs.2015.03.011

El Kontar, R., Polly, B., Charan, T., Fleming, K., Moore, N., Long, N., & Goldwasser, D. (2020). *URBANopt: An open-source software development kit for community and urban district energy modeling: Preprint*. https://www.osti.gov/biblio/1677416.

Fonseca, J. A., Nguyen, T.-A., Schlueter, A., & Marechal, F. (2016). City Energy Analyst (CEA): Integrated framework for analysis and optimization of building energy systems in neighborhoods and city districts. *Energy and Buildings, 113*, 202–226. https://doi.org/10.1016/j.enbuild.2015.11.055

He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2017). Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Hong, T., Chen, Y., Lee, S. H., & Piette, M. A. (2016). *CityBES: A web-based platform to support city-scale building energy efficiency* (p. 9).

Hong, T., Chen, Y., Luo, X., Luo, N., & Lee, S. H. (2020a). Ten questions on urban building energy modeling. *Building and Environment, 168*, Article 106508. https://doi.org/10.1016/j.buildenv.2019.106508

Hong, Y., Ezeh, C. I., Deng, W., Hong, S.-H., Peng, Z., & Tang, Y. (2020b). Correlation between building characteristics and associated energy consumption: Prototyping low-rise office buildings in Shanghai. *Energy and Buildings, 217*, Article 109959. https://doi.org/10.1016/j.enbuild.2020.109959

HosseiniHaghighi, S., De Uribarri, P. M.Á., Padsala, R., & Eicker, U. (2022). Characterizing and structuring urban GIS data for housing stock energy modelling and retrofitting. *Energy and Buildings, 256*, Article 111706. https://doi.org/10.1016/j.enbuild.2021.111706

Hu, S., Zhang, Y., Yang, Z., Yan, D., & Jiang, Y. (2022). Challenges and opportunities for carbon neutrality in China's building sector—Modelling and data. *Building Simulation, 15*(11), 1899–1921. https://doi.org/10.1007/s12273-022-0912-1

Jin, X., Zhang, C., Xiao, F., Li, A., & Miller, C. (2023). A review and reflection on open datasets of city-level building energy use and their applications. *Energy and Buildings, 285*, Article 112911. https://doi.org/10.1016/j.enbuild.2023.112911

Kavgic, M., Mavrogianni, A., Mumovic, D., Summerfield, A., Stevanovic, Z., & Djurovic-Petrovic, M. (2010). A review of bottom-up building stock models for energy consumption in the residential sector. *Building and Environment, 45*(7), 1683–1697. https://doi.org/10.1016/j.buildenv.2010.01.021

Lu, Z., Im, J., Rhee, J., & Hodgson, M. (2014). Building type classification using spatial and landscape attributes derived from LiDAR remote sensing data. *Landscape and Urban Planning, 130*, 134–148. https://doi.org/10.1016/j.landurbplan.2014.07.005

Luo, X., Hong, T., & Tang, Y.-H. (2020). Modeling thermal interactions between buildings in an urban context. *Energies, 13*(9), 2382. https://doi.org/10.3390/en13092382

Nouvel, R., Zirak, M., Coors, V., & Eicker, U. (2017). The influence of data quality on urban heating demand modeling using 3D city models. *Computers, Environment and Urban Systems, 64*, 68–80. https://doi.org/10.1016/j.compenvurbsys.2016.12.005

Oraiopoulos, A., & Howard, B. (2022). On the accuracy of urban building energy modelling. *Renewable and Sustainable Energy Reviews, 158*, Article 111976. https://doi.org/10.1016/j.rser.2021.111976

Österbring, M., Mata, É., Thuvander, L., Mangold, M., Johnsson, F., & Wallbaum, H. (2016). A differentiated description of building-stocks for a georeferenced urban bottom-up building-stock model. *Energy and Buildings, 120*, 78–84. https://doi.org/10.1016/j.enbuild.2016.03.060

Pasichnyi, O., Levihn, F., Shahrokni, H., Wallin, J., & Kordas, O. (2019a). Data-driven strategic planning of building energy retrofitting: The case of Stockholm. *Journal of Cleaner Production, 233*, 546–560. https://doi.org/10.1016/j.jclepro.2019.05.373

Pasichnyi, O., Wallin, J., & Kordas, O. (2019b). Data-driven building archetypes for urban building energy modelling. *Energy, 181*, 360–377. https://doi.org/10.1016/j.energy.2019.04.197

Peng, J., Kimmig, A., Wang, J., Liu, X., Niu, Z., & Ovtcharova, J. (2021). Dual-stage attention-based long-short-term memory neural networks for energy demand prediction. *Energy and Buildings, 249*, Article 111211. https://doi.org/10.1016/j.enbuild.2021.111211

Perwez, U., Shono, K., Yamaguchi, Y., & Shimoda, Y. (2023). Multi-scale UBEM-BIPV coupled approach for the assessment of carbon neutrality of commercial building stock. *Energy and Buildings, 291*, Article 113086. https://doi.org/10.1016/j.enbuild.2023.113086

Prototype Building Models. (2021). *Building energy codes program*. https://www.energycodes.gov/prototype-building-models#Commercial.

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

Reinhart, C. F., Dogan, T., Jakubiec, J. A., Rakha, T., & Sang, A. (2013). *Umi-an urban simulation environment for building energy use, daylighting and walkability*.

Remmen, P., Lauster, M., Mans, M., Fuchs, M., Osterhage, T., & Müller, D. (2018). TEASER: An open tool for urban energy modelling of building stocks. *Journal of Building Performance Simulation, 11*(1), 84–98. https://doi.org/10.1080/19401493.2017.1283539

Robinson, D., Haldi, F., Kämpf, J., Leroux, P., Perez, D., Rasheed, A., & Wilke, U. (2009). *Citysim: Comprehensive micro-simulation of resource flows for sustainable urban planning*.

Shanghai Municipal Commission of Housing and Urban-Rural Development. (2021). *Analysis Report on Shanghai Municipality's National Government Office Buildings and Large Public Buildings in 2021* (in Chinese).

ShangHai Statistical Yearbook 2022. (2022). https://tjj.sh.gov.cn/tjnj/20230206/804acea250d44d2187f2e37d2e5d36ba.html.

Swan, L. G., & Ugursal, V. I. (2009). Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. *Renewable and Sustainable Energy Reviews, 13*(8), 1819–1835. https://doi.org/10.1016/j.rser.2008.09.033

Wang, C., Ferrando, M., Causone, F., Jin, X., Zhou, X., & Shi, X. (2022a). Data acquisition for urban building energy modeling: A review. *Building and Environment, 217*, Article 109056. https://doi.org/10.1016/j.buildenv.2022.109056

Wang, D., Landolt, J., Mavromatidis, G., Orehounig, K., & Carmeliet, J. (2018). CESAR: A bottom-up building stock modelling tool for Switzerland to address sustainable

energy transformation strategies. *Energy and Buildings, 169*, 9–26. https://doi.org/10.1016/j.enbuild.2018.03.020

Wang, M., Yu, H., Yang, Y., Jing, R., Tang, Y., & Li, C. (2022b). Assessing the impacts of urban morphology factors on the energy performance for building stocks based on a novel automatic generation framework. *Sustainable Cities and Society, 87*, Article 104267. https://doi.org/10.1016/j.scs.2022.104267

Wu, H., Ding, R., Zhao, H., Chen, B., Xie, P., Huang, F., & Zhang, M. (2022). Forging multiple training objectives for pre-trained language models via meta-learning. *CoRR*. https://doi.org/10.48550/arXiv.2210.10293. *abs/2210.10293*.

Yu, Z., Geng, Y., He, Q., Oates, L., Sudmant, A., Gouldson, A., & Bleischwitz, R. (2021). Supportive governance for city-scale low carbon building retrofits: A case study from Shanghai. *Climate Policy, 21*(7), 884–896. https://doi.org/10.1080/14693062.2021.1948383