Geo-spatial Information Science

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tgsi20

Large-scale urban building function mapping by integrating multi-source web-based geospatial data

Wei Chen, Yuyu Zhou, Eleanor C. Stokes & Xuesong Zhang

To cite this article: Wei Chen, Yuyu Zhou, Eleanor C. Stokes & Xuesong Zhang (31 Oct 2023): Large-scale urban building function mapping by integrating multi-source web-based geospatial data, Geo-spatial Information Science, DOI: 10.1080/10095020.2023.2264342

To link to this article: https://doi.org/10.1080/10095020.2023.2264342

GEO-SPATIAL INFORMATION SCIENCE
https://doi.org/10.1080/10095020.2023.2264342

Large-scale urban building function mapping by integrating multi-source web-based geospatial data

Wei Chen [a], Yuyu Zhou [b,c], Eleanor C. Stokes [d] and Xuesong Zhang [e]

[a]Department of Geological and Atmospheric Sciences, Iowa State University, Ames, USA; [b]Department of Geography, The University of Hong Kong, Hong Kong, China; [c]Urban Systems Institute, The University of Hong Kong, Hong Kong, China; [d]NASA Headquarters, Washington, USA; [e]USDA-ARS Hydrology and Remote Sensing Laboratory, Beltsville, USA

ABSTRACT

Morphological (e.g. shape, size, and height) and function (e.g. working, living, and shopping) information of buildings is highly needed for urban planning and management as well as other applications such as city-scale building energy use modeling. Due to the limited availability of socio-economic geospatial data, it is more challenging to map building functions than building morphological information, especially over large areas. In this study, we proposed an integrated framework to map building functions in 50 U.S. cities by integrating multi-source web-based geospatial data. First, a web crawler was developed to extract Points of Interest (POIs) from Tripadvisor.com, and a map crawler was developed to extract POIs and land use parcels from Google Maps. Second, an unsupervised machine learning algorithm named OneClassSVM was used to identify residential buildings based on landscape features derived from Microsoft building footprints. Third, the type ratio of POIs and the area ratio of land use parcels were used to identify six non-residential functions (i.e. hospital, hotel, school, shop, restaurant, and office). The accuracy assessment indicates that the proposed framework performed well, with an average overall accuracy of 94% and a kappa coefficient of 0.63. With the worldwide coverage of Google Maps and Tripadvisor.com, the proposed framework is transferable to other cities over the world. The data products generated from this study are of great use for quantitative city-scale urban studies, such as building energy use modeling at the single building level over large areas.

ARTICLE HISTORY
Received 29 December 2022
Accepted 25 September 2023

KEYWORDS
Building functions; geospatial data; TripAdvisor; Google Static Maps

1. Introduction

Buildings, the main venues of urban social and economic activities, are the most important component and the finest measurement unit for urban studies (Hu, You, and Neumann 2003). Information on the morphology (e.g. shape, size, and height) and function (e.g. working, living, and shopping) of buildings is highly needed for urban and regional planning as well as other quantitative urban studies (Kunze and Hecht 2015). Building functions are shaped by the actual use of inhabitants, and therefore they can be viewed as a bridge, enabling researchers to monitor socioeconomic space. For example, with the help of data on building functions, we can investigate city-wide air quality (Xu et al. 2021), public health (Kousa et al. 2002), energy consumption (Davila, Reinhart, and Bemis 2016; Li et al. 2018), energy heat emission (Luo et al. 2020), and disaster loss (Yeh, Loh, and Tsai 2006).

Considering the importance of building information on urban studies, many efforts have been devoted to mapping building footprints, heights, and functions. One way to capture building morphological information is to use physical features (e.g. spectrum, texture, and geometry) extracted from remote sensing data, including very high-resolution satellite imagery, aerial imagery, and Laser-scanner data (LiDAR). With large coverage and rich physical features, remote sensing data can be used to extract building morphological information over large-scale areas. For instance, Microsoft Maps team (Anon 2018) produced high-quality building footprint data in the United States, Canada, Uganda, and Tanzania by applying deep learning algorithms into Bing imagery. Li et al. (2020) estimated building height in 500 m resolution over the United States using Sentinel-1 synthetic aperture radar data. However, remote sensing data lack information about the economic and social function of buildings, limiting its ability to distinguish between buildings with similar spatial forms, such as, hotel and shopping mall.

To overcome the aforementioned limitations, social sensing data (e.g. social media check-in records, taxi trajectories, mobile phone calls, transit smart card records) have been widely used to differentiate buildings with different socio-economic functions (Arunplod et al.

CONTACT Yuyu Zhou yuyuzhou@hku.hk
Supplemental data for this article can be accessed online at https://doi.org/10.1080/10095020.2023.2264342.

2017; Liu et al. 2018; Niu et al. 2017; Zhuo et al. 2019). The information of building function can be derived by integrating multi-source social sensing data created by individuals (Liu et al. 2015). For example, Niu et al. (2017) integrated social media users' real-time location records and taxi trajectory data to identity buildings with shopping, hotel, office, hospital and residential functions. However, the high cost of collecting social sensing data limits its potential to extract building functions over large areas. Furthermore, the utilization of the shared local governmental database, such as the accessor's parcel data, can provide valuable information of building functions (Chen et al. 2020). However, in many regions, especially in countries in the global south, access to such database is severely limited or not available at all. Even in regions where a shared local government database exists, collecting this data proves to be a time-consuming and labor-intensive process, as it involves accessing disparate governmental databases specific to each city. Consequently, there is an urgent need to develop a transferable and integrated framework capable of effectively addressing the lack of building function data over large areas, while accommodating the data availability challenges encountered in countries of the global south.

In this study, TripAdvisor.com and Google Maps were chosen as two web platforms that can be used to map building functions over a large area based on their advantages. First, both of them contain the location and socio-economic function of buildings. TripAdvisor.com is an online travel company that contains the address of hotels, shops, and restaurants. Many studies have utilized web crawler tools to analyze users' reviews and visualized geographical locations of restaurants and hotels to improve their business' service and quality of products (Chang, Ku, and Chen 2019). Google Maps is a web mapping platform that provides locations of offices, hospitals, schools, etc. It offers a series of Application Programming Interfaces (APIs), allowing users to utilize Google Maps services and to conduct place information queries. Second, both of them operate across many countries and update timely. TripAdvisor.com operates in 49 countries and has 463 million average monthly unique visitors. Google Maps is used by over 1 billion people every month in 104 countries around the world.

Based on those two web platforms, an integrated framework was proposed to map building functions over 50 cities in the United States. First, we proposed two workflows to extract geospatial data including Points of Interest (POIs), roads, and land use parcels from Google Maps and TripAdvisor.com. Second, we identified residential buildings using an unsupervised machine learning algorithm. Third, we identified six non-residential functions using type ratio of POIs and area ratio of land use parcels. The remainder of this paper describes the study area, dataset used in this study (Section 2), the proposed workflow to collect geospatial data (Section 3.1), workflow to identify building functions (Section 3.2), and results (Section 4). After description, there is a discussion (Section 5) of results and conclusion (Section 6).

2. Study area and datasets

2.1. Study area

We selected 50 U.S. cities (Figure 1) to test the scalability of the proposed framework. Fifty U.S. cities were chosen based on two reasons. First, the selected cities had different sizes, from small (Decatur, Georgia), middle (Des Moines, IA) to large (Boston, MA). Second, the selected cities were at risk of being hit by natural disasters according to the frequency of weather hazards archived in the Storm Events Database from National Oceanic and Atmospheric Administration (NOAA) (NOAA 2021). The resultant building function maps in those cities are important in disaster management, for example, estimation of disaster loss and vulnerability to disaster.

2.2. Data

Main data used in this study include city boundary, building footprints, and Sentinel-1 building height datasets. We downloaded 50 city boundaries from OpenStreetMap (OSM) and extracted building footprints and building height within these city boundaries. OSM is a volunteer geo-information project founded in 2004. Administrative boundaries in OSM were delineated by volunteers with reference data from state or county GIS websites. Building footprints were downloaded from Microsoft Maps (Anon 2018), a country wide open building footprints dataset that provide the location and geometry of individual buildings across all 50 states of the United States. We employed these data as the fundamental mapping units within our framework to generate the resultant building function map. Furthermore, we leveraged these data to develop a series of metrics that effectively capture the shape, size, and uniformity of housing within each individual parcel. Building height was estimated at 500 m resolutions from Sentinel-1 data in 2015 using a method proposed by Li et al. (2020) and the global resultant data can be download from https://figshare.com/s/7f2b254ed18fac8eb7a0 (Zhou et al. 2022). First, a building height model was developed using the reference height from LiDAR and dual-polarization information (i.e. VV [copolarization] and VH [cross-polarization]) at 500 m resolution. Second, three parameters in building height model were calibrated through a cross-validation. The estimated building heights exhibits excellent performance in the United States, as indicated by a low Root Mean

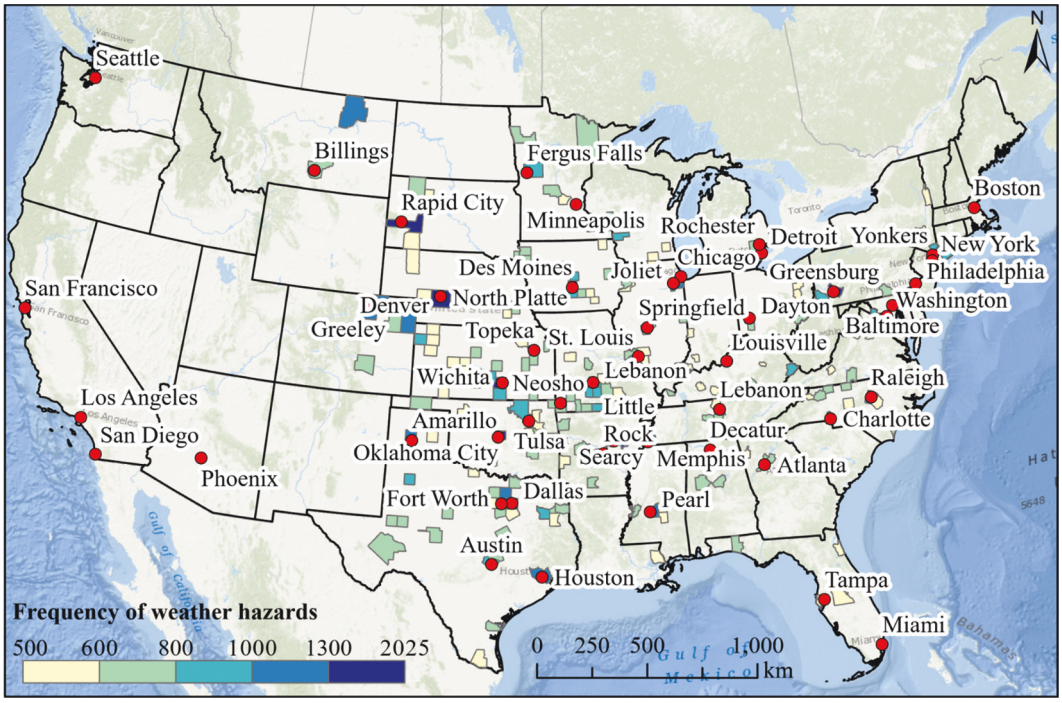

Figure 1. The selected 50 study cities in the United States.

Square Error (RMSE) of less than 0.50 m between the estimated urban built-up heights and the reference.

3. Methodology

We developed a framework to identify urban building functions including residence, office, school, shop, hotel, restaurant, and hospital (Figure 2). It includes two workflows to collect geospatial datasets and one workflow to identify building functions. More details are presented in the following sections.

3.1. Collection of web-based geospatial data

3.1.1. Web crawler

We designed a web crawler to automatically collect addresses of hotels, restaurants, and shops from web contents of TripAdvisor.com for each city and to convert these addresses to POIs using the geocoding technique. POIs are points with longitude-latitude coordinate and specific building functions. Three web crawlers are included in Figure 3. First, we designed a web crawler to collect 50 U.S. cities' Uniform Resource Locators (URLs) (e.g. https://www.tripadvisor.com/Hotels-g32655-Los_Angeles_California-Hot

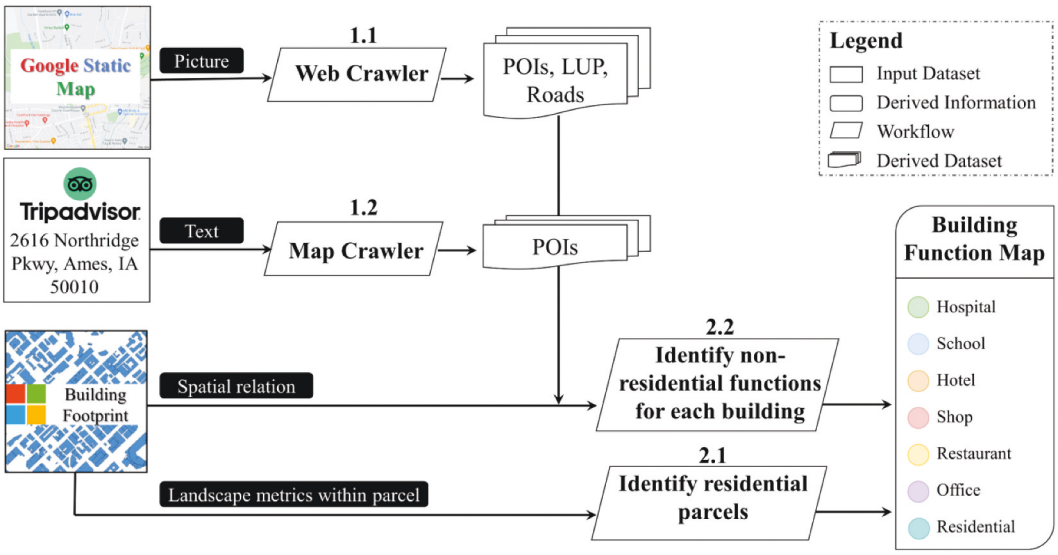

Figure 2. The overall framework of this study.

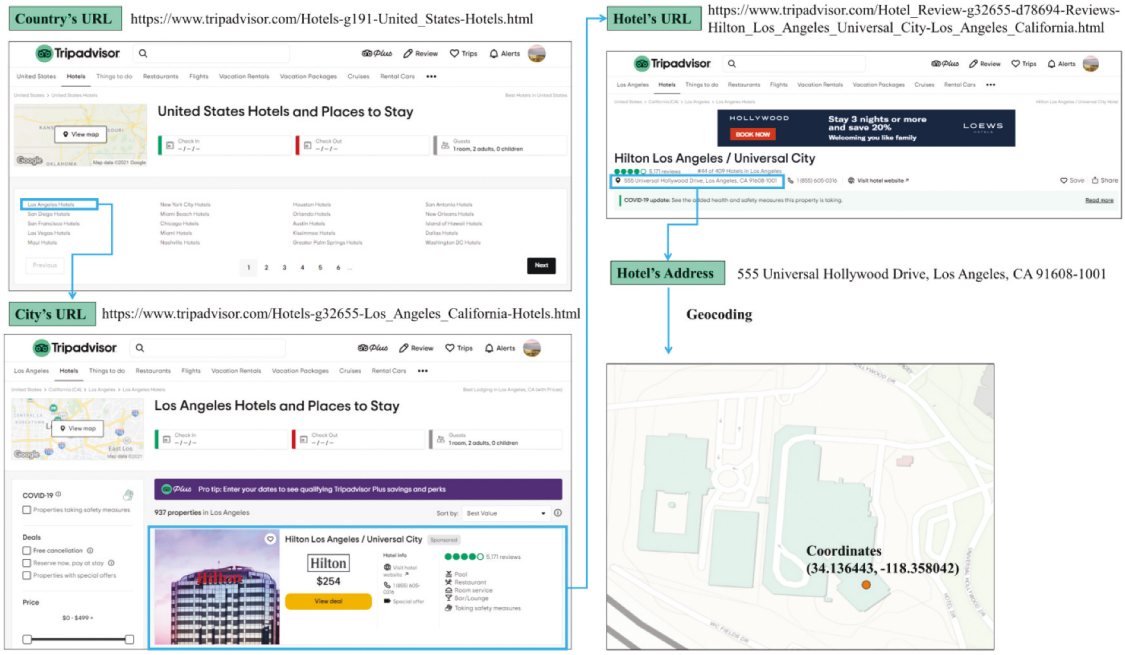

Figure 3. An example of collecting a POI with the function of hotel using web crawlers.

els.html) from U.S. country's URLs (e.g. https://www.tripadvisor.com/Hotels-g191-United_States-Hotels.html). Second, we designed a web crawler to collect all hotels' URLs (e.g. https://www.tripadvisor.com/Hotel_Review-g32655-d78682-Reviews-The_Garland-Los_Angeles_California.html) from each city's URL. Third, we designed a web crawler to collect hotel addresses from each hotel's URL. Finally, the geocoding technique was used to convert 9076 hotels, 107,935 restaurants, 3827 shop addresses to POIs with corresponding building functions.

3.1.2. Map crawler

We developed a map crawler to automatically collect geospatial data including roads, land use parcels, and POIs using Google Maps Static APIs (Figure 4). The Maps Static APIs service can return Google static maps as an RGB image according to the defined zoom level, map style, image size, and coordinates of the central point. The map crawler includes four key components. First, a fishnet covering the whole city was generated using ArcPy provided by ArcGIS Pro 2.7. One fishnet grid represents one RGB image

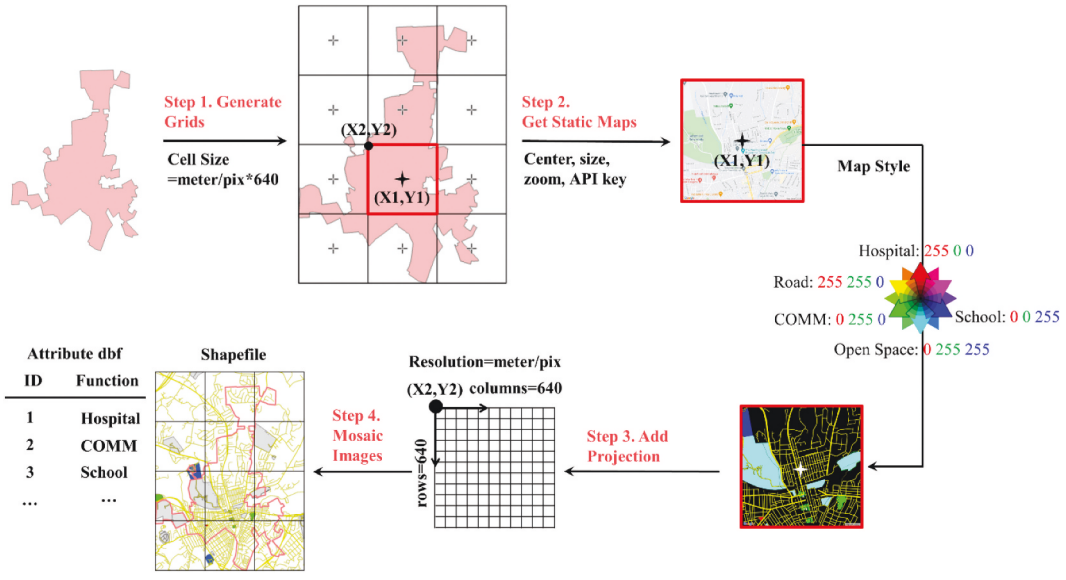

Figure 4. An example of collecting land use parcels using map crawler. Meters/pixel for roads and land use parcels is 4.756 and for POIs is 0.597.

requested from the Google Maps Static API. Therefore, the cell size of fishnet is meters/pixel multiplied with the image size (640 × 640). Second, a static map of each grid was obtained using Google Maps Static API with coordinates of the central point of each grid, zoom level (15 for extraction of roads and land use parcels and 18 for extraction of POIs), and the customized map style. The map style customization options can be accessed through the "Map Styles" tab on the Google Maps Platform. In our map style, we assigned distinct colors to various types of POIs, such as hospitals, offices, schools, hotels, shops, and restaurants, as well as different land use parcel types, including commercial corridors, open spaces, hospitals, and schools. The static map at zoom level of 18, incorporating the customized map style, can be found in Figure S1(b). Third, we added the projection of for each RGB image by assigning upper-left corner coordinates of RGB image as upper-left corner coordinates of the corresponding fishnet grids as well as assigning resolution of RGB image as meter/pixel (a projected grid is depicted in Figure S1(c)). Fourth, the projected images were mosaicked and converted to polygons or points as land use parcels and POIs with specific socioeconomic functions (the resulting POIs are displayed in Figure S1(d)). The mosaic and conversion processes were effectively processed on high-performance computing cluster with sufficient processing power and memory capacity. Additionally, we employed parallel computing techniques to optimize the computational efficiency and reduce processing time.

3.2. Identification of building functions

We developed a workflow (Figure 5) to identify functions of building footprints by fusing the collected web-based geospatial data. It includes two key steps. First, we identified residential parcels based on building footprint-derived landscape features and assigned buildings within the identified parcels as residential buildings. Second, we identified functions of non-residential buildings by fusing the collected POIs and land use parcels.

3.2.1. Identification of residential building functions

We first calculated a set of building footprint-derived landscape metrics (Table 1) for each parcel. These parcels have relatively homogeneous socioeconomic functions (W. Chen et al. 2022; Liu and Long 2016; Yuan, Zheng, and Xie 2012; Zhang et al. 2017) and can be segmented by roads collected by the map crawler.

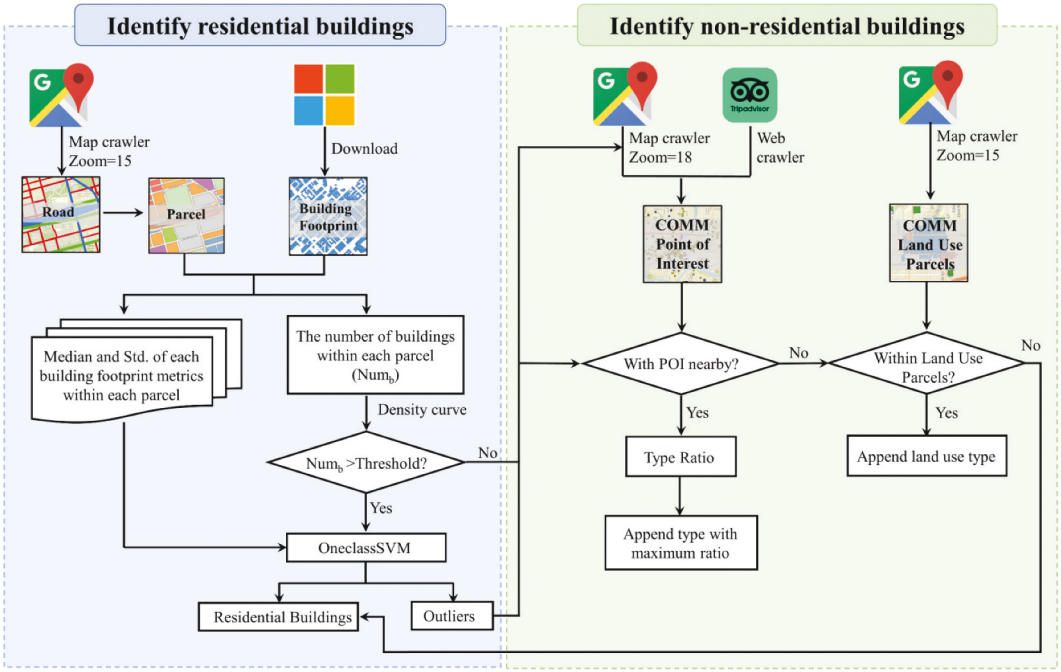

Figure 5. The workflow to identify function of residential and non-residential buildings.

Table 1. Metrics of building footprints.

Metrics	Definition
Area (A)	The areal extent of a building polygon
Perimeter (P)	The distance around a building polygon
Length (l)	The length (l) and width (W) of the minimum bounding rectangle enclosing the building polygon
Width (W)	
Elongatedness (ELG)	*ELG=l/W*
Height	500 m building height raster

Parcel-based summary statistics were generated by calculating a measure of central tendency (median) and variability (standard deviation) for each of the metrics. The former allows us to measure characteristics of the typical residential parcel, while the latter allows us to measure the uniformity of residential parcels. To ensure that the machine learning algorithm applied equal weight to each metric, we normalized each metric with a mean of zero and one standard deviation (Durst et al. 2021).

We then applied the OneClassSVM algorithm to parcels with the number of buildings in the parcel (Num_b) larger than the threshold and classified them into residential or non-residential parcels. OneClassSVM, an unsupervised outlier detection algorithm, was chosen to classify residential and non-residential (outliers) parcels because of the unequal distribution of binary classes (80% of residential vs. 20% of non-residential). The threshold of Num_b was used to obtain higher classification accuracy because we found that there were distinct differences between binary classes in the parcel larger than a threshold. For example, when the threshold is 24 (Figure 6(b)), the difference between binary classes was more distinct compared to the threshold is 1 (Figure 6(a)).

The optimal threshold of Num_b can be determined for different cities by following a three-step process. In Figure 6(c), the upper, middle, and bottom portions respectively provide examples for each step, allowing us to gain insights into the determination of the optimal threshold. First, the density curve of Num_b was generated using probability density function (Equation (1)) for Num_b in each city dataset. Second, the cumulative density curve (Equation (2)) of Num_b was obtained by integrating the probability density function. Third, as shown in the bottom portion of Figure 6(c), a consistently error rate below 10% was achieved for the four cities we examined when the value of Num_b was approximately at the turning point of cumulative density curve. This error rate is calculated as the ratio between the number of non-residential buildings in residential parcels and the total number of non-residential buildings. Therefore, the turning point of cumulative density curve was determined as the threshold of Num_b. Using a Cartesian plane, we determined the turning point on the cumulative density curve by calculating the shortest distance between the curve's points and the line connecting the cumulative probability for minimum and maximum values of Num_b.

$$P(a < x < b) = \int_a^b f(x)dx \tag{1}$$

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt, \text{for all } x \in \mathbb{R} \tag{2}$$

Let x be the continuous random variable with the probability density function f. The probability is calculated by finding the area under its curve and the X-axis within the lower limit (a) and upper limit (b). The cumulative distribution function F is found by integrating the f.

3.2.2. Identification of non-residential building functions

We utilized area percentage and type ratio to identify non-residential building functions. First, the building footprint layer was intersected with the land use parcel layer, and the intersected area percentage was calculated. If the intersected area percentage was larger than 50%, building footprints were assigned with land use parcel type. Second, the building footprint layer was intersected with the POIs layer and the intersected POIs were used to calculate type ratio (Equation (3)) (Chen et al. 2020) to determine the POI type that can be appended to buildings. This ratio was calculated as

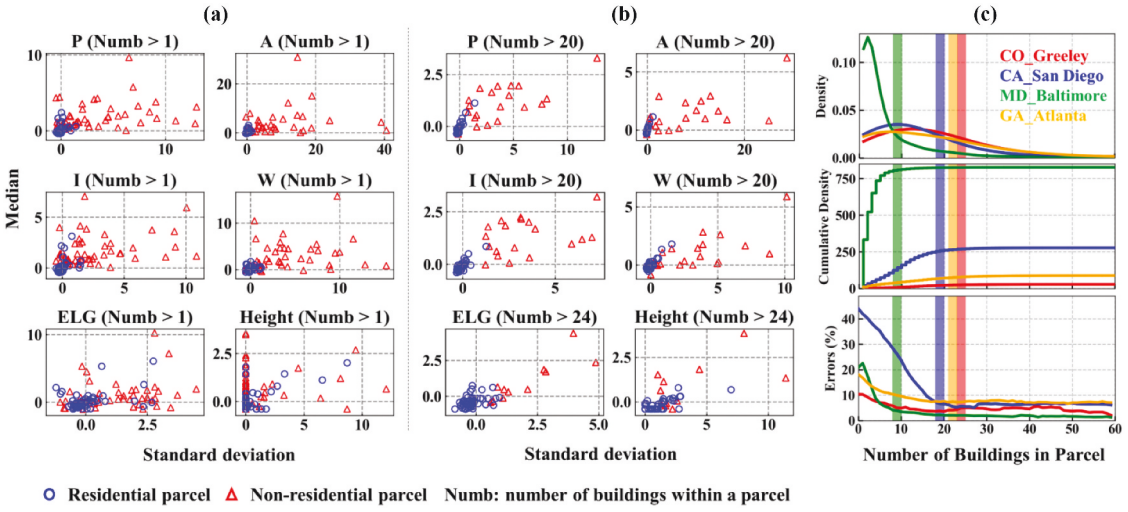

Figure 6. Normalized parcel-based building footprint metrics in Greeley, CO when Num_b larger than 1 (a) and 20 (b). The determined Num_bthreshold of four example cities based on their different density curves (c).

the percentage of POI types among the total number of POIs in the building's buffer region. The radius of the building's buffer is the half of the Euclidean distance between this building and its nearest building because a POI is not likely to be shared by two buildings. Type ratio was calculated for POIs with the major functions (i.e. hospital, school, and hotel) in each building buffer. The function with the maximum type ratio was appended to building footprints.

$$TR_i = \frac{n_i}{N_j} \times 100\% \quad (3)$$

where TR_i is the type ratio of POI function i; n_i is the number of POIs with function i; and N_j is the total number of POIs in the building's buffer region j.

3.3. Accuracy assessment

We collected POIs and land use parcels from OSM in 50 U.S. cities to evaluate results of the geospatial data collection from the web-based platform. We generated 500 m × 500 m grids to calculate POIs density in each city for data from OSM and Google Maps and produced building function density map (i.e. total floor areas in 500 m grid) for four socioeconomic functions (i.e. commercial, office, institutional, and residential types). We calculated the area of land use parcels within the city boundary for school, hospital, commercial corridors, and open space from OSM and Google Maps.

Boston, MA, a metropolitan city in the northeastern coastal area and Des Moines, IA, a medium-sized city in the middle west were selected to evaluate the performance of the proposed framework in detail. We collected assessor's parcels in Boston and Des Moines from the Boston government online data portal (Anon 2017) and the Polk County Assessor Database (Anon 2019) to identify the functions of reference building footprints. Approximately 5% of building footprints in Boston and Des Moines were not mapped by our framework. In Des Moines, the unclassified building footprints consisted of mobile homes (43%), multi-functional buildings (28%), and industrial structures (10%). Meanwhile, in Boston, the unclassified building footprints included other exempt buildings (33%), multi-functional buildings (32%), and industrial structures (21%). For the 95% of building footprints that were successfully mapped by our framework, we collected a total of 10,000 reference building footprints in a stratified manner to evaluate the performance of building function identification. The numbers of reference buildings in Boston and Des Moines were listed in Table 2. These 10,000 reference buildings were sampled to conduct accuracy assessment 1000 times. Confusion matrices, overall accuracy, Producer's Accuracy (PA), User's Accuracy (UA), and kappa coefficient were calculated in each accuracy assessment.

4. Results

4.1. Accuracy of building function identification

The accuracy assessment indicates that our framework performed well in identifying building functions (Figure 7). In Des Moines and Boston, the average overall accuracy achieved was 93.9% and 93.4%, respectively, accompanied by an average kappa coefficient of 0.62 and 0.63 (Figure 7(c)). As a dominant function, residential function donmonstrated remarkable accuracies, exhibiting both UA and PA values greater than 0.9 (Figure 7(a,b)). Within the category of non-residential functions, our framework showcased exceptional performance in accurately classifying schools, achieving an average UA and PA exceeding 0.7. This successful identification of schools is visually depicted in Figure S2(b-d). However, when it came to classifying offices, our framework demonstrated relatively lower performance, with average UA and PA hovering around 0.5 and 0.6, repectively. As illustrated in Figure S2(c), several buildings were classified as shops and restaurants in our framework, whereas they were labeled as offices in the reference map. Furthermore, the restaurant function demonstrated a notable PA of approximately 0.75, but a comparatively lower UA of around 0.6. Conversely, shops demonstrated a remarkable UA of approximately 0.7, but a significantly lower PA of around 0.4, suggesting that our framework faced challenges in identifying some buildings with the shop function. For example, in the upper area of Figure S2(d), our framework failed to identify several buildings with shop functions. In terms of hospitals and hotels, our framework showed better performance in Des Moines compared to Boston. Especailly for hotels, we achieved an impressive performance in Des Moines, with an average PA of 0.68 and UA of 0.75.

4.2. Evaluation of data collection workflow

Compared to POIs from OSM, POIs from Google Maps in 50 cities had larger coverage and higher density (Figure 8), which can largely benefit building

Table 2. Sample size for seven building functions in representative cities of Boston and Des Moines.

Function	Residence	Office	Shop	Hotel	Hospital	School	Restaurant
Boston	9000	440	320	20	40	130	50
Des Moines	9000	455	350	15	20	90	70

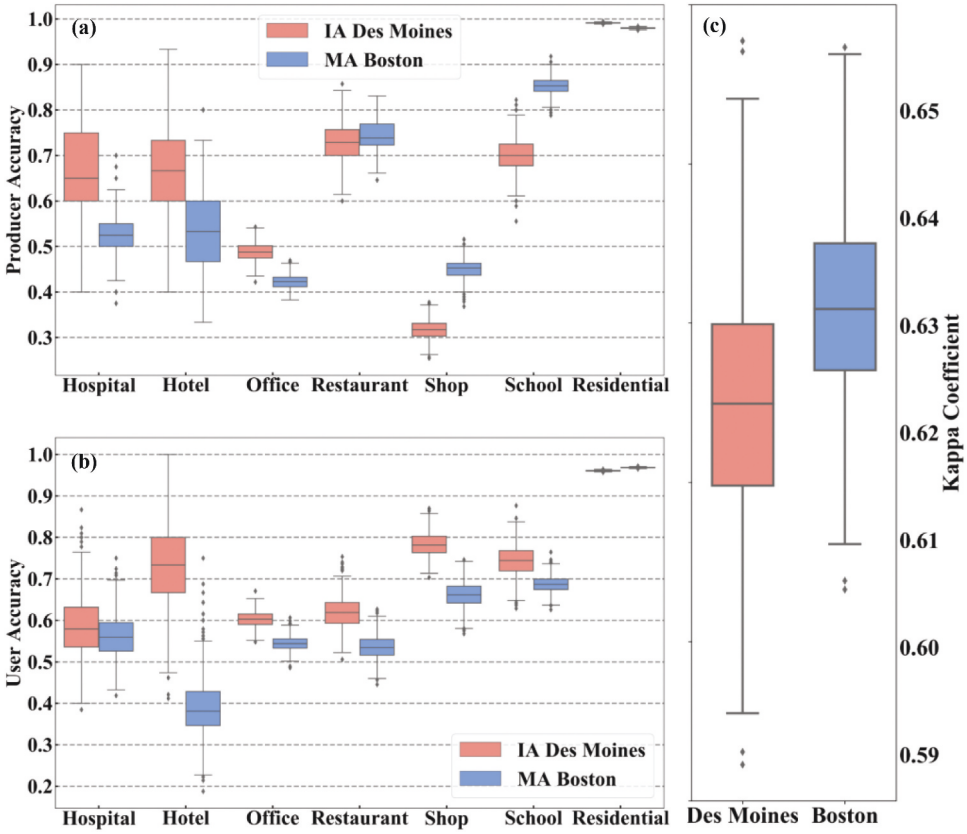

Figure 7. The evaluation of building function maps in Boston, MA and Des Moines, IA including producer accuracy (a), user accuracy (b), and kappa coefficient (c).

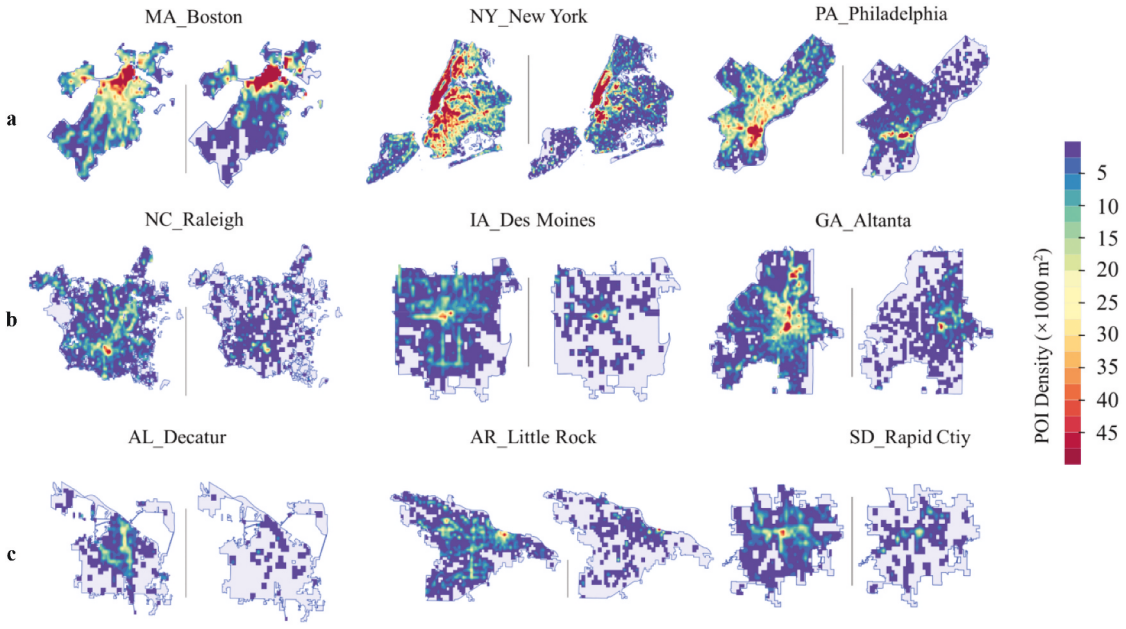

Figure 8. The spatial pattern of POIs density (i.e. the number of POIs in 500 m grids) collected from Google Maps (left) and OSM (right) in large (a), middle (b), and small (c) size cities.

function mapping. We collected a total of over 120,000 POIs via web crawler and a total of 550,000 POIs via map crawler. As Figure 8 shows, POIs from Google Maps were obviously denser than those from OSM in large, middle, and small size cities. The difference of POI density between Google Maps and OSM was small in central business districts of large cities. In middle and small size cities, POIs collected from Google Maps were distributed around the whole city but those from OSM were only distributed in the city center, especially in Des Moines and Rapid City.

Compared to land use parcels from OSM, land use parcels from Google Maps performs better in identifying commercial corridors (Figure 9(d)). The total area of school parcels from OSM was larger than that of Google Maps in big cities, such as Chicago and Houston. In addition, the total area of hospital parcels from OSM was slightly larger than Google Maps in big cities, such as Dayton and New York. There are two reasons leading to smaller areas of land use parcels from Google Maps than OSM. First, as shown in the top of Figure 9(e,f), school and hospital parcels in Google Maps were split by roads with a smaller extent. Second, some school and hospital buildings were not represented by parcels in Google Maps because they only had single buildings in them and these buildings were represented by POIs (Figure 9(e,f) bottom).

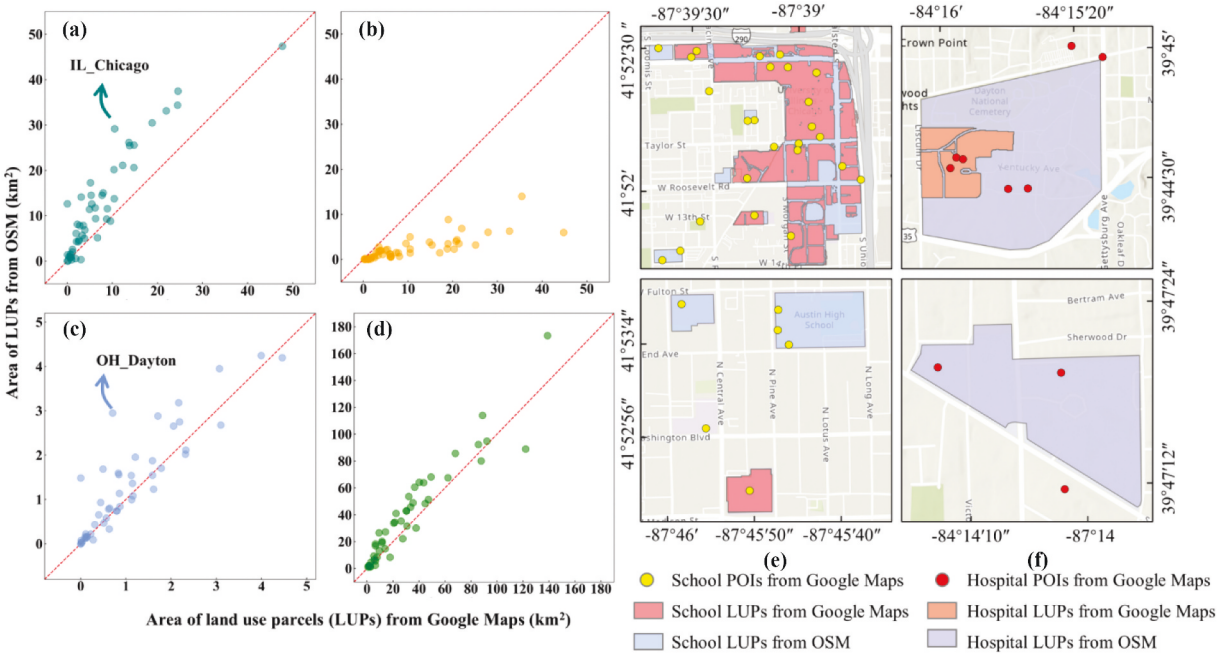

Figure 9. The areas comparison of land use parcels from Google Maps and OSM in (a) school, (b) commercial corridors, (c) hospital, (d) open space. Chicago, IL and Dayton, OH were two cities with the largest area difference of school parcels (e) and hospital parcels (f) from OSM and Google Maps.

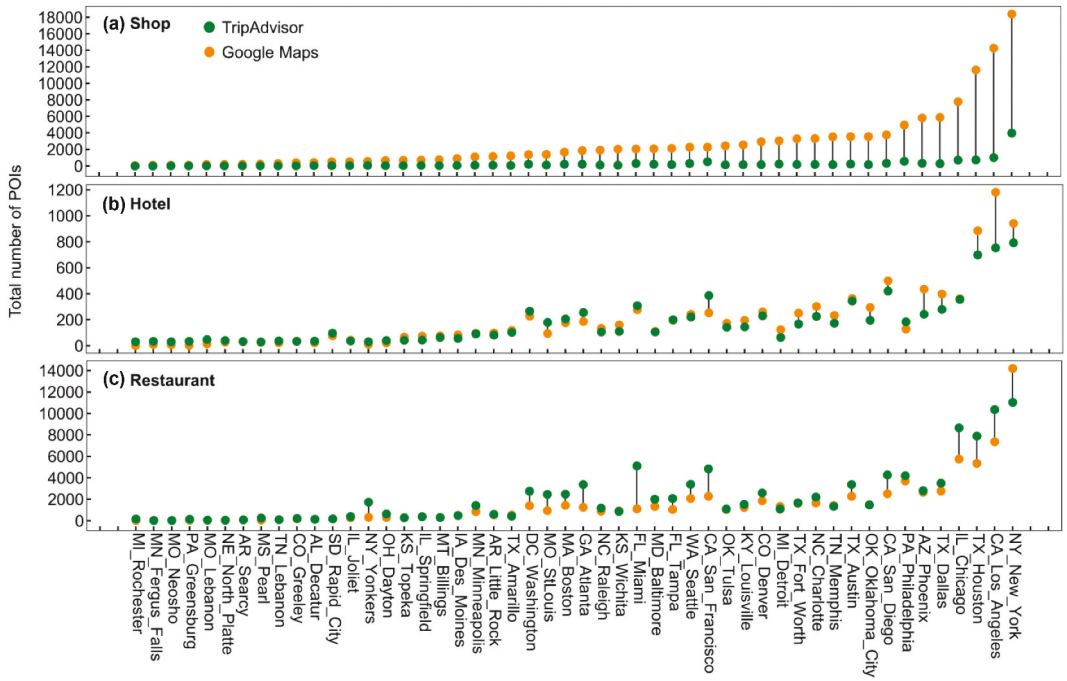

Figure 10. The comparison of (a) shop, (b) hotel, and (c) restaurant POIs derived from Google Maps and TripAdvisor.

The total number of POIs derived from Google Static Maps exceeded those from TripAdvisor, particularly in the case of shop POIs. As shown in Figure 10(a), the disparity in the number of shop POIs between Google Static Maps and TripAdvisor exceeds 10,000 in major metropolitan cities such as New York, Houston, and Los Angeles. This distinction is reasonable since the shops gathered by TripAdvisor predominantly consist of appealing establishments that cater to travelers such as shopping malls, antique stores, and street markets, while the shops derived from Google Static Maps predominantly comprise various stores that are part of our daily lives, such as grocery stores, small retail shops, and drugstores. However, TripAdvisor has demonstrated an advantage in acquiring restaurant POIs in major metropolitan cities and hotel POIs in small cities (Figure 10(b,c)). For instance, in cities such as Los Angeles, Houston, and Chicago, the number of restaurant POIs derived from TripAdvisor exceeded those from Google Maps by over 2000. In cities such as Rochester, Fergus Falls, and Greensburg, the number of hotel POIs derived from TripAdvisor slightly exceeded those from Google Maps.

4.3. Urban building function maps

The spatial patterns of buildings with different functions can be captured by the generated urban building function maps and have similar spatial patterns with reference data. The density of residential buildings identified by our maps shows a similar spatial pattern with the reference data. Figure 11d shows that the densest residential buildings areas with total floor areas larger than 2.7×10^4 were mainly distributed in southeastern Boston and northern Des Moines. In addition, both our map and reference map revealed the spotty pattern of institutional buildings (i.e. hospitals and schools). Figure 11c shows that hospitals and schools with total floor areas larger than 0.9×10^4 were clustered together and sparsely distributed across the city.

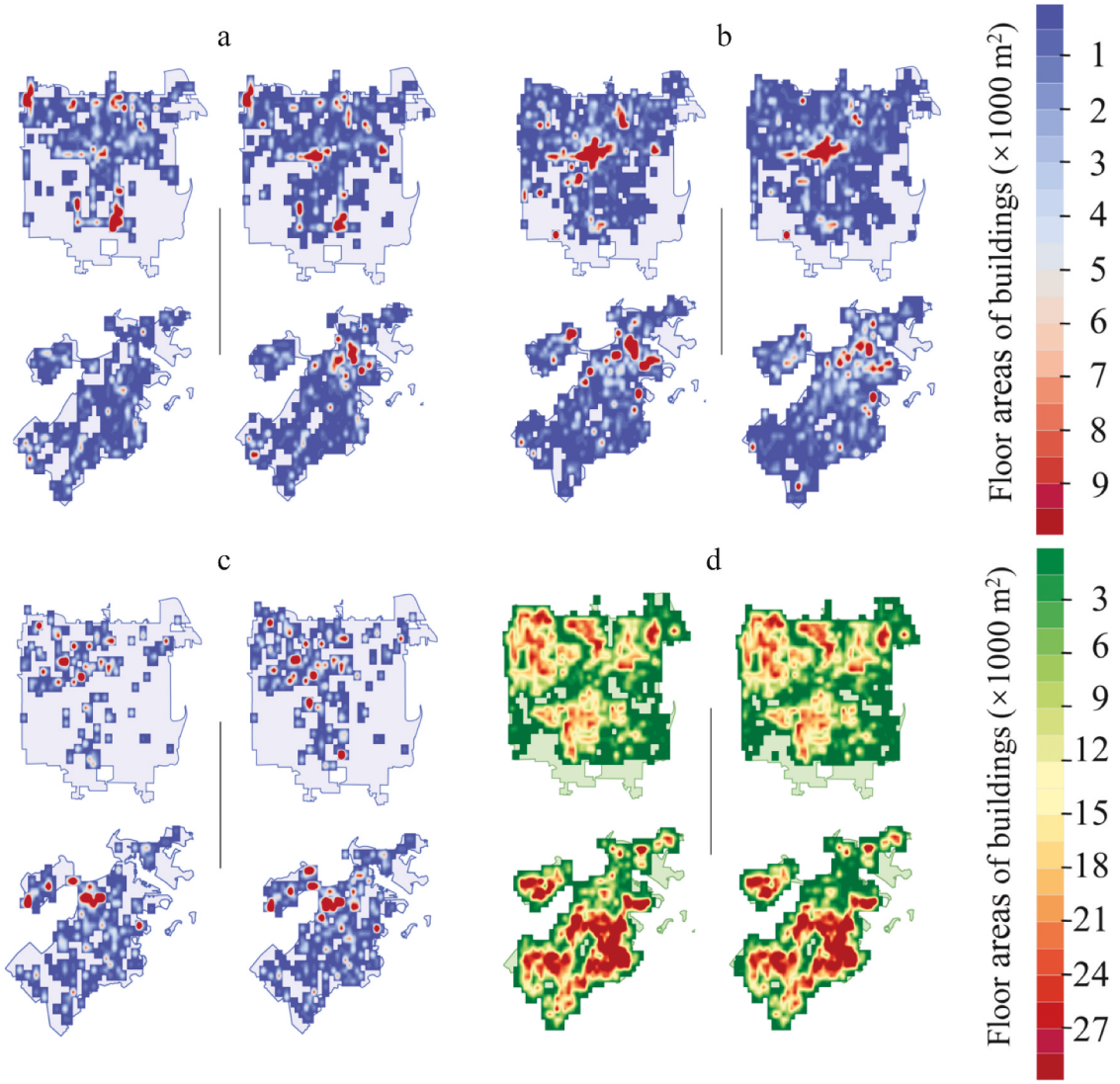

Figure 11. The total floor areas of buildings in 500 m x 500 m grids for (a) commercial, (b) office, (c) institutional, and (d) residential types in Des Moines, IA (upper) and Boston, MA (bottom).

The density of office and commercial types (Figure 11(a,b)) was slightly different with reference maps, especially in Boston. The reason is that restaurants and shops with high PA but low UA were easy to be misclassified as offices, leading to lower density of commercial types and higher density of office types.

5. Discussion

5.1. Uniqueness of the proposed methodology

This study aims to identify non-residential building functions on a large scale through an integration of the data obtained from open-source web platforms. The use of TripAdvisor.com, a life-service website, providing not only reviews and rates for hotels, shops, and restaurants but also their detailed addresses. Employing a combination of geocoding technique and web crawler workflow, these text-based addresses can be converted into POIs, enabling the inference of building functions. Similarly, other life-service websites, such as Realtor.com can be leveraged to extract POIs with residential functions using a comparable approach. Another valuable resource is Google Maps, a popular web mapping platform, which offers static map images encompassing various features such as POIs, land use parcels, and building footprints. By implementing a map crawler workflow, it becomes possible to convert the image-based information into geospatial data, facilitating the inference of building functions. This methodology can be extended to extract additional data from Google static maps, including bus stations, road networks, and parking lots, among other features.

This study also aims to identify residential building functions on a large scale through building footprint-derived landscape metrics. First, we use a dataset of building footprints developed by Microsoft to measure the size, placement, and uniformity of housing. These comprehensive Microsoft data cover the entire nation, making them a potential data source for future expansion into a nationwide study. Second, we investigated suitable machine learning algorithm according to the characteristics of the residential parcels, such as unequal distribution of binary classes and optimal threshold of Num_b. By capturing how these building footprint-derived metrics varies within each parcel, we used the OneClassSVM algorithm and self-adaptive Num_b to successfully distinguish between residential and non-residential parcels.

5.2. Practical applications of the proposed methodology

5.2.1. Workflows for web-based data collection

Our workflows for web-based data collection are highly adaptable with relatively low cost, enabling execution with multiple iterations to maintain up-to-date building function maps. Among Google Maps APIs (e.g. street view, place detailed, and maps static APIs), Google Maps Static API was the most cost- and computation-effective one. Specifically, compared to 28,000 free monthly downloads for Google static street view APIs that many urban studies utilized (Richards and Wang 2020; Zeng et al. 2018), Google Maps Static APIs can have requested for free up to 100,000 downloads per month. In addition, although both Google places detailed and Maps Static APIs had 100,000 free monthly downloads, one Google places detailed API request can only be used to obtain information of one POI. By using our workflow of data collection, one Google Maps Static API request can be used to obtain the location and type of many POIs. Therefore, by utilizing a web crawler on TripAdvisor.com at no cost and employing a map crawler on the Google Maps Static API at a low cost, we have the capability to collect current POIs and land use parcels on an annual basis. These collected data sets are then used to update the building functions for existing buildings, ensuring our building function maps up to date.

Our workflows for web-based data collection were replicable across geographies, which can support geospatial data-driven urban studies over large areas. Web-based mapping platforms are ideal data sources for geospatial data-driven urban studies with its copiousness, large area coverage and reliability (Chao et al. 2018; Wang, Li, and Shi 2017). With the help of the map crawler, important map elements such as building footprints, roads, parks, and bus stops in Google Maps can be retrieved by assigning them with different RGB codes in map styles, providing useful geospatial datasets for urban sustainable studies at a large scale. For example, compared to OSM road networks, road polygons with width and types (i.e. highway, arterial, drivable local, and trail traffic) can be collected by the map crawler, supporting the quantification of pedestrian exposure to traffic particulate matter (Qiu et al. 2017). Bus stops and parks collected by the map crawler can be used to assess the equitable availability of public open space (Timperio et al. 2007) and optimize public transportation systems for increasing accessibility of urban green space (Chen and Chang 2015).

5.2.2. Web-based building function mapping

Our framework exhibits high transferability to other cities due to its ease of execution and the utilization of data sources with worldwide coverage. Google Static Maps offer almost complete coverage around the world, even cities located in the southern hemisphere of the globe (e.g. Rustenburg in South Africa, Rocha in Uruguay). TripAdvisor.com, as the largest travel site in the world operating in 26 countries, can provide travel-related POIs (i.e. shops, restaurants, and hotels)

in tourist cities around the world. To meet the demand for mobile navigation and life services in daily life, geospatial data from web mapping platforms need to be frequently updated (Chen et al. 2020). In addition, with the goal to increase the coverage of building footprint data available to public, Microsoft have recently released building footprints in South America, Africa, and Australia. Our framework incorporates straightforward geospatial analysis and machine learning algorithms, allowing for easy execution. As the data resources receive updates, our framework enables the identification of building functions for new structures and the updating of building functions for existing ones through multiple iterations.

Our framework can contribute to quantitative urban studies, especially on a large scale. For example, bottom-up urban building energy use or heat emission models (Y. Chen et al. 2022; Li et al. 2017) need detailed building information such as size, height, and function to estimate spatial and temporal patterns of energy consumptions or heat emissions at a fine scale. Considering the existing large scale building footprint (Anon 2018) and height (Li et al. 2020) datasets, large scale building function maps can contribute to quantifying large-scale building energy uses or heat emissions. Therefore, our building function maps have potential to pave a way for urban building energy modeling to investigate urban building energy uses and heat emissions under different climate backgrounds, offering support for government policy making and sustainable city development planning.

5.3. Legal regulations on web crawler technology

In this study, we have used two ways to retrieve data from web-based platforms. One approach was to utilize the Maps Static APIs, which are provided by Google to establish a connection with their Google Maps service. The Google Maps API operates on a pay-as-you-go pricing structure, offering a monthly $200 USD credit for each billing account within the Google Maps platform. The $200 USD credit enables us to perform up to 100,000 requests to the Maps Static API or 50,000 requests to the Geocoding API. In this study, we used a total of $800 USD free credit over a span of four months, which enabled us to conduct 121,631 requests to the Geocoding API, 809,090 Maps Static API requests at the zoom level of 18, and 97,501 Maps Static API requests at the zoom level of 15. Providing APIs has become a common practice among various web mapping platforms, including Baidu Maps, MapQuest, and Bing Maps, etc., enabling users to connect with their mapping services. Therefore, the proposed map crawler proves to be practical and effective, as long as the web mapping platforms maintain their current method of data sharing through APIs.

Furthermore, there exists publicly accessible data on the internet that has not been structured for direct downloading or is inaccessible through an API. To obtain this content, it is necessary to scrape it from websites using programming code, as it is accessible and viewable within web browsers. Therefore, specialized packages such as "Beautiful Soup" and "Scrapy" offer effective solutions in this regard. On 17 April 2019, the European Union introduced a legal framework for Text and Data Mining (TDM) on copyright and related rights in the Digital Single Market (DSM Directive) (Egger, Kroner, and Stöckl 2022). The DSM Directive grants TDM an exception in regard to reproductions and extractions made by research organizations and cultural heritage institutions in order to carry out, for the purposes of scientific research to which lawful access is acknowledged. Thus, the proposed web crawler designed to extract publicly displayed addresses from TripAdvisor.com is exclusively intended for scientific research purposes. To date, TripAdvisor.com has been a valuable source for conducting big data analysis, as numerous researchers have utilized its extensive information on online review ratings to understand customer behaviors (Khorsand, Rafiee, and Kayvanfar 2020; Mariani, Borghi, and Laker 2023; O'connor 2008). By implementing our proposed web crawler, TripAdvisor.com will become an invaluable resource for inferring building functions.

APIs and web scraping are two standard methods to collect data from websites but the preference leans toward utilizing APIs. The operation of APIs is typically governed by the terms and conditions outlined by the provider. This framework ensures that the likelihood of encountering legal complications is minimized if we remain in alignment with these terms and guidelines. In contrast, web scraping necessitates strict adherence to data privacy regulations stipulated by commercial entities to maintain the legality. Krotov and Silva (2018) have generated a list of inquiries to assess whether web scraping projects can potentially result in lawsuits or ethical controversies, serving as a valuable resource for gauging the likelihood of lawsuits or ethical disputes arising from such projects. The inquiry list includes the following important questions: does "terms of use" of the websites explicitly forbid web crawling? Could crawling and scraping potentially result in substantial damage to the website? Additionally, does the acquisition of data from the website have the potential to undermine personal privacy?

5.4. Future work

This study opens future research avenues of mapping multi-function buildings and building functions with limited geospatial data. First, the proposed framework

only identified single function of buildings, resulting in its low performance on correctly classifying offices. In cases where a building encompasses multiple POIs, our framework determines the building's function based on the majority function among the POIs. Leveraging Google Street View images holds great promise in identifying buildings with multiple functions, as these images offer building profile pictures captured from diverse fields-of-view (Li, Zhang, and Li 2017). The profile view of street-level images can be effectively utilized to assess the socio-economic functions of an individual building across various aspects, such as discerning a restaurant on the left side and an office on the right side, as well as vertical levels, such as identifying a restaurant at the ground level and offices on the upper levels.

Second, the proposed framework could perform better in industrialized nations because cities in these nations tend to have abundant geospatial data in web-based platforms (Anguelov et al. 2010). Although data in Google Maps and TripAdvisor.com exist in the southern hemisphere of the globe (e.g. São Paulo in Brazil and Kampala in Africa), the coverage is less dense in those areas compared to industrialized nations (Anguelov et al. 2010). Therefore, the proposed framework with low density of geospatial data may have limited ability to identify building functions in the Global South. However, before the global coverage of Google Maps and TripAdvisor databases expanded to developing nations (Richards and Wang 2020), auxiliary datasets (e.g. remote sensing observations) could be used to improve building function mapping, especially in the Global South.

6. Conclusion

Building function map can provide an important source of data for characterizing human activities in the complex urban environment. Although social sensing-based methods are capable of identifying detailed building functions such as hospitals and schools for a large area, collecting social sensing datasets is expensive and difficult. In this paper, we present a framework for mapping building functions based on web-based geospatial datasets and implemented this framework in 50 U.S. cities with different sizes. Additionally, we allocated approximately one week of computing time on a server equipped with 50 threads for the purpose of web-based data collection and the construction of function identification (https://researchit.las.iastate.edu). The accuracy assessment indicates that the proposed framework performed well with average overall accuracies of 93.9% and 93.4% and average kappa coefficients of 0.62 and 0.63 in Des Moines and Boston, respectively. The mapped building functions can contribute to quantitative urban modeling studies, such as city-scale building energy modeling. In addition, the proposed workflows for web-based data collection have potential to support a variety of urban environmental studies, such as evaluation of urban green space availability and pedestrian exposure to traffic particulate matter. Considering that the Google Maps and TripAdvisor.com did not have residential POIs and had limited spatial coverage in the Global South, future research can focus on improving accuracy of building function mapping by identifying multi-function buildings and building functions in the southern hemisphere of the globe.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work is supported by the National Science Foundation [grant numbers 1854502 and 1855902]. Publication was made possible in part by support from the HKU Libraries Open Access Author Fund sponsored by the HKU Libraries. USDA is an equal opportunity provider and employer. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture.

Notes on contributors

Wei Chen received her M.S. degree in cartography and geographic information system from the University of Chinese Academy of Sciences, Beijing, China, in 2019. She is currently pursuing the Ph.D. degree at the Department of Geological and Atmospheric Sciences, Iowa State University, Ames, IA, USA. Her research interests include spatial data analysis, urban remote sensing, building energy use, and weather modelling.

Yuyu Zhou received the Ph.D. degree in environmental sciences from the University of Rhode Island, Kingston, RI, USA, in 2008. He is currently a Professor with the Department of Geography, University of Hong Kong, Hong Kong, China. His research interests include the applications of remote sensing, GIS, integrated assessment modeling, and spatial analysis to understand the problems of environmental change and their potential solutions.

Eleanor C. Stokes received the Ph.D. degree in environmental science from the Yale University School of Forestry and Environmental Studies, New Haven, CT, USA, in 2018. She is currently a senior scientist in NASA and the acting science PI of the NASA Black Marble nighttime light product. She is a remote sensing and urban land scientist interested in the planetary impacts of urbanization.

Xuesong Zhang received the Ph.D. degree in hydrology & watershed management from Texas A&M University, College Station, TX in 2008. He is currently a research physical scientist at USDA-ARS Hydrology and Remote

Sensing Laboratory, Beltsville, MD, USA. His research interests include hydrologic and agroecosystem models, ecosystem characterization and modeling, as well as terrestrial and aquatic biogeochemical cycles.

ORCID

Wei Chen http://orcid.org/0000-0002-5431-8837
Yuyu Zhou http://orcid.org/0000-0003-1765-6789
Eleanor C. Stokes http://orcid.org/0000-0002-0204-8847
Xuesong Zhang http://orcid.org/0000-0003-4711-7751

Data and code availability statement

The derived data that support the findings of this study are available at https://maps.google.com and https://www.tripadvisor.com with the permission of Google and TripAdvisor. The code examples of web crawler and map crawler for extracting geospatial data from TripAdvisor and Google Maps are available at https://github.com/Vivid-Urban/Web-based-data-collection.git. The resultant building function map is available from the corresponding author upon reasonable request.

References

Anguelov, D., C. Dulong, D. Filip, C. Frueh, S. Lafon, R. Lyon, A. Ogale, L. Vincent, and J. Weaver. 2010. "Google Street View: Capturing the World at Street Level." *Computer* 43 (6): 32–38. https://doi.org/10.1109/MC.2010.170.

Anon. 2017. *Boston Government Online Data Portal*. https://data.boston.gov/.

Anon. 2018. *Microsoft US Building Footprints*.https://github.com/Microsoft/USBuildingFootprints.

Anon. 2019. *Polk County Assessor Database*. http://www.assess.co.polk.ia.us/web/exports/basic/occgroupI.html.

Arunplod, C., M. Nagai, K. Honda, and P. Warnitchai. 2017. "Classifying Building Occupancy Using Building Laws and Geospatial Information: A Case Study in Bangkok." *International Journal of Disaster Risk Reduction* 24:419–427. https://doi.org/10.1016/j.ijdrr.2017.07.006.

Chang, Y.-C., C.-H. Ku, and C.-H. Chen. 2019. "Social Media Analytics: Extracting and Visualizing Hilton Hotel Ratings and Reviews from TripAdvisor." *International Journal of Information Management* 48:263–279. https://doi.org/10.1016/j.ijinfomgt.2017.11.001.

Chao, H., Y. Cao, J. Zhang, F. Xia, Y. Zhou, and H. Shan. 2018. "Population Density-Based Hospital Recommendation with Mobile LBS Big Data." In *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)*. Shanghai, January 15-17.

Chen, J., and Z. Chang. 2015. "Rethinking Urban Green Space Accessibility: Evaluating and Optimizing Public Transportation System Through Social Network Analysis in Megacities." *Landscape and Urban Planning* 143:150–159. https://doi.org/10.1016/j.landurbplan.2015.07.007.

Chen, Y., J. Yang, R. Yang, X. Xiao, and J. C. Xia. 2022. "Contribution of Urban Functional Zones to the Spatial Distribution of Urban Thermal Environment." *Building & Environment* 216:109000. https://doi.org/10.1016/j.buildenv.2022.109000.

Chen, W., Y. Zhou, Q. Wu, G. Chen, X. Huang, and B. Yu. 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, W., Y. Zhou, Y. Xie, G. Chen, K. J. Ding, and D. Li. 2022. "Estimating Spatial and Temporal Patterns of Urban Building Anthropogenic Heat Using a Bottom-Up City Building Heat Emission Model." *Resources Conservation & Recycling* 177:105996. https://doi.org/10.1016/j.resconrec.2021.105996.

Davila, C. C., C. F. Reinhart, and J. L. Bemis. 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.

Durst, N. J., E. Sullivan, H. Huang, and H. Park. 2021. "Building Footprint-Derived Landscape Metrics for the Identification of Informal Subdivisions and Manufactured Home Communities: A Pilot Application in Hidalgo County, Texas." *Land Use Policy* 101:105158. https://doi.org/10.1016/j.landusepol.2020.105158.

Egger, R., M. Kroner, and A. Stöckl. 2022. "Web Scraping: Collecting and Retrieving Data from the Web." In *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications*, edited by R. Egger, 67–82. Cham: Springer International Publishing.

Hu, J., S. You, and U. Neumann. 2003. "Approaches to Large-Scale Urban Modeling." *IEEE Computer Graphics and Applications* 23 (6): 62–69. https://doi.org/10.1109/mcg.2003.1242383.

Khorsand, R., M. Rafiee, and V. Kayvanfar. 2020. "Insights into TripAdvisor's Online Reviews: The Case of Tehran's Hotels." *Tourism Management Perspectives* 34:100673. https://doi.org/10.1016/j.tmp.2020.100673.

Kousa, A., J. Kukkonen, A. Karppinen, P. Aarnio, and T. Koskentalo. 2002. "A Model for Evaluating the Population Exposure to Ambient Air Pollution in an Urban Area." *Atmospheric Environment* 36 (13): 2109–2119. https://doi.org/10.1016/s1352-2310(02)00228-5.

Krotov, V., and L. Silva. 2018. "Legality and Ethics of Web Scraping." In *Americas Conference on Information Systems 2018: Digital Disruption*. New Orleans, August 16-18.

Kunze, C., and R. Hecht. 2015. "Semantic Enrichment of Building Data with Volunteered Geographic Information to Improve Mappings of Dwelling Units and Population." *Computers, Environment and Urban Systems* 53:4–18. https://doi.org/10.1016/j.compenvurbsys.2015.04.002.

Li, W., Y. Zhou, K. Cetin, J. Eom, Y. Wang, G. Chen, and X. Zhang. 2017. "Modeling Urban Building Energy Use: A Review of Modeling Approaches and Procedures." *Energy* 141:2445–2457. https://doi.org/10.1016/j.energy.2017.11.071.

Li, W., Y. Zhou, K. S. Cetin, S. Yu, Y. Wang, and B. Liang. 2018. "Developing a Landscape of Urban Building Energy Use with Improved Spatiotemporal Representations in a Cool-Humid Climate." *Building & Environment* 136:107–117. https://doi.org/10.1016/j.buildenv.2018.03.036.

Li, X., C. Zhang, and W. Li. 2017. "Building Block Level Urban Land-Use Information Retrieval Based on Google Street View Images." *GIScience & Remote Sensing* 54 (6): 819–835. https://doi.org/10.1080/15481603.2017.1338389.

Li, X., Y. Zhou, P. Gong, K. C. Seto, and N. Clinton. 2020. "Developing a Method to Estimate Building Height from Sentinel-1 Data." *Remote Sensing of Environment* 240:111705. https://doi.org/10.1016/j.rse.2020.111705.

Liu, X., and Y. Long. 2016. "Automated Identification and Characterization of Parcels with OpenStreetmap and Points of Interest." *Environment & Planning B: Planning & Design* 43 (2): 341–360. https://doi.org/10.1177/0265813515604767.

Liu, X., N. Niu, X. Liu, H. Jin, J. Ou, L. Jiao, and Y. Liu. 2018. "Characterizing Mixed-Use Buildings Based on Multi-Source Big Data." *International Journal of Geographical Information Science* 32 (4): 738–756. https://doi.org/10.1080/13658816.2017.1410549.

Liu, Y., X. Liu, S. Gao, L. Gong, C. Kang, Y. Zhi, G. Chi, and L. Shi. 2015. "Social Sensing: A New Approach to Understanding Our Socioeconomic Environments." *Annals of the Association of American Geographers* 105 (3): 512–530. https://doi.org/10.1080/00045608.2015.1018773.

Luo, X., P. Vahmani, T. Hong, and A. Jones. 2020. "City-Scale Building Anthropogenic Heating During Heat Waves." *Atmosphere* 11 (11): 1206. https://doi.org/10.3390/atmos11111206.

Mariani, M. M., M. Borghi, and B. Laker. 2023. "Do Submission Devices Influence Online Review Ratings Differently Across Different Types of Platforms? A Big Data Analysis." *Technological Forecasting & Social Change* 189:122296. https://doi.org/10.1016/j.techfore.2022.122296.

Niu, N., X. Liu, H. Jin, X. Ye, Y. Liu, X. Li, Y. Chen, and S. Li. 2017. "Integrating Multi-Source Big Data to Infer Building Functions." *International Journal of Geographical Information Science* 31 (9): 1871–1890. https://doi.org/10.1080/13658816.2017.1325489.

NOAA. 2021. *Storm Events Database*. https://www.ncdc.noaa.gov/stormevents/.

O'connor, P. 2008. "User-Generated Content and Travel: A Case Study on Tripadvisor. Com." In *Information and Communication Technologies in Tourism 2008*, edited by P. O'Connor, W. Höpken, and U. Gretzel, 47–58. Vienna: Springer.

Qiu, Z. W., X. Q. Xu, J. H. Song, Y. P. Luo, R. N. Zhao, B. H. Xiang, W. C. Zhou, X. X. Li, and Y. Z. Hao. 2017. "Pedestrian Exposure to Traffic PM on Different Types of Urban Roads: A Case Study of Xi'an, China." *Sustainable Cities and Society* 32:475–485. https://doi.org/10.1016/j.scs.2017.04.007.

Richards, D., and J. W. Wang. 2020. "Fusing Street Level Photographs and Satellite Remote Sensing to Map Leaf Area Index." *Ecological Indicators* 115:106342. https://doi.org/10.1016/j.ecolind.2020.106342.

Timperio, A., K. Ball, J. Salmon, R. Roberts, and D. Crawford. 2007. "Is Availability of Public Open Space Equitable Across Areas?" *Health & Place* 13 (2): 335–340. https://doi.org/10.1016/j.healthplace.2006.02.003.

Wang, C., Y. Li, and X. Shi. 2017. "Information Mining for Urban Building Energy Models (UBEMs) from Two Data Sources: OpenStreetmap and Baidu Map." *Energy Buding* 157:166–175. https://doi.org/10.26868/25222708.2019.210545.

Xu, X., J. Ou, P. Liu, X. Liu, and H. Zhang. 2021. "Investigating the Impacts of Three-Dimensional Spatial Structures on CO2 Emissions at the Urban Scale." *Science of the Total Environment* 762:143096. https://doi.org/10.1016/j.scitotenv.2020.143096.

Yeh, C.-H., C.-H. Loh, and K.-C. Tsai. 2006. "Overview of Taiwan Earthquake Loss Estimation System." *Natural Hazards* 37 (1): 23–37. https://doi.org/10.1007/s11069-005-4654-z.

Yuan, J., Y. Zheng, and X. Xie. 2012. "Discovering Regions of Different Functions in a City Using Human Mobility and POIs." In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Beijing, August 12-16.

Zeng, L., J. Lu, W. Li, and Y. Li. 2018. "A Fast Approach for Large-Scale Sky View Factor Estimation Using Street View Images." *Building & Environment* 135:74–84. https://doi.org/10.1016/j.buildenv.2018.03.009.

Zhang, Y., Q. Li, H. Huang, W. Wu, X. Du, and H. Wang. 2017. "The Combined Use of Remote Sensing and Social Sensing Data in Fine-Grained Urban Land Use Mapping: A Case Study in Beijing, China." *Remote Sensing* 9 (9): 865. https://doi.org/10.3390/rs9090865.

Zhou, Y., X. Li, W. Chen, L. Meng, Q. Wu, P. Gong, and K. C. Seto. 2022. "Satellite Mapping of Urban Built-Up Heights Reveals Extreme Infrastructure Gaps and Inequalities in the Global South." *Proceedings of the National Academy of Sciences* 119 (46): e2214813119. https://doi.org/10.1073/pnas.2214813119.

Zhuo, L., Q. Shi, C. Zhang, Q. Li, and H. Tao. 2019. "Identifying Building Functions from the Spatiotemporal Population Density and the Interactions of People Among Buildings." *ISPRS International Journal of Geo-Information* 8 (6): 247. https://doi.org/10.3390/ijgi8060247.