Assessing the Potential of Open Source Data for Expanding the Building Database: A case study in Huangpu District, Shanghai

Chengcheng Song[1], Yixing Chen[1,2], Zhang Deng[1] , Yue Yuan[1]
[1]College of Civil Engineering, Hunan University, Changsha 410082, China;
[2]Key Laboratory of Building Safety and Energy Efficiency of Ministry of Education, Hunan University, Changsha 410082, China

Abstract

Urban Building Energy Modelling necessitates extensive building information. This study examines the potential application of open data for urban building energy modeling, with a specific focus on the Huangpu District of Shanghai, China. This study initially gathered open data for the area and subsequently merged them using geographic information system information processing software. Finally, it was assessed how much data on building size could be provided by combining retrieved open data with current open data processing methods. The results showed that open data has great potential for expanding building data.

Highlights

- Proposed an evaluation framework to assess the potential of open data for UBEM
- 14,083 building footprints were covered by this study
- 70% of building footprint information in Huangpu district could be extended by open data

Introduction

Cities are a significant component of energy consumption, and China's urbanization process has been growing year by year(*STATISTICAL COMMUNIQUÉ OF THE PEOPLE'S REPUBLIC OF CHINA ON THE 2022 NATIONAL ECONOMIC AND SOCIAL DEVELOPMENT*, n.d.), and China has committed to reaching carbon peak by 2023. Therefore, it is crucial to adjust the energy consumption of urban buildings. Urban building energy modeling (UBEM) is the foundation of urban building energy-saving renovation, which refers to bottom-up, physics-based building energy models from a city scale (Reinhart & Cerezo Davila, 2016). UBEM takes into account the building's physical characteristics, such as its shape, category, systems, and other factors that affect energy use. By simulating the building's energy use under different conditions, energy modelers can identify the most effective strategies for reducing energy consumption and improving the urban buildings' efficiency.

However, UBEM requires large amounts of building information. Acquiring this information can be challenging, as detailed information for buildings on an urban scale is difficult to obtain. Generally speaking, input data consists of both geometric and non-geometric data(Wang, 2022). Geometric data can be used to build a 3D model of the building, including its footprints, height, and window-to-wall ratios (WWRs). Non-geometric data includes information about the building envelope, building material, and the heating, ventilation, and air conditioning (HVAC) systems.

Open data refers to publicly available data, which has the potential to help solve the challenge of acquiring building information for UBEM, but the potential varies between different cities due to differences in the coverage of open data and the fact that it cannot be directly used in UBEM. Open datasets are datasets that have been sorted and can be partially applied to UBEM. Such datasets often appear in a few developed countries, such as the U.S., the U.K., Ireland, and Singapore(Jin et al., 2023). However, in other cities, the development of open datasets is still in its early stages. Nevertheless, there are a lot of open data resources worth utilizing in these cities, and therefore, it is necessary to make full use of these open data resources to build open data. Various map service providers, such as Google Earth(*Google Earth*, n.d.), Baidu Map(*Baidu Map*, n.d.), Amap(also known as GaoDe Map)(*Amap*, n.d.), and open source map OpenStreetMap(OSM) (*OpenStreetMap*, n.d.), are the most widely used open urban building data sources. Additionally, public information from some real estate websites and government public websites can also be obtained. The data available from these sources mainly includes building shapefiles, building GIS information, satellite imagery, street view images, points of interest (POIs), and areas of interest (AOIs). These open data, however, need to be processed and integrated before it can be used in UBEM. The processing of open data has received considerable scholarly attention in recent years.

For the obtaining of geometric data, there are some building footprints shapefiles can be directly downloaded and used, like OSM, However, the timeliness and effectiveness of OSM data are both limited, especially in developing countries. Shapefiles can be directly used for UBEM, but to ensure the validity of the data, it is necessary to merge and align shapefiles from multiple sources. The footprints can also be extracted from the satellite imagery, Yang et al.(Yang et al., 2018) compared four deep neural network architectures for extracting building footprints from the satellite imagery at city scale, and proposes an efficient CNN-based framework for generating high-quality building maps in United States. Wang(Wang et al., 2021) estimated building heights using building shadows on satellite imagery and WWRs using street view images. Szcześniak et al.(Szcześniak et al., 2022) proposes an automated method to extract facade

https://doi.org/10.26868/25222708.2023.1355

windows layouts and calculated the WWRs automatedly. Huang et al.(Huang et al., 2019) presented an approach detect the roof area of buildings and then estimated the city's solar potential. Lee et al. (Lee et al., 2019) proposed a data-driven approach called DeepRoof that utilizes satellite images to evaluate rooftop solar potential, which identifies the roof geometry and estimates the solar potential for every planar roof segment.

Non-geometric data require complicated handling. For example, the building envelope and the type of HVAC system in a building are often related to the standards when the building was constructed, and therefore obtaining the construction year of the building can be helpful. Additionally, the building's category can significantly impact its energy consumption, making it crucial to determine the building's category. However, it is extremely difficult to determine the year built and category of all buildings using open data alone. Obtaining such data often requires historical satellite images, POIs, and AOIs. Deng et al (Deng et al., 2021) determine the category and year built of 68,966 buildings in Changsha by using POIs within buildings that include AOIs and combining them with Community Boundary Datasets. Chen et al.(Chen et al., 2020) used natural language processing and text similarity methods to cluster POIs and then determine the category of a building using the POIs located within it. The building envelope of urban buildings is one of the most difficult information to obtain. In existing research(Deng et al., 2022; Liu et al., 2022), such data is mainly acquired by utilizing the standards executed during the construction period of the building.

Although the use of open data for building energy consumption simulation may have great potential, the reliability of these data deserves discussion. Wang et al. (Wang et al., 2022) argue that the quality of these data should be mainly considered from three aspects, namely, performance, availability, and cost. The cost of open data is not worth discussing, but open data is not the same as open-source data, and sometimes the acquisition of open data also can be hard.

Overall, it is very challenging to process and incorporate all open data into the modeling process of UBEM. To the best of the author's knowledge, most of the existing studies only introduce the acquisition of one or several data, and there is a lack of comprehensive evaluation of the availability of open data for a city. Therefore, this paper proposes a comprehensive evaluation framework to assess the coverage of building information based on open data in a city, and conducts a case study on the buildings in Huangpu District, Shanghai. It should be noted that the main focus of this paper is to assess the potential availability of open data within Huangpu District. The actual acquisition and processing of such data are beyond the scope of this paper.

Methodology

Workflow of this study

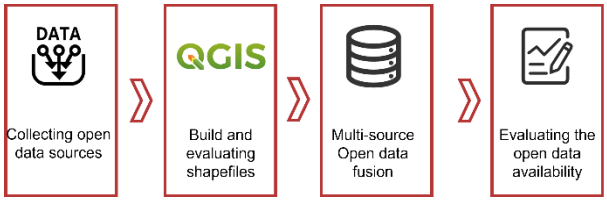

Figure 1: The workflow of this paper

Figure 1 outlines the research methodology for this study. Initially, open data was collected and acquired, and subsequently, building shapefiles were established using Q-GIS, a free and open source geographic information system that can create, edit, visualise, analyse and publish geospatial data. A basic building database was built and the suitability of various types of open data applications was evaluated based on the agglomeration and geographic distribution of the building structures. The resulting analysis allows for assessing the potential of open data to be employed in different contexts.

Introduction of the case study buildings

As is shown in Figure 2, Shanghai Huangpu District is located in the central part of Shanghai, China. It covers an area of approximately 20.46 square kilometers and has a population of about 658,000. The district has a humid subtropical climate, with hot and humid summers and cool, dry winters. Huangpu District in Shanghai presents a unique architectural landscape that includes traditional Chinese, modern skyscrapers, European-style, and industrial buildings. The district is a good case study for UBEM to analyse the environmental impacts of different building types and promote energy-efficient and sustainable urbanization.

https://doi.org/10.26868/25222708.2023.1355

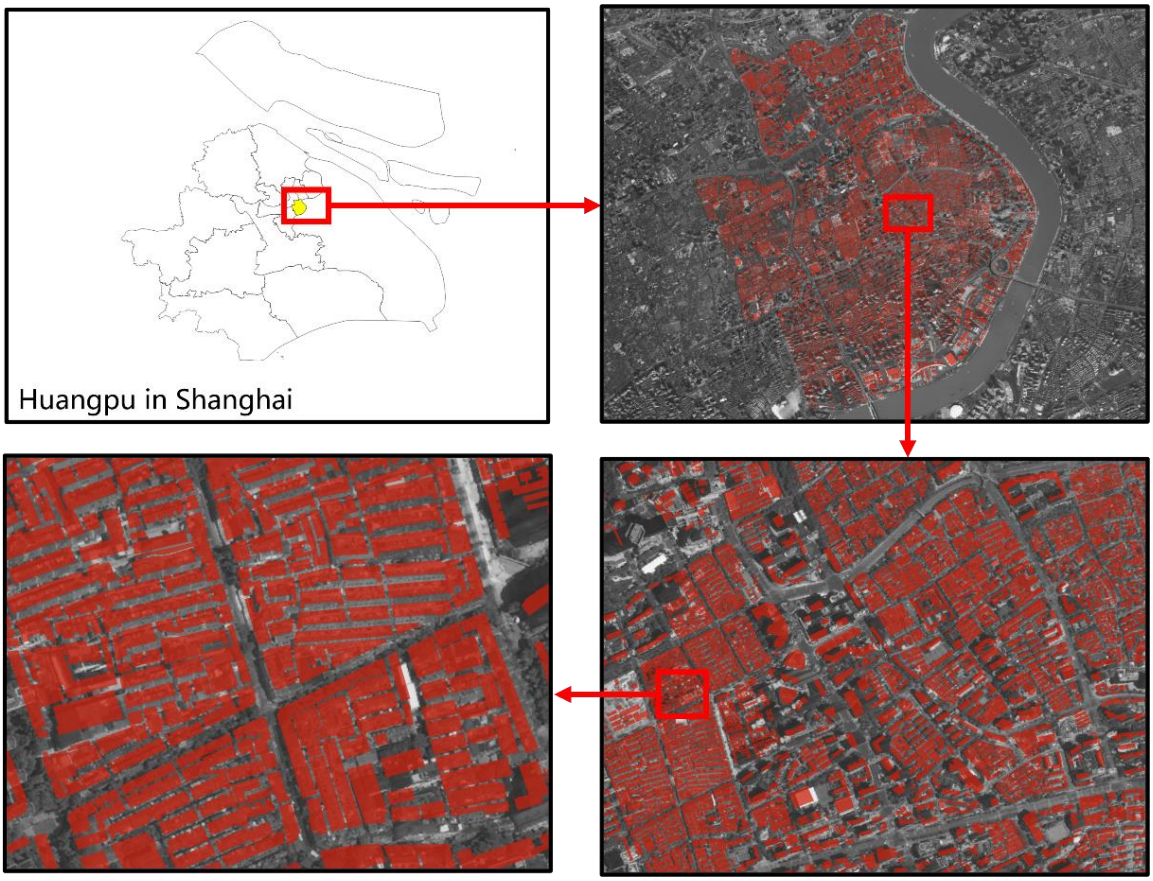

Figure 2: Introduction of the case study buildings

Acquisition and assessment of open data

In existing studies, the main open data utilized includes satellite imagery, historical satellite imagery, building shapefiles, street view images, POIs, AOIs, and real estate website information. The following outlines the acquisition methods for each open-source data type, including the reasons behind selecting one channel over another if multiple options are available. It should be noted that the data collected in this study only represents publicly available resources, and its quality and timeliness may vary. It does not reflect the actual performance of the data source websites and cannot be used as a reference to evaluate various information sources.

Satellite images

The selection of satellite images comprises the following maps: Amap, Google Earth, Baidu Map, and Tianditu, with Tianditu being developed by the Chinese government and launched in 2010, while the other maps are provided by internet map service providers. The selection of the reference satellite images considers two factors, namely the shooting time and the resolution. Google Earth has the highest resolution (0.25m/pixel), while the others have a resolution of 0.5m/pixel. However, not all maps provide information on the shooting time of the satellite images. For this study, the construction progress of a super high-rise building under construction in the Huangpu district was selected as a reference, with its resolution and timeliness being listed in Table 1, where "1" represents the newest and "4" represents the oldest.

Table 1 Satellite Images Table

Source	Highest Resolution	Timeliness
Amap	0.5m/pixel	2
Tianditu	0.5m/pixel	1
Google Earth	0.25m/pixel	4
Baidu	0.5m/pixel	2

After comprehensive consideration, this study has opted for satellite imagery from Tianditu as the reference satellite imagery, with a clear timeline in the fourth quarter of 2022.

Building shapefile

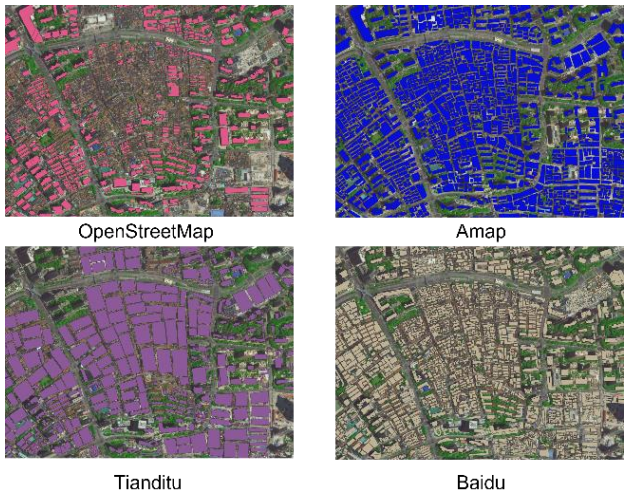

Figure 3 Shapefiles in different sources

https://doi.org/10.26868/25222708.2023.1355

The shapefile is a commonly used geospatial data format, and the Building shapefile contains various data such as the geometric positions and attribute information of buildings, which can be read and edited by GIS software. In this paper, we use the shapefile format to store the geometric information of buildings for subsequent analysis and modeling. The acquisition of the base shapefile can be done through computer algorithms and satellite images, or by downloading building data from some literature sources(Zhang et al., 2022). The most convenient way to download is through OSM, while other building outline data is obtained from partially open and downloadable websites.

Street view

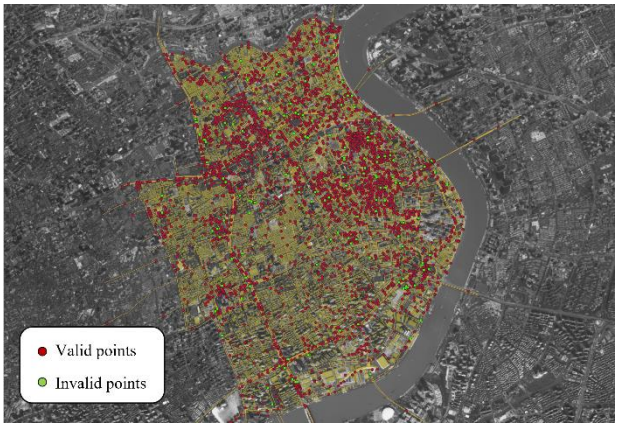

Figure 4 Sampling points for street view maps

In street view images, Google Street View is almost unavailable in Huangpu District. Commercial permission is required for using Amap and Baidu Map, but they cover almost all road networks. In this study, we used ArcGIS to construct sightlines to sample the road network of Huangpu District. To avoid obstructed sightlines, we sampled at every turn and every 200 meters. We also constructed a preview request for the GIS coordinates of each sampling point. If a preview image was returned, it indicated that the area had street-view images. A total of 3670 sampling points were obtained, of which 183 were invalid points.

Pois

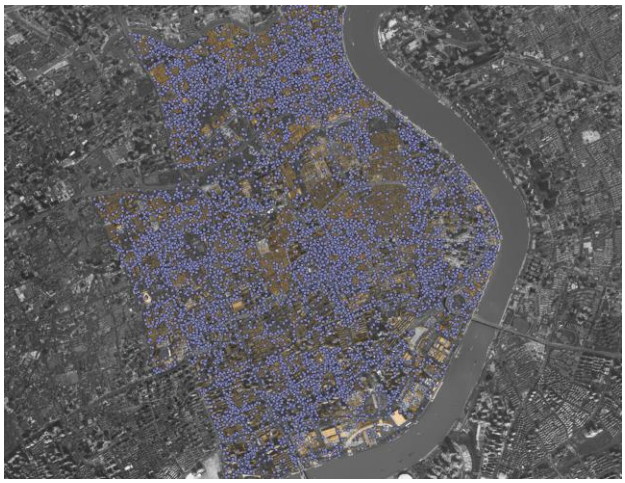

Figure 5 Pois in Huangpu

A point of interest (POI) is a specific location or landmark that is of interest or importance to people. POIs can be anything from historical landmarks, tourist attractions, restaurants, shops, or any geographic point that someone may want to find or explore. In the context of maps or navigation systems, POIs are commonly used to help users find the locations they are interested in. POIs can be obtained either through the free API from Amap or by downloading from Tianditu and Guihuayun. In this work, these types of POIs were collected, and after deduplication, 17032 pois were retained.

Aois

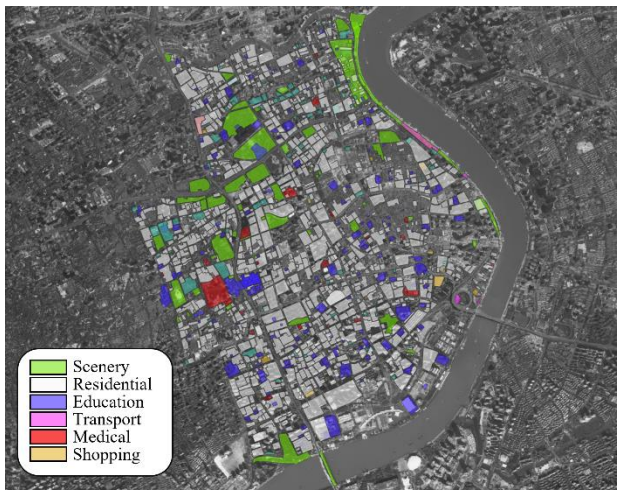

Figure 6 AOI category

The method used to obtain the AOIs was referenced from (Deng et al., 2021). Firstly, we obtained the names of residential communities from real estate websites and then acquired the names of education, scenic areas, and hospitals from government websites. Afterward, we constructed a request to obtain the AOI information. A total of 1,209 AOIs were collected, including 62 for scenery, 13 for shopping, 5 for transport, 163 for education, 998 for residential, and 29 for medical facilities. The majority of residential AOIs were sourced from real estate websites.

Assessment standards for building information usability

The information that can be obtained or inferred based on these sources primarily includes the building's GIS information, footprint, construction age, height, category, window-to-wall ratio, and roof type. These pieces of information and sources of data may not always correspond one-to-one and may sometimes be one-to-many or many-to-one. For instance, the building category may be determined as a shopping mall because it contains multiple commercial POIs, or as a residential dwelling because it falls within a residential AOI. Taking this into account, this paper determines whether building information can be determined using certain open data according to the following rules:

Building footprint: Building footprint refers to the area occupied by the bottom of the building, including the bottom shape and GIS information. The following two types of information can provide building footprint:

https://doi.org/10.26868/25222708.2023.1355

- Shapefiles containing building information.
- High-resolution satellite images(0.5m/pixel) that cover the building area.

Year built: The acquisition of year built is essential for establishing standards for HVAC systems in buildings. According to the Energy Conservation Standards for Buildings (GB 50189-2005) and GB 50189-2015, the construction age of a building can be divided into three categories: pre-2005, 2006-2015, and post-2015. The following two sources of information can provide the year built of a building:

- Real estate information containing the year of construction or renovation.
- Historical satellite images covering the period from 2004-2016.

Building Height: The acquisition of building height can be challenging as it is difficult to obtain directly. Although previous studies have used satellite imagery to identify building shadows to estimate building heights, this method is limited by its dependence on the time of image acquisition and may require manual adjustments for different areas. Additionally, some very high-resolution satellites (e.g., Sentinel-2) may provide data about building heights, but the accuracy of such data is often low. In China, the accuracy of such datasets is limited to only 10 meters(Wu et al., 2023), which renders it unusable in the context of UBEM. Therefore, this section considers the following three sources of information:

- Real estate information containing information on building height.
- Shapefiles containing information on building height.
- Buildings that are located within the coverage area of street view maps.

Building Type: The determination of building type mainly depends on two factors: the POIs contained within its footprint and the AOIs that contain it. These pieces of information can be used to infer the building type, as described in detail in Deng et al.'s work [1]. The following three types of information are required, where 1st is necessary, and 2nd and 3rd are optional:

- Building footprint (necessary)
- Three or more POIs are contained within the footprint.
- AOIs that contain the building.

Window-to-Wall Ratio: As of now, the only known way to obtain the window-to-wall ratio information through open source is by using street view maps. To obtain this information, the building must be within the coverage area of street view maps, and only the building closest to the street view road is considered to avoid obstruction. The following information is required:

- The building is located within the coverage area of street view maps.

Building Envelope: Existing research mainly acquires parameters for building envelopes based on the age of the building and the partially mandatory regulations implemented during its construction. Also, facade color could affect the performance of the envelope, thus, the street view could help.

- Year built information
- The building is located within the coverage area of street view maps.

Evaluation of Open Data Quality

The quality evaluation of open data is very important. In this article, we will evaluate the obtained open data through the following two aspects:

Data source: The reliability and credibility of the data source mainly need to consider factors such as the credibility of the data provider, the frequency of data updates, and the type of data source.

GIS information quality: Environmental Systems Research Institute （ESRI） believes that it mainly includes the following types(*Identify Data Quality Requirements—ArcGIS Pro | Documentation*, n.d.):

- Completeness
- Logical consistency
- Spatial accuracy
- Thematic accuracy
- Temporal quality
- Data usability

Where, completeness refers to the truth value of GIS data, which means whether the buildings in the GIS data correspond to the actual existing buildings. Logical consistency refers to whether the elements in the GIS data are in a reasonable area, for example, if a building is in the middle of the road, it indicates that there is a problem with the consistency of the data. Spatial accuracy refers to whether the position of the buildings in the GIS data deviates from their actual position on the earth. Thematic accuracy refers to whether the information in the feature is accurate, such as whether a residential building is set as a commercial building. Temporal quality refers to the timeliness of GIS data. Data usability refers to whether GIS data is used for its correct purpose, such as whether the data of a certain community is used to evaluate the city.

Since the UBEM dataset in this article is mainly based on the data provided by map service providers and government data, the data of map service providers should be the latest, because many people use it every day, and if there is any error, there will be feedback and correction at any time. Secondly, it is the government's data. The open data provided by the government may not be as good in timeliness, but it is relatively robust in data reliability. The timeliness of satellite maps, the effectiveness of shapefiles, and the selection process have been mentioned in the previous section. The dataset constructed based on such data should be quite reliable.

Results

According to the above method, a total of 14,803 footprints were collected and manually calibrated, of which 13,613 were obtained from publicly available shapefiles, accounting for 92%, and approximately 8% were manually calibrated buildings. As shown in Figure 7, 9,480 buildings were covered by street view,

https://doi.org/10.26868/25222708.2023.1355

accounting for 64%, 8,392 buildings were covered by aoi, accounting for 56.7%, and 6,167 buildings were covered by poi, accounting for 41.7%.

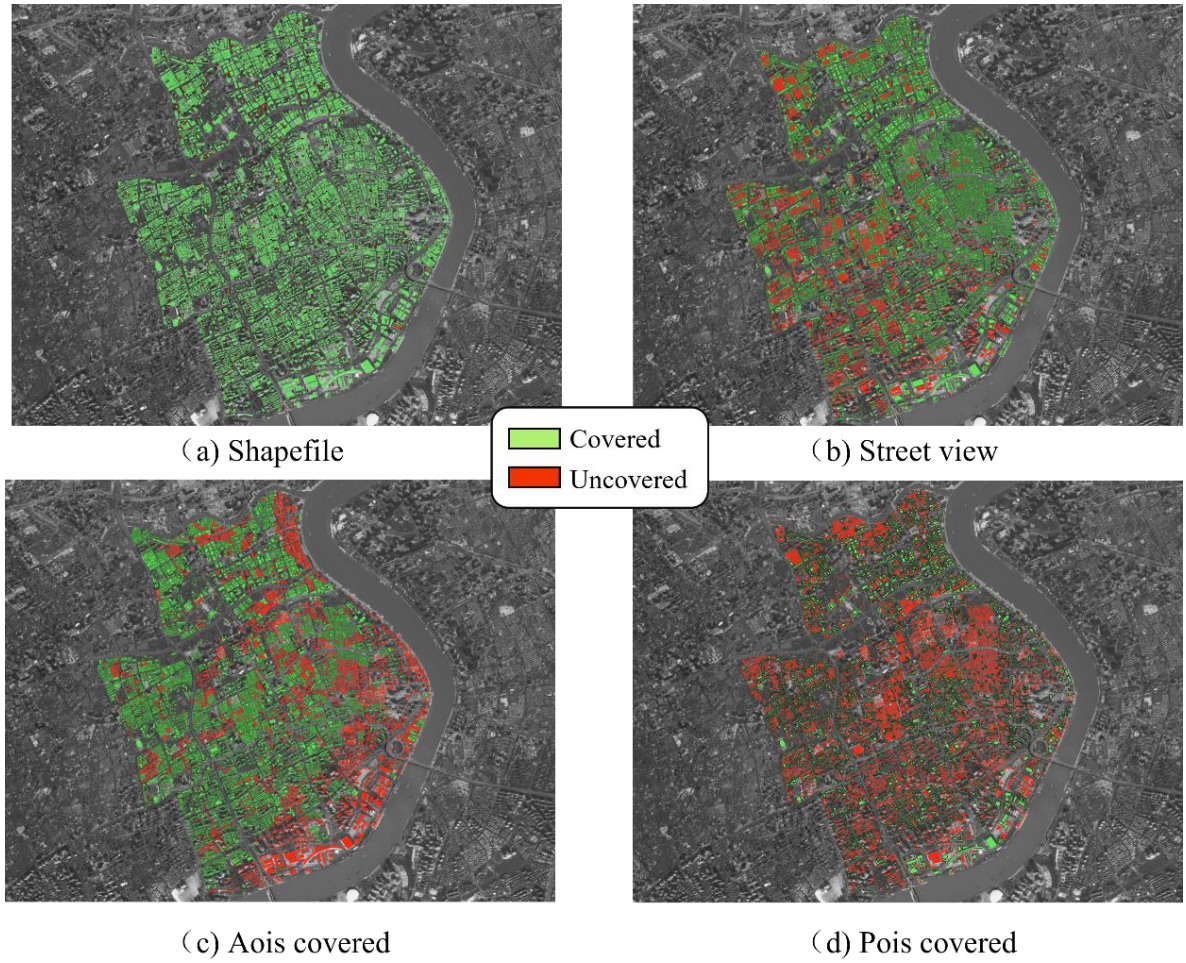

Figure 7 The coverage of the raw open data

After the processing described above, the year built data of 100% of the buildings can be supplemented, the height data of 99.5% of the buildings can be obtained, the category information of 71.5% of the buildings can be supplemented, and the WWRs information of 69.7% of the buildings can be supplemented.

As shown in Figure 8, it is evident that the acquired open-source data exhibits relatively low coverage. However, with proper integration and processing, the coverage of all UBEM data can exceed 70%.

For further discussion, an upset plot was plotted as shown in Figure 9 to display the combination of different types of information. The plot is divided into three parts: the bar chart in the upper right corner represents each grouping, and the image in the lower left corner represents the number of buildings with a certain attribute. The intersection of the two bar charts represents the categories in the upper right corner grouping that have different numbers of attributes, and The information within the red dashed box is the information needed by UBEM.. For example, there are 2,969 buildings that contain all the information and only 41 buildings that contain only the "Pois covered" and "Year built" information.

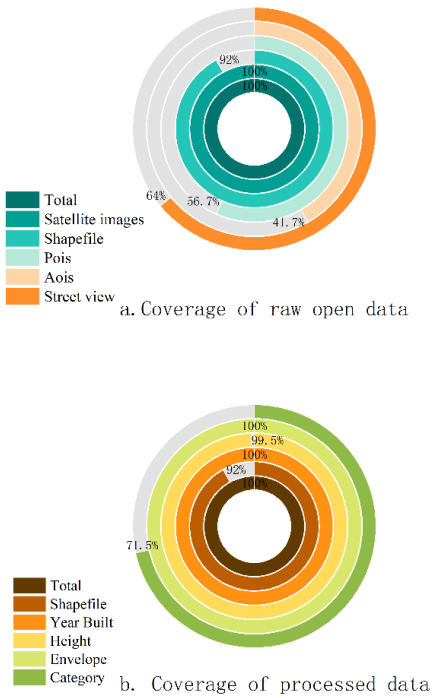

Figure 8 The coverage of raw data and processed data.

https://doi.org/10.26868/25222708.2023.1355

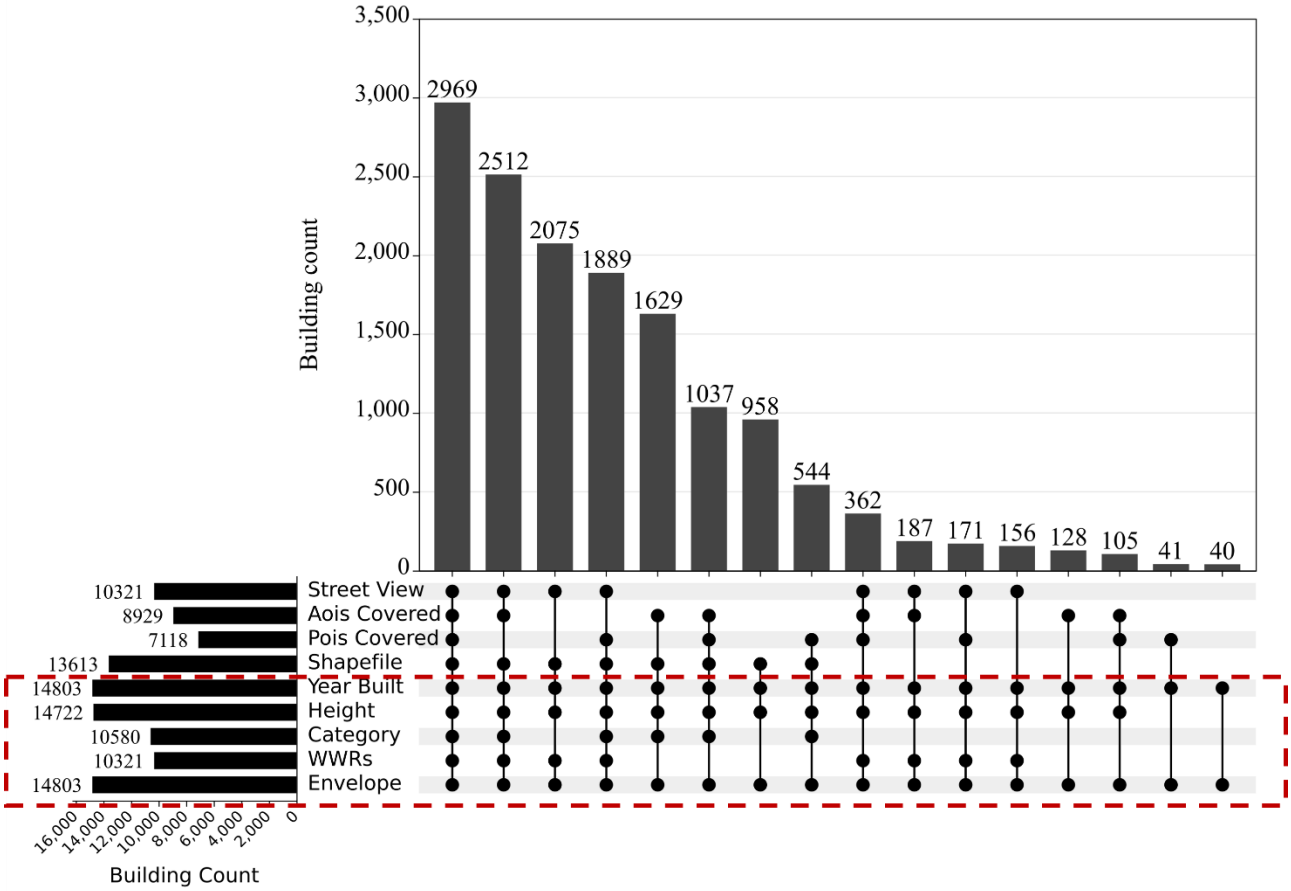

Figure 9 The upset plot of the processed dataset

Conclusion and discussion

This paper proposes a comprehensive evaluation framework to assess the potential availability of open data for urban building energy modeling. A case study on buildings in Huangpu District, Shanghai, is conducted to demonstrate the application of the framework. The results show that open-source data can provide a considerable amount of building information useful for UBEM., with proper integration and processing, the coverage of all UBEM data can exceed 70%, and the effectiveness and coverage of different types of information is considerable. However, it is challenging to obtain comprehensive data for all buildings.

However, this article still has the following shortcomings that need to be further improved: firstly, it only explores the potential of constructing a UBEM dataset using open-source data and presents a preliminary dataset, without releasing a complete and improved UBEM dataset. Secondly, the quality of open-source data should be emphasized and discussed in detail. However, due to space constraints, this article only briefly analyzes two directions. Thirdly, the article only discusses the potential of constructing a UBEM dataset using open data, without actually modeling the UBEM for the city. Therefore, we will address these issues in future research.

Additionally, it should be noted that although open-source data provides valuable information for urban building energy modeling, it is becoming increasingly limited for several reasons. Firstly, data quality may be affected by inaccuracies or errors since the data is sourced from different organizations or individuals. Secondly, open-source data may be slow to update, with delays ranging from months to years, which lags behind the pace of urban development. Thirdly, due to the large scale and high density of urban buildings, it is difficult to obtain complete information for all buildings from open-source data. Finally, map service providers are gradually limiting the number and scope of access to their free APIs, making it more difficult to access the latest open data.

Thus, it is essential to explore new data sources and develop more efficient data processing methods to better utilize open-source data for energy modeling.

Acknowledgment

This paper is supported by the National Natural Science Foundation of China (NSFC) through Grant (No. 51908204), and the Natural Science Foundation of Hunan Province of China through Grant (No. 2020JJ3008).

References

Amap. (n.d.). Retrieved 23 March 2023, from https://www.amap.com/

Baidu Map. (n.d.). Retrieved 23 March 2023, from https://map.baidu.com/@12577013,3254710,13z

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

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., & 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

Google Earth. (n.d.). Retrieved 23 March 2023, from https://earth.google.com/web/

Huang, Z., Mendis, T., & Xu, S. (2019). Urban solar utilization potential mapping via deep learning technology: A case study of Wuhan, China. *APPLIED ENERGY*, *250*, 283–291. https://doi.org/10.1016/j.apenergy.2019.04.113

Identify data quality requirements—ArcGIS Pro | Documentation. (n.d.). Retrieved 4 June 2023, from https://pro.arcgis.com/en/pro-app/latest/help/data/validating-data/identify-data-quality-requirements.htm

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*, 112911. https://doi.org/10.1016/j.enbuild.2023.112911

Lee, S., Iyengar, S., Feng, M., Shenoy, P., & Maji, S. (2019). DeepRoof: A Data-driven Approach For Solar Potential Estimation Using Rooftop Imagery. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2105–2113. https://doi.org/10.1145/3292500.3330741

Liu, Z., Zhou, X., Tian, W., Liu, X., & Yan, D. (2022). Impacts of uncertainty in building envelope thermal transmittance on heating/cooling demand in the urban context. *Energy and Buildings*, *273*, 112363. https://doi.org/10.1016/j.enbuild.2022.112363

OpenStreetMap. (n.d.). Retrieved 23 March 2023, from https://www.openstreetmap.org/

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

STATISTICAL COMMUNIQUÉ OF THE PEOPLE'S REPUBLIC OF CHINA ON THE 2022 NATIONAL ECONOMIC AND SOCIAL DEVELOPMENT. (n.d.). Retrieved 23 March 2023, from http://www.stats.gov.cn/english/PressRelease/202302/t20230227_1918979.html

Szcześniak, J. T., Ang, Y. Q., Letellier-Duchesne, S., & Reinhart, C. F. (2022). A method for using street view imagery to auto-extract window-to-wall ratios and its relevance for urban-level daylighting and energy simulations. *Building and Environment*, *207*. Scopus. https://doi.org/10.1016/j.buildenv.2021.108108

Wang, C. (2022). Data acquisition for urban building energy modeling: A review. *Building and Environment*.

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

Wang, C., Wei, S., Du, S., Zhuang, D., Li, Y., Shi, X., Jin, X., & Zhou, X. (2021). A systematic method to develop three dimensional geometry models of buildings for urban building energy modeling. *SUSTAINABLE CITIES AND SOCIETY*, *71*. https://doi.org/10.1016/j.scs.2021.102998

Wu, W.-B., Ma, J., Banzhaf, E., Meadows, M. E., Yu, Z.-W., Guo, F.-X., Sengupta, D., Cai, X.-X., & Zhao, B. (2023). A first Chinese building height estimate at 10 m resolution (CNBH-10 m) using multi-source earth observations and machine learning. *Remote Sensing of Environment*, *291*, 113578. https://doi.org/10.1016/j.rse.2023.113578

Yang, H., Yuan, J., Lunga, D., Laverdiere, M., Rose, A., & Bhaduri, B. (2018). Building Extraction at Scale Using Convolutional Neural Network: Mapping of the United States. *IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING*, *11*(8), 2600–2614. https://doi.org/10.1109/JSTARS.2018.2835377

Zhang, Z., Qian, Z., Zhong, T., Chen, M., Zhang, K., Yang, Y., Zhu, R., Zhang, F., Zhang, H., Zhou, F., Yu, J., Zhang, B., Lü, G., & Yan, J. (2022). Vectorized rooftop area data for 90 cities in China. *Scientific Data*, *9*(1), 66. https://doi.org/10.1038/s41597-022-01168-x