Chen and Biljecki *Urban Informatics* (2022) 1:12
https://doi.org/10.1007/s44212-022-00012-2

Urban Informatics

ORIGINAL ARTICLE

Open Access

Mining real estate ads and property transactions for building and amenity data acquisition

Xinyu Chen[1] and Filip Biljecki[2,3*]

Abstract

Acquiring spatial data of fine and dynamic urban features such as buildings remains challenging. This paper brings attention to real estate advertisements and property sales data as valuable and dynamic sources of geoinformation in the built environment, but unutilised in spatial data infrastructures. Given the wealth of information they hold and their user-generated nature, we put forward the idea of real estate data as an instance of implicit volunteered geographic information and bring attention to their spatial aspect, potentially alleviating the challenge of acquiring spatial data of fine and dynamic urban features. We develop a mechanism of facilitating continuous acquisition, maintenance, and quality assurance of building data and associated amenities from real estate data. The results of the experiments conducted in Singapore reveal that one month of property listings provides information on 7% of the national building stock and about half of the residential subset, e.g. age, type, and storeys, which are often not available in sources such as OpenStreetMap, potentially supporting applications such as 3D city modelling and energy simulations. The method may serve as a novel means to spatial data quality control as it detects missing amenities and maps future buildings, which are advertised and transacted before they are built, but it exhibits mixed results in identifying unmapped buildings as ads may contain errors that impede the idea.

Keywords: Webscraping, Deep learning, Crowdsourcing, Urban digital twin, GeoAI, Computer vision

1 Introduction

Geospatial data on buildings is important for a wide array of applications. For example, they can be used to study the urban fabric, while adding building attributes such as their type and height facilitates generating 3D building models, energy simulations, climate studies, disaster mitigation, land administration, and urban morphology analyses (Park and Guldmann 2019; Li et al. 2020b; Huang & Wang 2020; Agugiaro et al. 2020; Yuan et al. 2020; Palliwal et al. 2021; Abdelrahman et al. 2021; León-Sánchez et al. 2021; Ning et al. 2021; Wu & Biljecki 2021; Koeva et al. 2021; Bourdeau et al. 2019; Chen et al. 2020; Li et al. 2021; Florio et al. 2021; Hopf 2018). Information on nearby amenities (POIs) and the surroundings are also important in this context, as they are often associated with buildings, e.g. as indicators of housing value, demographics, and accessibility (Feng & Humphreys 2012; Kang et al. 2021; Yang et al. 2021; Mirkatouli et al. 2018; Szarka & Biljecki 2022; Su et al. 2021).

However, in practice, such data is still complex to obtain, and many issues prevail despite the significant developments in GIScience and remote sensing communities such as proliferation of Volunteered Geographic Information (VGI), i.e. OpenStreetMap (OSM), and advancements in data acquisition techniques. First, such data remains unavailable for most of the world, especially considering open data instances. Second, when such features are mapped, they often lack semantic information (attributes), e.g. year of construction, number of

*Correspondence: filip@nus.edu.sg

[2] Department of Architecture, National University of Singapore, Singapore, Singapore
Full list of author information is available at the end of the article

storeys, and type of building. This omission is most evident in recent efforts mapping buildings at a large-scale but without considering any descriptive information on them (Huang et al. 2020; Li et al. 2020a; Sirko et al. 2021). Third, once acquired, such data is challenging to maintain. With rapid urbanisation, new buildings and amenities are being built continuously while old ones are being demolished, making it difficult to keep databases up to date. Further, building data, like almost any other set of geospatial information, is not always entirely correct — data quality issues, such as completeness, attribute accuracy and positional accuracy continue to be of significant concern in many datasets and geographies around the world. Finally, features may change certain properties during their life-cycle, e.g. buildings may be repurposed, and such changes may not be reflected in a dataset.

On the other front, buildings, accompanied by amenities that serve them, are a prominent class of real estate. When they are advertised and acquired, various data records describing their characteristics from the real estate point of view are generated. Such real estate data, primarily advertisements (listings) posted on property websites (i.e. online marketplaces) and data on transactions after these properties are acquired, capture much of the characteristics of buildings. These properties are directly or indirectly overlapping with many attributes that are usually collected for typical GIS databases in the built environment and are used in a large number of analyses and urban studies. However, these datasets, which are dynamic and come in various flavours, sources and services, and which are often available openly, are almost entirely disconnected from spatial data infrastructures and geospatial developments.

First of all, in this paper, we posit that real estate data can be exploited to provide value for geospatial researchers and practitioners by collecting building data that may be relevant for a range of analyses in the urban context and geospatial tasks such as quality assessment of existing data. By unfolding this idea and providing a proof of concept and a prototype, we seek to bridge the gap between real estate and geospatial data.

The idea of collecting spatial data indirectly, finding proxies for their characteristics, and amplifying data from other domains to serve GIScience are not new, e.g. Yin et al. (2020) analyse movement trajectories to infer attributes of roads, Chen et al. (2021b) mine social media data to map and understand amenities, Lines & Basiri (2021) exploit obstructions in satellite signals to reconstruct the vertical extent of buildings, Wu & Biljecki (2022) map buildings from street networks, Milojevic-Dupont et al. (2020) infer the heights of buildings by developing a regression model that predicts them from the characteristics of the footprint and surrounding context, and Delmelle & Nilsson (2021) assess the ability of using property listing text for neighbourhood type prediction. However, to the extent of our knowledge, the potential of real estate data in building and amenity data acquisition remains uninvestigated despite their abundance, which is the key contribution of this paper.

An illustration of our hypothesis and the work is given in Fig. 1. Real estate data mainly span two forms: texts and images. For example, transaction data are mostly recorded in texts (e.g. address, price, and flat type), while typical rental or sale listings can have both text descriptions and images of the property and its attached amenities such as common spaces, gyms, and swimming pools. In most cases, such records are about subdivisions of buildings (e.g. units, flats). However, even when they pertain to a subset of a building (e.g. one out of a few hundreds of flats in a residential estate), they are representative of entire buildings, as by extension, they provide building information such as the building's year of construction, tenure, location, and common amenities. Some of these forms of data provide information that could be extracted directly without much effort, while some would require a degree of processing. For example, text and/or photos in ads often feature amenities such as sport courts that are part of or are near the advertised property. In such instances, these amenities could be detected from ads using computer vision approaches which are now mature and readily available for that purpose (Chen et al. 2021a). Further, textual data may require processing as well. Free-text descriptions about a property may contain valuable information that could be extracted using basic text mining or natural language processing techniques.

The same goes for transaction data. For example, while the information about the number of storeys of a building may be available in an ad, depending on the jurisdiction and other aspects, it may also be available indirectly from transaction data, from the address (i.e. unit number) of the apartment that is sold. That is, provided a long horizon of transaction data, we may be able to pick up the apartment sold on the highest floor (or at least very close to it), presenting an equivalent of the height of a building with an accuracy sufficiently reasonable for a number of analyses and to indicate the rough urban form (Manoli et al. 2019; Liu et al. 2020; Wang et al. 2021).

The extracted and processed data may provide value for multiple applications. Continuing elaborating the illustrated concept, here we provide four examples. First, the location of the building and presence of an amenity such as a swimming pool in its vicinity (e.g. deduced from a photo in the listing of a currently advertised apartment), may be used to check the content of an existing spatial database such as OSM for quality. If the feature is missing from the building's vicinity, the spatial dataset can be

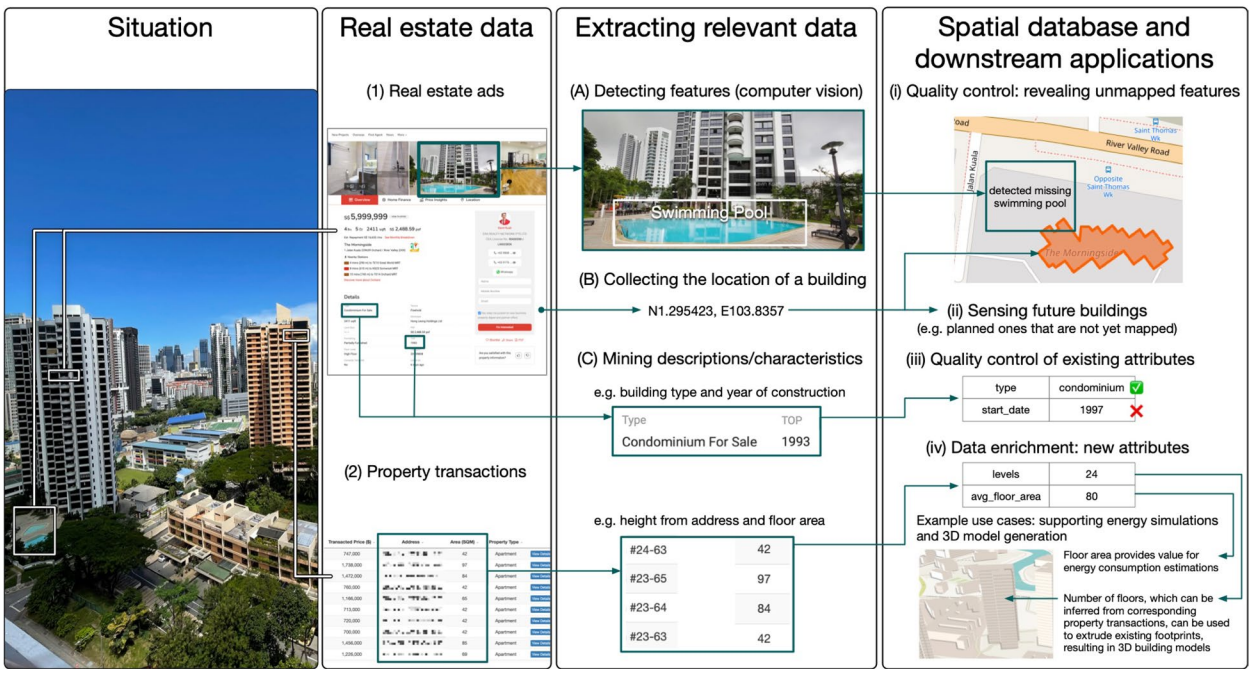


Fig. 1 The idea of our research: real estate datasets (i.e. property ads and transactions) provide a wealth of information on buildings but are a latent type of data in GIS that has a diverse set of untapped applications — ranging from quality control to rapid data updates thanks to their omnipresence, amount of information, and dynamic nature. This illustration is based on real data of a setting in Singapore. Notes: (i) the exemplified property transaction data has been adopted from another undisclosed building to preserve privacy; (ii) this example is simplified and it is not exhaustive, as it presents rather a subset of data that is in the scope of this research; (iii) some building information may be found in both ads and transactions (e.g. year of construction, floor area of apartments). Data sources: OpenStreetMap, PropertyGuru, Urban Redevelopment Authority of Singapore, and OSM Buildings

flagged as suffering from completeness issues with the ability of identifying issues at a high spatial resolution. Second, extracted descriptive information about buildings, such as their type and year of construction, may be compared to the existing attributes in a spatial database, potentially indicating a discrepancy warranting further attention (e.g. errors or outdated records). Third, properties may be listed and also sold even before the buildings started being constructed (off-plan, pre-sales properties). Therefore, ads and transactions may hold value as a signal for new or future buildings that are yet to be mapped, and provide information about the future situation that may support use cases in urban studies and democratisation of the planning process, and make such data available to a broader audience as data on future developments is rarely available openly, especially in datasets such as OSM. Fourth, the extracted information can be used to enhance the database, if they have not been available previously, e.g. number of storeys of a building, which is often the case. Expanding the set of attributes in a spatial database may open the door for further spatial analyses that require such information, but have not been possible previously due to the lack of such data. Since nowadays online real estate marketplaces are prevalent around the world, unlike (open) data on buildings, we believe that such idea holds great potential. Therefore, much of our work focuses on investigating how can we make the best out of such data in the geospatial realm, by developing a proof of concept and developing experiments for various scenarios.

Another characteristic of real estate data is that they are highly dynamic. Their continuous update is especially true for advertisement services in which every minute multiple ads representing a portion of real estate in a city may be added. As these data are uploaded by the users of the property websites, we postulate that real estate data may be considered as a latent form of volunteered geographic information (VGI), and one that warrants further investigations, as contemplated above (Goodchild 2007). To be more specific, for the first time, we deem that they are a type of passive and implicitly volunteered VGI (Craglia et al. 2012; See et al. 2016; Ghermandi & Sinclair 2019; Hopf 2018), as contributing spatial data is not the contributor's primary intention, similarly to social media and geo-tagged imagery such as Flickr (Yan et al. 2017, 2018). Considering real estate data from such an angle, we posit

that they may also double as reference data to enrich and verify building and amenity databases, which has not been investigated yet even though it has various application in a few other areas such as socio-economic studies (You et al. 2017; Liu et al. 2019; Kang et al. 2020; Su et al. 2021). This topic is also relevant in the context of the growing interest in VGI among the smart city and sustainable development communities (Milojevic-Dupont & Creutzig 2021; Nitoslawski et al. 2019).

With the aim of developing a new method of improving geospatial databases of buildings and amenities, we pursue the following research questions: What is the potential of using real estate data to update or create spatial databases of buildings and associated amenities? How can we develop an automated mechanism to collect, maintain, and ensure the quality of building and amenity information in spatial databases? For maintaining a database and assuring its quality, there are two lines of work that we consider. The first one is focused on developing a new data quality assessment method that is investigating whether we can leverage real estate data to examine the completeness of the building footprints and locations of amenities. The second one zeroes in on collecting new data on buildings: the use of real estate data to add unfilled attributes and new buildings or amenities into the existing database.

2 Background

2.1 Information on buildings and their update

Acquisition of building information, the primary focus of this paper, has been thoroughly investigated in the last decade. A variety of data sources and approaches, from satellite/aerial imagery to point clouds to street view imagery, have been used to extract information on buildings, in the form of points and footprints to semantically rich 3D building models, for diverse applications and at various scales (Bshouty et al. 2019; Xie et al. 2019; Zhang et al. 2021a; Gui & Qin 2021; Biljecki & Ito 2021; Frantz et al. 2021; Ledoux et al. 2021). Some parts of the world have benefited from these developments and an increasing number of jurisdictions is rich in data on buildings, which are often released openly. Nevertheless, in many other parts of the world, mapping buildings remains a manual task, e.g. by OpenStreetMap contributors, due to lack of input data such as high-resolution imagery, often leaving areas unmapped or partially mapped.

While the acquisition of building information has been a rapidly developing topic, and while there has been an increasing body of research focusing on developing mechanisms to update spatial databases automatically (Zhang et al. 2018; Guo et al. 2016; Cheng et al. 2008; Tian et al. 2012), update of building information also remains a challenge. Most of them rely on manual updates from cadastral data and change detection (Shi et al. 2020). In this paper, we also seek to understand whether we can take advantage of the dynamic nature of real estate data to fetch up-to-date information on buildings for the purpose of their update.

2.2 OpenStreetMap and quality control

OpenStreetMap gained considerable attention in the recent years, and it is now being used across academia, government, and companies as a reliable source of spatial data (Yan et al. 2020). In fact, in some cases, it is the only freely available source of spatial data (So & Duarte 2020). While OSM started with a focus on roads, now the community is increasingly spotlighting buildings, and in some locations they are fully mapped, together with attributes (Brovelli & Zamboni 2018; Biljecki 2020). As such, OSM ascended to support a variety of spatial analyses that require building data in academia and beyond (Westrope et al. 2014; Cerri et al. 2020; Schilling and Tränckner 2020; Braun et al. 2021; Ma et al. 2022; Zhang et al. 2022; Komadina & Mihajlovic 2022). Nevertheless, their completeness, including in developed countries, remains heterogeneous. Therefore, two lines of effort have emerged – ameliorating the data and assessing their quality. Both often require a *reference dataset*, another instance of assumedly sufficient reliability that can be freely used to either ingest it in OSM or use it to cross-check the content of OSM (Zielstra et al. 2013; Zheng & Zheng 2014; Brovelli et al. 2016; Juhász & Hochmair 2018; Zhou 2018; Witt et al. 2021; Majic et al. 2021). The same concepts apply for other instances beyond OSM.

In this study, we investigate the potential of property transactions and commercial real estate advertisement data to serve as reference data for both purposes. This aspect may be particularly interesting as another contribution in the topic of VGI quality, as using one form of VGI to assess the quality of another has not been documented much.

2.3 Applications of real estate data

Property transactions have been used routinely for various types of real estate analyses (Fesselmeyer & Seah 2018; Lee & Ooi 2018). Information obtained by scraping real estate websites, such as property listings and ads for short-term accommodation, have been used primarily for studies such as understanding patterns driving prices and analysing socio-economic distributions (Boeing & Waddell 2016; Li & Biljecki 2019; Boeing 2020; Delmelle & Nilsson 2021; Kang et al. 2020; Liang et al. 2021; Zhang et al. 2021b; Nowak & Smith 2016; Su et al. 2021). For example, You et al. (2017) estimate prices of houses from a large amount of house photos posted by real estate brokers in property websites.

Outside of the real estate focus, there are few papers making use of such data. To the extent of our knowledge, the paper most related to ours is the study of Hopf (2018) who extracted information from textual data in 8341 real estate advertisements in Switzerland to support predictive energy data analytics. Building attributes such as dwelling type, amount of rooms, dwelling level from the real estate advertisements are used for household classification. These attributes are attached with geographic coordinates of the buildings, and all real estate advertisements within a radius of 1000 m of each household address are considered. This work demonstrates how textual data in real estate ads can be exploited to extract some information on buildings. We take inspiration from the work, and considerably expand it by positioning the work in the geospatial domain and volunteered geographic information, considering property transactions for the first time (besides property listings only), increase the scope of the extracted information, investigate and elaborate a broad range of use cases (e.g. data quality control), and conduct a city-wide analysis to comprehensively investigate the potential of the approach and several aspects neglected hitherto such as the influence of the urban form.

Liu et al. (2019) analyse over 200,000 images in rental ads to study the geographical differences of interior decorations in ten major cities in United States, taking advantage of the rare opportunity that photos in ads give us an insight in homes at a large scale. In a related work using the same type of data, Rahimi et al. (2016) investigate the decor of home spaces to study the presence of geographic culture and globalisation trends. Finally, Chu et al. (2016) take advantage of floor plans, which are often included in real estate ads, to generate indoor 3D models. However, the work is not standalone as it also requires other data.

These studies demonstrate that real estate data, which is often widely and easily available, contains a rich set of information pertaining to buildings in the shape of text and photos covering a large geographical area, and therefore they may hold much potential in geospatial research. However, to the extent of our knowledge, there has been no research on using real estate data to systematically and comprehensively extract information on buildings for purposes such as geodatabase update.

3 Methodology

3.1 Study Area

Singapore is a densely populated city-state in Southeast Asia with a vibrant real estate market. Buildings in Singapore are classified into three main categories: residential, commercial, and industrial. For residential buildings, there are three subcategories: Housing & Development Board (HDB) (public housing), private housing, and hybrid housing. Buildings managed by HDB accommodate more than 80% of residents, and their quality in OSM has been deemed as very high with near full completeness (Chen 2020; Biljecki 2020).

3.2 Data

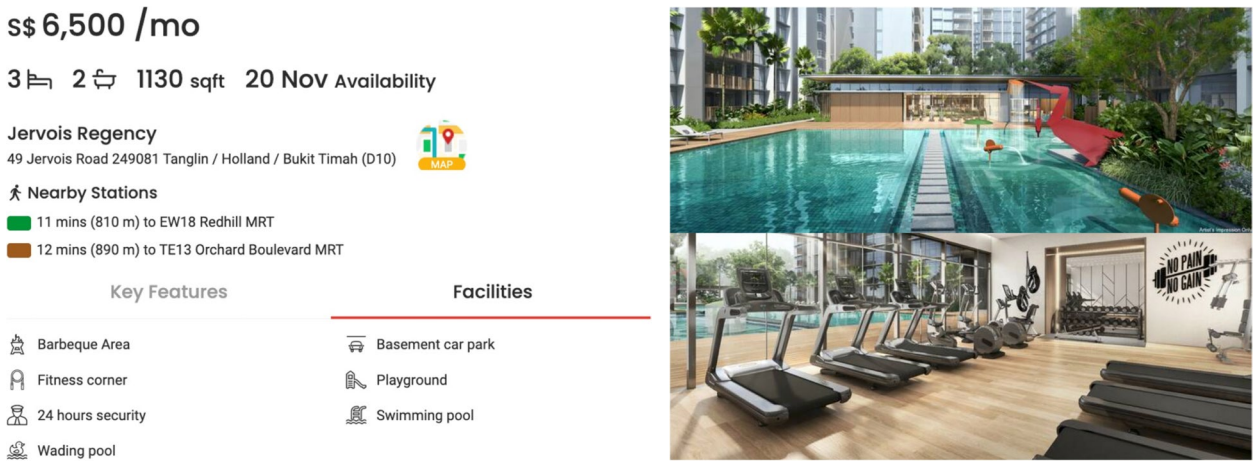
3.2.1 Data collection, cleaning, and processing

For real estate data, we use two instances: property sales transactions and real estate ads (listings). In the former, we have two datasets, which are similar but come from two different sources: transaction data of public housing are downloaded from the Singapore Government's open data portal, and transaction data of private/hybrid properties (private housing, commercial buildings, and industrial buildings) are obtained from REALIS (Real Estate Information System), which is a real estate database maintained by the Urban Redevelopment Authority (URA). Transaction data of public housing properties are available for a 20-year period, and transaction data from REALIS are of the last five years. They contain the address, price, type of flat, floor area, and storey of each transacted property (Fig. 1), and the dataset is similar to what many other governments elsewhere provide.

On the other hand, listings data, including rent and sale properties, are scraped from PropertyGuru, a popular property website in Singapore (see Figs. 1 and 2), also similar to those that are available in many other countries around the world.

The total number of collected transactions and listings will be described later. In our analysis, we will also take a subset of only data available in the last month, to understand the continuous potential of this work, e.g. collecting data on a monthly basis and understanding how many buildings can be captured with only one month worth of transactions and ads. There is a certain overlap between these sources in term of information they provide, and in the workflow we describe them together, but in the results, we consider these sources separately, to provide an interpretation of the results per each source, as not all of these might be available in other locations at the same time.

Regarding transaction data, building locations (addresses) and attributes of sold properties can be extracted directly from the datasets, thanks to the clean and structured dataset. For real estate advertisements, besides the locations of the building (available directly as coordinates), the descriptions and photos of the listings contain much information about building characteristics, but require a degree of processing. Further, unlike transactions, ads provide information about the attached amenities. In our work, we focus on sports facilities (primarily gyms and swimming pools), but the method is generalisable to further types of amenities such as playgrounds and parking lots.



(a) Text description pertaining to amenities.

(b) Photos included in the ad.

Fig. 2 Example of a listing for a property for rent. Besides the location and building characteristics, the ads contain information about associated amenities in both textual and photo forms. Source: PropertyGuru

To ensure the reliability of the mechanism, the reference data require automatic cleaning and processing to make it usable for harmonisation with a spatial database. There are three main steps of data cleaning and processing in this study: localisation, removal of duplicates, and extraction of information from text and photos.

First, the transaction data collected from authoritative resources include the address instead of geographic coordinates. Geocoding is applied to obtain the geographic coordinates of each building by using the Google Maps Platform API. For listing data, geographic coordinates are already available.

Second, we have found multiple records pertaining to the same property in all sources of data. In the transactions, this is because the same unit was sold multiple times in the timeframe of the dataset. Regarding the listings, the reason is that one or multiple real estate agents may advertise the same property concurrently. Such duplicates were detected, and only the latest record was preserved.

Third, the extraction of relevant information from the transactions and the textual portion of listings was trivial using existing approaches and requiring little processing (thus, much of the paper will be devoted to the results and discussion rather than the method). The photos in the listings were used to extract the presence of amenities in them. These are detected using computer vision (object detection) using the Google Cloud Vision API, giving reliable results without the need to develop an own prediction model, making the method accessible to researchers who have little expertise with such techniques.

3.2.2 Summary of the datasets at stages of data processing

After the steps of localisation and removal of duplicates, we have identified that rent and sale listings cover 4371 and 8865 buildings with at least one listing, respectively (Table 1). This amount represents about one tenth of the building stock of Singapore, and about half of residential buildings. Transactions covered more than 90% of public housing buildings (9316), and 14,192 private/hybrid residential buildings (as well a high share) (Table 1). 1800 commercial and industrial buildings have been identified as well (Table 1). The information extraction will be based on these cleaned datasets.

3.3 Connecting the extracted data with a spatial database (OSM)

The first step towards making use of relevant information from real estate data is to associate the extracted items with the ones in OpenStreetMap (or any other spatial database). The idea is to first identify the locations of the extracted items in the OSM, and then compare them with the existing data in OSM. For buildings, the listings are converted to data points, which indicates the locations; while for amenities, we define the nearest buildings in OSM as the approximate locations. Next, buffering and spatial intersection are applied to compare the two databases. We defined the intersected points as 'covered' points and unintersected ones as 'uncovered' points, indicating whether the extracted items can be matched. The methodology of this process for both buildings and related amenities is illustrated in Figs. 3 and 4. Following this method, the functions of the developed mechanism

Table 1 Summaries of cleaned datasets

Data		Original datasets	Geolocated datasets	Datasets without duplicates
Property listing data	Rent listings	14170	13778	4371
	Sale listings	47966	46752	8865
Transaction data	Transaction data of public housings	838477	838278	9316
	Transaction data of private/hybrid housings	108015	108015	14192
	Transaction data of commercial buildings	2907	2907	892
	Transaction data of industrial buildings	5310	5310	908

Original datasets contain complete data downloaded from the resources; *geolocated datasets* contain data with geographic coordinates converted from addresses; *datasets without duplicates* only contain the latest record of each building

are outlined in Table 2. Buildings are described first in this section.

The features that cannot be matched may indicate those that are missing from the spatial database, and thus, the process doubles as a method to detect unmapped instances, because they are overlooked (omission), have been demolished in the meantime, or because they have not yet been constructed. While from the real estate data we are not able to extract building footprint polygons (the most common geometric form of building information), the locations of buildings (as points) are sufficient for this purpose. Further, they are also sufficient for those databases in which buildings are mapped as points.

For both covered and uncovered building points, the same set of building information can be extracted. Table 3 highlights the building attributes that can be extracted from the three reference datasets based on our exploration. While our work is focusing on a particular study area, we assert that it is generalisable because similar services and records in other countries contain comparable information.

There are two kinds of methods of attributes extraction for different building information. First, attributes including building address, lease commence date (or year of construction and completion year), tenure, and building type can be extracted directly, and for which only the latest record/listing in the datasets is necessary. Second, the remaining two building attributes that can be extracted from the real estate datasets are approximate building levels and statistics of floor area. They are extracted indirectly and require multiple records/listings in a building, as the accuracy of extraction converges towards the true value. Figure 5 illustrates the approach of estimating the number of floors of a building. Among all of the records/listings of transacted units in a building, the one with the highest level corresponds to the approximate building levels. As there will be more records/listings created continuously, the approximate building levels will be updated if a unit of higher level appears over time, ultimately capturing a property at the highest floor, or otherwise one that is close to the highest floor, at least resulting in an approximate building form (i.e. distinguishing between mid-rise and high-rise buildings). The method of extracting statistics of floor area applies similar idea to the previous method. Mean, maximum and minimum floor area of units in the building can be established from all of the records/listings of units in the building.

Moving on to amenities, real estate data can be used to check the correctness and completeness of them in OSM. However, the unmapped amenities cannot be added to the database, as only their approximate location can be determined. When viewed as part of buildings, the presence of amenities, on the other hand, could be added as an attribute of buildings.

To identify amenities from real estate data, both text and image data from listings are used (Fig. 2). The texts of the listings include labels of amenities that the buildings have in their surroundings. Besides extracting amenity information from texts, nearly all ads contain also photos of amenities associated with the buildings. After the classification, each image will have a list of description labels. The image labels are then attached to the corresponding listings which contain the geographic coordinates of the buildings.

Next, to compare these amenity labels with the existing data of OSM, both building dataset and amenity dataset of OSM are used. The rationale of the approach is that if an amenity appears in the photo of an ad of an apartment of a building, they must be very close to the building, if not part of it. Thus, while we will not be able to map their exact location, we are able to use it for quality assessment purposes, i.e. verify whether there is such an amenity in the immediate vicinity of the mapped building. The process is straightforward – it consists of buffering the corresponding building and locating whether the same amenity is inside the buffer. If not, the amenity can be flagged as unmapped.

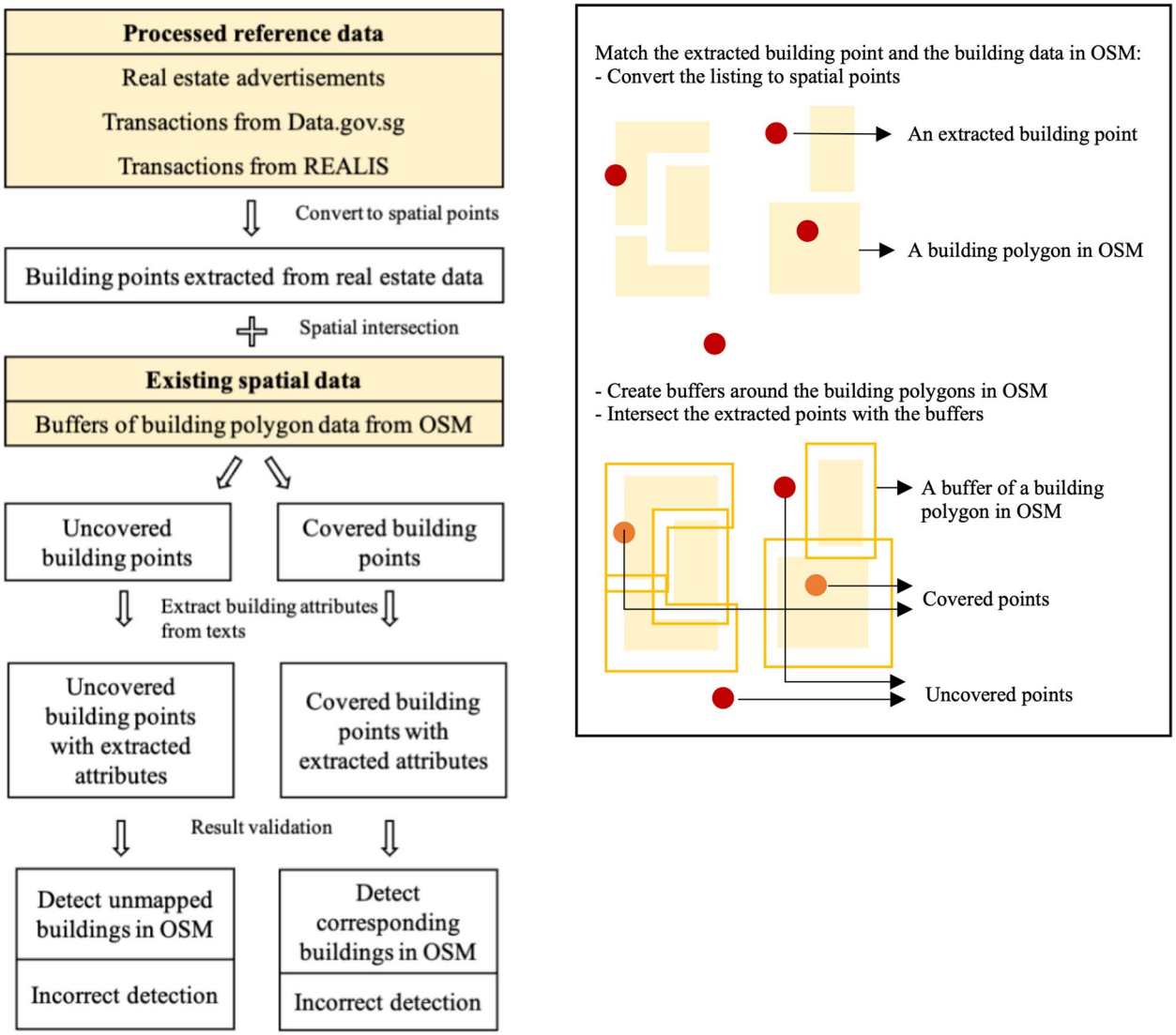


Fig. 3 Workflow (left) and spatial operations (right) of matching extracted data of buildings with a spatial database (OSM). The yellow squares in the left represent the data used in the method and implementation

3.4 Validation

To evaluate the performance of the described mechanism and the feasibility of the idea of using real estate data for such purpose, samples of the results of building and amenity databases updating are manually checked by comparing them with the ground truth (mostly satellite and street view imagery). For building database updating, two perspectives are checked: (i) if the building points from real estate data uncovered by OSM building footprints can identify the locations of unmapped buildings; and (ii) if the building points from real estate data covered by OSM building footprints can detect the corresponding buildings in OSM? For each category in three reference datasets, more than 50 samples are selected (e.g. 50 for uncovered extracted points from sale ads are checked). In total, 200, 200, and 300 samples are selected from three datasets respectively (Tables 11, 12 and 13). For amenities, the accuracy of the approximate locations of unmapped swimming pools and fitness centres are checked. The sample size is 50 for each kind of amenity. The samples are randomly selected from all over the city.

4 Results

4.1 Extraction of building information

4.1.1 Matching between real estate data and OSM data

Table 4 outlines the result of geometry matching between the three real estate datasets and OSM building data. The total number of points in the table indicates the number

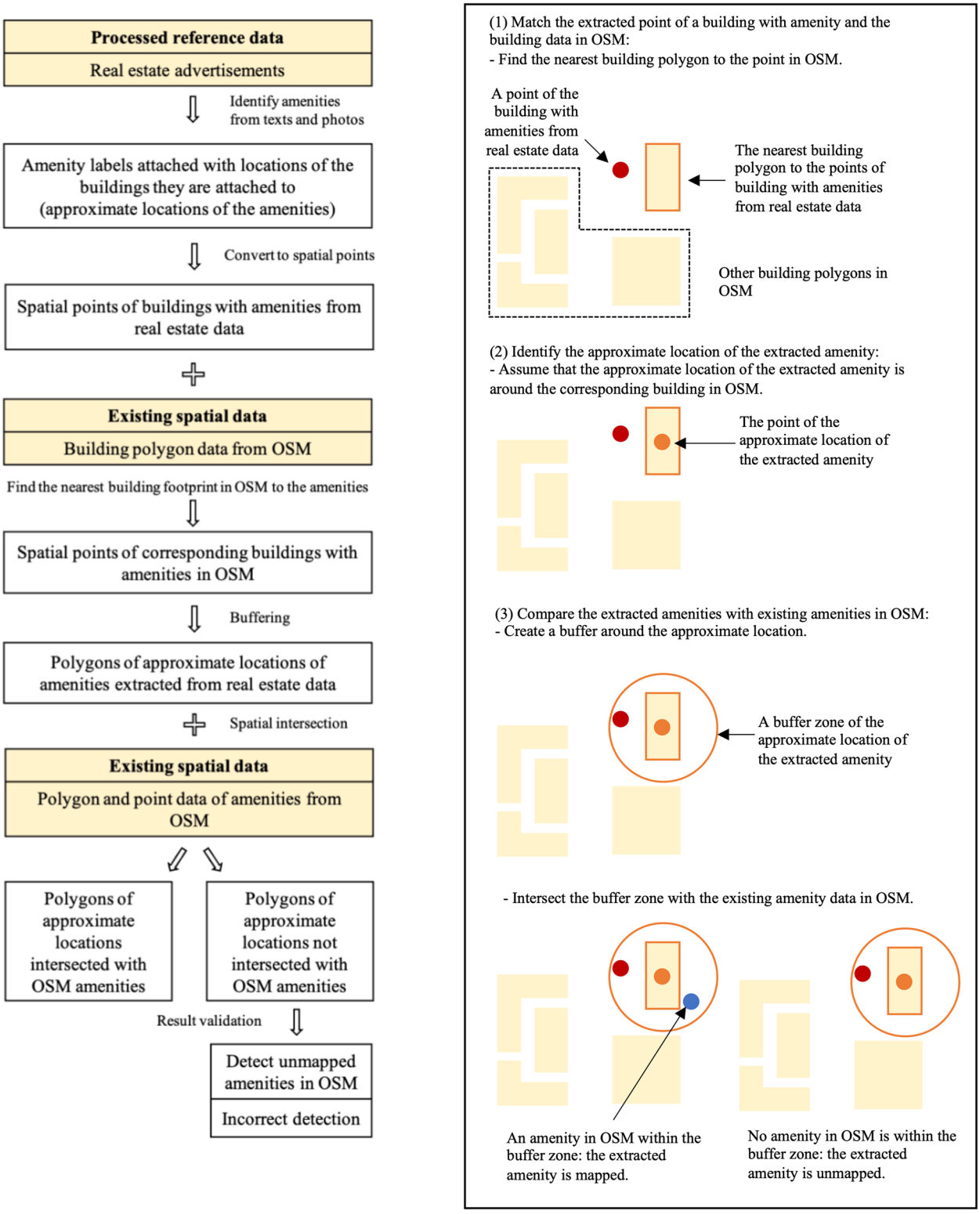


Fig. 4 Workflow (left) and spatial operations (right) of matching extracted data of amenities with a spatial database (OSM). The yellow squares in the left represent the data used in the method and implementation

Table 2 Functions of the developed approach

		Quality check	Data updating
Buildings	Locations	✓	✓
	Attributes	✓	✓
Amenities	Locations	✓	✗

Note: ✓ indicates that point locations of unmapped buildings can be identified, but they may not be always added into the database, e.g. if the database models buildings as footprints or 3D models, as there is no information of building shapes in the real estate data, thus, point-based building data may only provide a hint of the location of the building, which is sufficient for assessing completeness and updating attributes

of ads, and afterwards the number of buildings detected in the processed datasets. The table also includes the number of matched and unmatched buildings, suggesting potentially unmapped buildings or those that are yet to be constructed, and thus, are missing in OSM. A majority of buildings could be matched to counterparts in OSM, likely due to the high completeness of OSM in the study area. Nevertheless, these results suggest the performance of the method also in areas that are not mapped well in the considered spatial database such as OSM, as we determined that about one tenth of the building stock could be inferred

Table 3 Building attributes extracted from real estate reference data

	Transactions		Listings
	Public housing	Private/hybrid	
Directly	• Lease commence date • Tenure • Building type	• Building name • Lease commence date • Completion year • Tenure • Building type	• Completion year • Tenure • Building type
Indirectly	• Approximate building levels • Min / Mean / Max floor area	• Approximate building levels • Min / Mean / Max floor area	

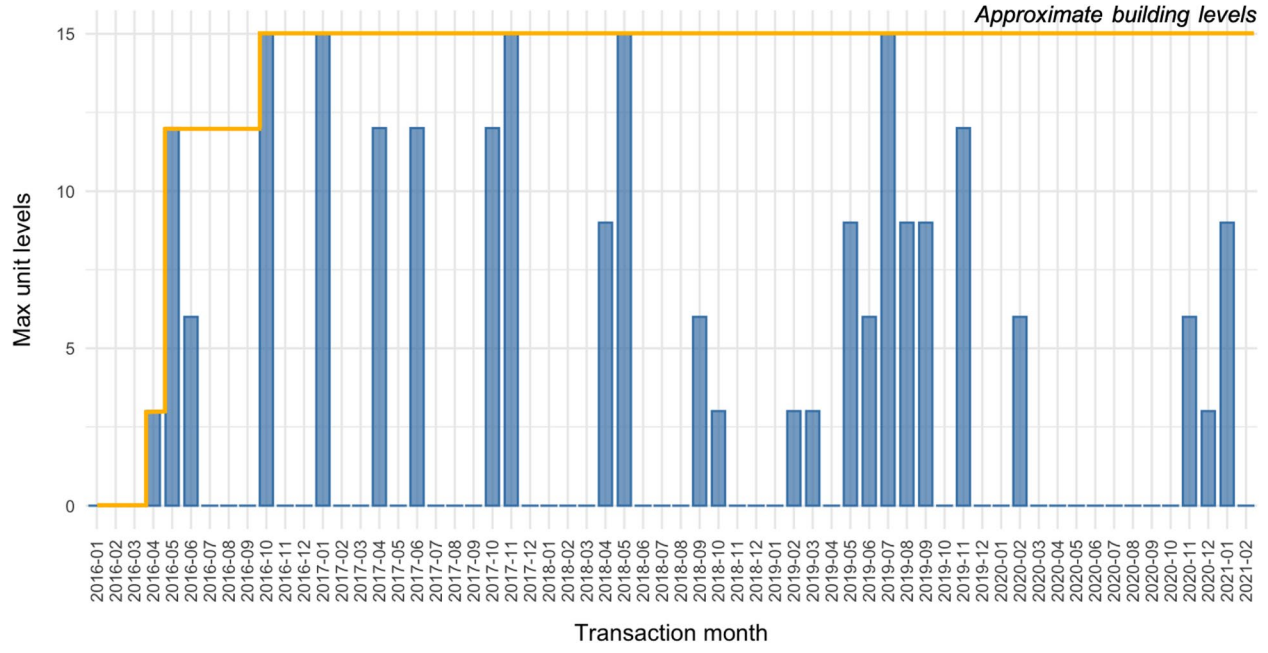


Fig. 5 Indirect method of estimating building levels. This example is based on the transaction data of a high-rise residential building (Blk 702 West Coast Rd, Singapore). It shows the maximum unit levels that appear in the transaction data in each month in a five-year period and their convergence to the true value over time. The approximate level of this building is the highest unit level that appears in the transaction records. In this example, as it is the case in many other instances, the method identified the building level correctly, and about one year of transactions is usually enough to obtain accurate information on the vertical extent of the building, which provides indispensable value to some use cases

Table 4 The results of matching buildings extracted from real estate data with those in OpenStreetMap

		Listings		Transactions			
				Public housing	Private/hybrid		
		For rent	For sale	R	R	C	I
Properties		13,778	46,752	838,278	108,015	2907	5310
Buildings		4371	8865	9316	14,192	892	908
Matched buildings	#	3411	6019	8765	10,784	788	753
	%	**78**	**68**	**94**	**76**	**88**	**83**
Unmatched buildings	#	960	2846	551	3408	104	155
	%	**22**	**32**	**6**	**24**	**12**	**17**

Number of building footprints in OSM: 115,305

Notes: *R* residential, *C* commercial, *I* Industrial. The building completeness of OSM in Singapore is high, thus, the number of footprints can be used as a ballpark figure of the building stock and to put the numbers above in perspective

Table 5 Samples of building information extracted from rent listings, which have been matched to counterparts in OSM, and can be used to update the existing record with previously unavailable attributes

Location	Address	Year	Type	Tenure
1.29882, 103.88493	33 Fort Road	2016	Residential (non-landed, private)	Freehold
1.37270, 103.85723	774 Bedok Reservoir View	1979	Residential (public housing)	99-year Leasehold

Table 6 Summary of extracted information from rent and sale listings in a period of one month

	Attributes		Location	Address	Year	Type	Tenure
Covered (8115)	Rent (3032)	Number of missing values	0	16	513	0	401
		Number of extracted values	3032	3016	2519	3032	2631
	Sale (5083)	Number of missing values	0	72	1551	0	1453
		Number of extracted values	5083	5011	3532	5083	3630
Uncovered (3806)	Rent (960)	Number of missing values	0	12	381	0	334
		Number of extracted values	960	948	579	960	626
	Sale (2846)	Number of missing values	0	97	1897	0	1853
		Number of extracted values	2846	2749	949	2846	993

from real estate data, a value that is useful for any scenario discussed in the paper.

4.1.2 Building information extracted from real estate advertisements

For both rent and sale listings, four building attributes can be extracted, and a sample of the extraction is given in Table 5 as an example. The method proves to be useful to borrow such attributes and enrich existing building records.

To put the results in perspective, and pronounce the idea of periodically or continuously scraping real estate websites to mine data on buildings that can be used to enrich spatial databases, we focus only on the last one month of scraped data to understand how much information can be extracted from solely one month of listings. The results in Table 6 suggest that listings posted in a span of one month can update the attributes of 8115 buildings (around 7% of the city's building stock) and identify 3806 uncovered buildings (around 3.3% of the city's building stock). However, it should be noted that listings do not always contain all these attributes, which we indicate in the same table. The unmatched buildings will be discussed in later sections.

4.1.3 Building information extracted from public housing transaction data

There are nine building attributes that can be extracted from HDB transactions data (Table 7). The data have a

high degree of matching with OSM, and the attributes extracted from it do not suffer from any missing value, unsurprisingly as the source of the transactions is from the government. Table 7 indicates that HDB transactions that have occurred in a single month can update the attributes of 1535 buildings, and identified 101 unmapped buildings.

4.1.4 Building information extracted from transaction data from private/hybrid properties

For hybrid/private housing transaction data, ten attributes can be extracted (Table 8). Building name, completion year, lease commence date, and approximate level are the four attributes with most missing values, but nevertheless, the majority of transactions contains such information. Interpreting the results, private/hybrid housing transactions generated in one month can update the attributes of 721 buildings (around 0.6% of the building stock), and it identified 450 unmapped buildings. The results for commercial and industrial buildings are congruent, and are not included here for space considerations.

4.2 Extraction of amenity information

Fitness centres and swimming pools are extracted from both text descriptions and photos in the listings: 3051 unique fitness centres and 2176 unique swimming pools are extracted from texts, and 99 unique fitness centres and 1452 unique swimming pools are extracted from photos. After removing the duplicates between amenities from texts and photos, 3096 fitness centres and 2947 swimming pools remain.

After extracting the amenities from listing data, we converted these listings with amenities into spatial points. To compare them with the amenity data of OSM, the nearest building in OSM to each of the spatial point is identified (Fig. 4). This is because the location of each spatial point indicates the location of the building nearby the amenity rather than of the amenity itself. Finding the nearest buildings in OSM is detecting the locations of corresponding buildings with amenities extracted from ads. We assumed that the locations of the nearest buildings are the approximate locations of extracted amenities.

In some cases, the amenities are fairly far (over 100 metres) from their nearest buildings, which is uncommon. This might be because the building locations in the listings are inaccurate or the buildings in the listings have not been mapped in OSM yet. To avoid the influence of these cases and identify the corresponding buildings, amenities with distances longer than a certain threshold to their nearest buildings are removed. To decide on the threshold, results with various distances have been checked. When the distances are shorter than 25 metres, in most of the cases the nearest building is the corresponding building which pertains to the amenity, while when distances are as large as around 30 and 40 metres, usually the nearest buildings are not the ones where the amenities are attached. Besides, according to the results, 85% of fitness centres and 88% of swimming pools have a distance shorter than 25 metres to their nearest buildings. Hence, to ensure the accuracy and completeness of the results, amenities with distances longer than 25 metres from the building are removed.

After processing, there are 2623 fitness centres and 2600 swimming pools extracted from listing data that remain. Besides, there are 2512 and 2272 unique buildings corresponding to the fitness centres and swimming pools respectively. These numbers are less than the numbers of amenities, which means some amenities extracted by listings are close to each other and share the same nearest buildings.

After detecting the approximate locations of extracted amenities in OSM, buffering and spatial intersection are applied to identify if the extracted amenities are already mapped in OSM (Fig. 4). To achieve a higher accuracy of the comparison, the buffer distance should be long enough to cover the attached amenities of the buildings but should not be too long to cover other amenities. Hence, quantiles of distances between amenities and their nearest neighbours and amenities and their nearest buildings in OSM are calculated. In this study, 20 metres are selected as the buffer distance. Because for both fitness centre and swimming pool, distances of over 75% amenities to their nearest amenities is larger than 20 metres and distances of less than 25% amenities to their nearest buildings is longer than 20 metres.

After identifying the buffer distances, intersections are applied between the buffers of the buildings with amenities and the existing amenities of OSM. Table 9 shows that 99% of the fitness centres in the approximate locations are not mapped in OSM yet, while 19% of the swimming pools are already mapped in OSM.

While this method cannot detect the exact coordinates and shapes, the results suggest that it can signal omission issues in the database with a high degree of reliability and approximate expected location of the omitted feature (Fig. 6).

4.3 Validation of the results

4.3.1 Potential of updating and validating a spatial database of buildings

Table 10 illustrates a few samples of the validation. The table also exposes some issues in real estate data, which affect the performance of the method. Tables 11, 12 and 13 outlines the performance of the method for the validation set.

In general, the method seems to be successful for using real estate data to enrich existing building data

Table 7 Summary of extracted information from HDB transactions in a period of one month. For all records, all attributes are extracted

	Attributes	Loc	Addr	Lease commence date	Type	Tenure	Approx levels	Mean floor area	Max floor area	Min floor area
Cov (1535)	Number of extracted values	1535	1535	1535	1535	1535	1535	1535	1535	1535
Uncov (101)	Number of extracted values	101	101	101	101	101	101	101	101	101

Note: *COV* covered, *UNCOV* uncovered, *LOC* location, *ADDR* address

Table 8 Summary of extracted information from private/hybrid housing transaction data in a period of one month

	Attributes	Loc	Addr	Name	Year	Tenure	Lease commence date	Approx levels	Type	Mean floor area	Max floor area	Min floor area
Cov (721)	Number of missing values	0	0	72	58	0	359	286	0	8	8	8
	Number of extracted values	721	721	649	663	721	362	435	721	713	713	713
Uncov (450)	Number of missing values	0	0	25	16	0	149	82	0	8	8	8
	Number of extracted values	450	450	425	434	450	301	368	450	442	442	442

Note: *COV* covered, *UNCOV* uncovered, *LOC* location, *ADDR* address

Table 9 Comparison between extracted amenities and existing amenities data of OSM

	Fitness centre		Swimming pool	
	#	%	#	%
Locations of mapped amenities	15	0.73	436	19.19
Locations of unmapped amenities to add	2052	99.27	1836	80.81
Total number of approximate locations of amenities	2067		2272	

with previously unavailable attributes, but it is less so in detecting unmapped buildings. The validation set suggests that for building points from real estate data uncovered by building polygons of OSM, most of them are unmatched because of the inaccurate building locations in real estate data (case exhibited in Table 10(a)). For uncovered points of residential building, no more than 15% of them can detect unmapped buildings in OSM (Table 10(b)), but this low figure does not necessarily suggest the limited performance of the method, as it may rather reflect that there are simply not many buildings that are left unmapped in the study area. For commercial and industrial buildings, the percentages are slightly higher (18% and 34%). This difference might be because that there are more unmapped commercial and industrial buildings than residential buildings in OSM in the study area. It is worth noting that, while the method in some instances is able to detect buildings that are missing from the targetted spatial database, some buildings from older transactions can represent demolished building rather than those that are unmapped or yet to be constructed (Table 10(c)), which may be both an advantage or disadvantage, depending on the perspective: on the one hand, it may be possible to reconstruct historical data, while on the other hand, this may be undesirable if only the current or future situations are sought. Whether a building is unmapped because it was demolished or because it is unbuilt could be distinguished by checking the year of completion of the building, which is usually available in both the ads and transactions.

These results should also be placed in the context of the very high building completeness of OSM data in Singapore. Applying the method in areas of partial and heterogeneous completeness may result in detecting many more unmapped buildings. Thus, the high performance of matching buildings in the real estate data with their counterparts in OSM (or any other spatial database), can be interpreted also as the method having high potential of detecting missing buildings.

For buildings points covered by building polygons in OSM, the accuracy rate of detecting the corresponding buildings in OSM (Table 10(d)) is fairly high. Authoritative real estate data including transactions can detect over 94% of OSM buildings correctly (Tables 12 and 13), and for these buildings a bridge can be established to transfer the semantic information from one to the other dataset. The building locations in real estate advertisements are less reliable but can still achieve an accuracy rate of 86% and 88% for rent and sale listings, respectively (Table 11), suggesting a high potential as a rapid acquisition method for areas lacking completeness.

Among the real estate records that can be correctly associated to buildings in OSM, there are some points covered by multiple building polygons. This is because these building footprints are clustered together, and the buffers of them overlap with each other. Tables 12 and 13 indicate that public housing and industrial buildings have much less overlap than private/hybrid housings and commercial buildings, caused by the urban form of our study area — some properties (e.g. terraced, semi-detached houses) and shophouses are close to each other.

4.3.2 Potential of validating a spatial database of amenities

The performance of the method for collecting data on amenities is outlined in Table 14. The method proves to be highly effective: among the samples of unmapped swimming pools and fitness centres, 90% and 76% of them are detected correctly. The inaccuracies of the mechanism include two causes. One is that the approximate locations of the amenities are correct, but they are already mapped. The other one is that the mechanism does not detect the approximate location accurately, and there is no amenity. Most of the inaccuracies pertain to the second situation, which are caused by the errors of the locations in real estate data.

In conclusion, the strongest point of the method is the enrichment of spatial databases with descriptive information of buildings, highly valuable for a variety of use cases, and the detection of unmapped amenities associated with buildings. In theory, the method holds value for detecting unmapped buildings, but its performance in practice is burdened by imperfect real estate data and it depends on the level of completeness of existing data.

5 Discussion

5.1 Observations, limitations, and opportunities for further investigations

The results uncover the high potential of real estate data to serve spatial data infrastructures and affirm the role

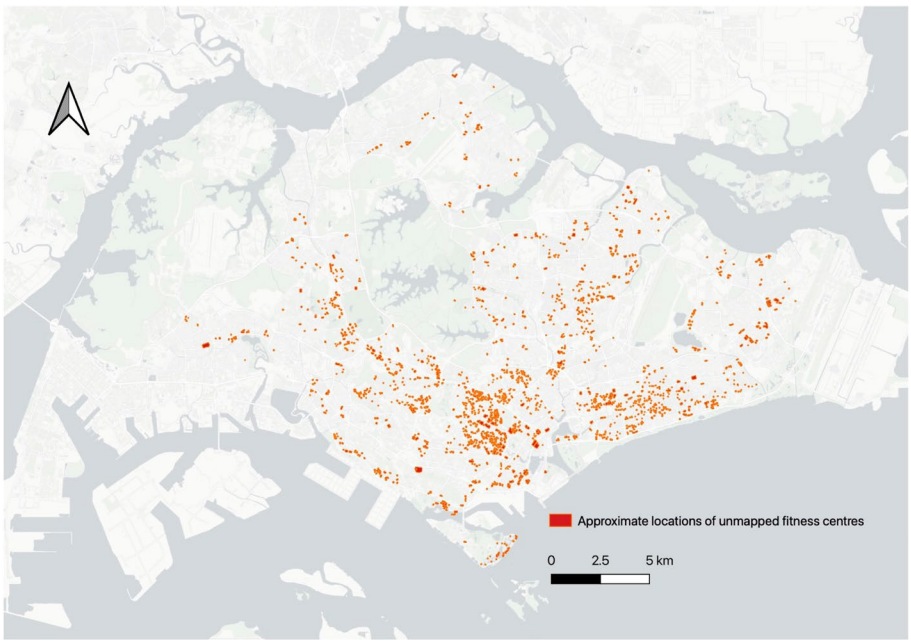


Fig. 6 Approximate locations of unmapped fitness centres detected thanks to our method. Basemap: OpenStreetMap contributors

of such data as a latent form of implicit VGI. In a relative simple way, real estate data offers filling gaps in the data that otherwise are not easily obtainable, and alleaviates the challenging problem of acquiring spatial data of fine and dynamic urban features. Considering that such approach has been engaged for the first time, there are limitations and research opportunities.

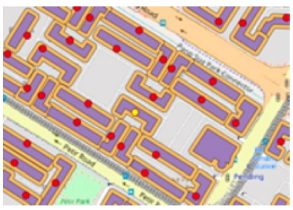
In the development of the method, we have regarded its applicability in other study areas. Real estate markets are highly digital in many parts of the world, and datasets such as listings and transactions are quite similar in content and structure, so for its large part, the method can be applied elsewhere with little modifications. Nevertheless, a limitation of the work is that a specific study area was investigated, which may have its own particularities and the results will inevitably differ elsewhere. In the next section, we will discuss applications and scalability of the method in other study areas. Another limitation of the work is related to transactions — they suffer from *survivorship bias*, as transaction data contains only the apartments that have been sold successfully, not all that are on the market.

Next, in our work, we use a mix of authoritative data (property transactions) and user-generated data (real estate ads). The first one is used routinely in assessing the quality of VGI. The second one, however, is not, and essentially, in this paper, we are using one form of VGI to check another one. Further, this instance of VGI can also be used to check authoritative datasets, another uncommon exercise, but a potentially viable novelty and contribution (e.g. using listings to detect whether the function of a building remains the same as recorded in the cadastral registry). Such approaches would be interesting to consider for further investigations.

Regarding amenities, a small set of amenities was considered (gyms and swimming pools) that are specific to the study area, and with limited coverage (only those that are in the immediate vicinity of the property that is sold could have been captured), and we could only use them for quality assessment rather than mapping them (since their exact location is not known). In this work, the listings are predominantly advertising residential real estate, so we focused on amenities that pertain to such properties. However, these amenities are not common in many other countries, and there are many more amenities that are relevant in GIS and urban studies that could be investigated in the future, e.g. parking lots. Further, some amenities may be more characteristic to commercial properties. Perhaps an ad on selling commercial space may include a photo of a restaurant in the same building as an added value of the property. Next, we noticed that some ads feature amenities in the wider catchment area of the building, e.g. shopping malls and public transportation in the neighbourhood. While detecting these is not a challenge, because of the much larger radius that is entailed, it may be meaningless to include them in this work.

In terms of amenities, another limitation is that many of these have restricted access, thus, their benefit of

Table 10 Examples of matching the extracted real estate data for updating and/or validating a building database. The OSM data we have used is denoted in purple, while the basemap is also from OSM, but from a few months later with some unmapped buildings added in the meantime

	OpenStreetMap	Satellite image	Descriptions
(a)			The building point is **uncovered** by the building polygon of OSM. However, real estate data contains a positional error.
(b)			The building point is **uncovered** by the building polygon of OSM. The building point can indicate the location of an unmapped building, which in the meantime has been mapped in OSM.
(c)			The building point is **uncovered** by the building polygon of OSM. The building point can indicate the location of a building which is yet to be built or one that has been demolished.
(d)			The building point is **covered** by the building polygon of OSM. The corresponding building in OSM is detected, thus, it can be used to fill the missing attributes in OSM.

Note: is the inferred building location. is the selected building point. is the unselected building point. is the building footprint in OSM. is the buffer of the building footprint in OSM we have used to match buildings (7m). Source of basemaps: OpenStreetMap contributors and Google Maps

Table 11 Potential of updating and validating a spatial database of buildings: using real estate advertisements (listings)

		Uncovered			Covered			
		Detect unmapped buildings	Location error	Total	Correct detection		Incorrect detection	Total
					Overlap	No overlap		
Rent	#	4	46	50	11	32	7	50
	%	8%	92%	100%	22%	64%	14%	100%
Sale	#	2	48	50	5	39	6	50
	%	4%	96%	100%	10%	78%	12%	100%

being mapped remains dependent on a downstream application. While a substantial portion of Singapore's housing landscape is public, with buildings and their surroundings accessible without restrictions, it is mostly the private properties (e.g. Fig. 1) that include amenities that we have detected in this work. The difference and importance between a gym in a condominium and a private gym available to the public at a fee depends on the mapping and application context. In our case, we noticed that both tend to be mapped in OSM, but it can be argued that the relevance of the former may not be at the same level as the one of the latter chiefly due

Table 12 Potential of updating and validating a spatial database of buildings: using public housing transactions

	Uncovered			Covered			
	Detect unmapped buildings	Location error	Total	Correct detection		Incorrect detection	Total
				Overlap	No overlap		
#	14	86	100	2	98	0	100
%	14%	86%	100%	2%	98%	0%	100%

Table 13 Potential of updating and validating a spatial database of buildings: using private property transactions

		Uncovered			Covered			
		Detect unmapped buildings	Location error	Total	Correct detection		Incorrect detection	Total
					Overlap	No overlap		
Residential	#	1	49	50	15	34	1	50
	%	2%	98%	100%	30%	68%	2%	100%
Commercial	#	17	33	50	24	23	3	50
	%	34%	66%	100%	48%	46%	6%	100%
Industrial	#	9	41	50	4	45	1	50
	%	18%	82%	100%	8%	90%	2%	100%

Table 14 Result validation of amenity database updating

		Detecting unmapped amenities	Detecting correct amenity locations that are already mapped	No amenity at the locations	Total
Swimming pool	#	45	1	4	50
	%	90%	2%	8%	100%
Fitness centre	#	38	0	12	50
	%	76%	0%	24%	100%

to public access. At the same time, it might be very useful to map features such as automated external defibrillator units in private areas despite their restricted access. Therefore, this is not necessarily a limitation, but it may certainly result in an imbalance in mapping, which on the other hand, also reflects the state of OSM. However, there are other types of amenities that may be possible to detect, such as childcare and parks, which are public but integrated in residential estates, and thus, such limitation will not apply. Nevertheless, amenities are the secondary purpose of this work, and it does not affect the main goal — extracting data on buildings.

As another direction for future work, we suggest developing a measure of 'confidence' of the results that would encapsulate the accuracy of the derived data. For example, the reliability of the method on deriving the number of building levels (Fig. 5) much depends on the number of transactions and time, aspects that may be taken into account by such a metric.

5.2 Applications elsewhere

5.2.1 Influence of the urban form

The mechanism in this study is built based on the buildings and amenities in Singapore, which is a city that has experienced intense urban development and has high building density (Fig. 7). As much as possible, we investigated this idea in general with scalability in mind, but inevitably, the results are tied to the study area and the performance of the method elsewhere remains to be investigated.

Many landed properties (private housing, terraced houses) and shophouses (historic commercial buildings) in Singapore are standing shoulder to shoulder, and minor building location inaccuracies in real estate data can cause errors in identifying corresponding buildings in OSM, and overlapping buffers may cause mismatching the derived information. For high-rise buildings, which are usually well separated from each other, the mechanism is more effective in both detecting unmapped buildings and finding corresponding existing buildings in

(a) Landed properties (b) Shophouses (c) High-rise public housing

Fig. 7 Variable urban form in Singapore. The performance of this method is driven by urban morphology. The photos are courtesy of Unsplash contributors

OSM (Tables 11, 12 and 13). Hence, the approach might be more reliable in regions with lower building density and those that have detached real estate, but at the same time the high-rise urban form ensures that there are many data points for a single building, maximising the data acquisition. A convenient particularity of our study area is that for most buildings, at any moment, there is at least one apartment being advertised for rent or sale, which is sufficient for our method to work and collect data on the entire building.

Another influence of the urban form on the mechanism is about the estimation of approximate building levels, which are calculated as the highest levels of the units in one building appearing in all of the records/listings according to the mechanism. Urban areas mainly consisting of low-rise houses rather than high rises might not be very suitable to use this method. Because there might not be many records/listings of the buildings to estimate their total levels. At the same time, a small number of listings in a building is precisely something that could indicate its size as well, i.e. it can also serve as a proxy for its height.

5.2.2 Reference data

This study used multiple sets of real estate data that is adopted as reference data — advertisements and property sale transactions from both commercial and authoritative sources. In many countries, there are property websites providing abundant rent or sale listings containing the same building information, for example, Redfin in the US, Rightmove in the UK, Funda in the Netherlands, and Beike in China, ensuring replicability elsewhere. However, the authoritative transaction data might not be available in as many countries. In our analysis, we have considered each source of data in isolation to give an understanding of the results when only one of them is available.

5.3 Legal matters

Web scraping is an important step to collect the listings and transform them into data useful for our approach. While much of the method relies on data such as transactions, web scraped listings add much value because they contain additional data such as on amenities and may indicate buildings that will be constructed in the future, something rarely available in conventional spatial data sources. However, the legality of web scraping is still in a 'grey area' (Krotov et al. 2020), despite its abundant use in academia and elsewhere. It might constitute copyright infringement or a breach of contract of the website's terms of use, but it is also a question whether an online marketplace actually owns copyright on ads posted by its users[1]. In the context of this study, the Copyright Act 2021 of Singapore permits copying copyrighted work specifically for the purpose of computational data analysis[2], and it appears that the online marketplace does not prohibit web scraping for non-commercial use, which may be the case in many other countries.

Another legal concern is that the data that is added to certain spatial resources such as OSM should not be from any copyrighted data sources. Hence, only real estate data sources with licenses compatible with the targetted database (i.e. OSM) can be used for the mechanism. It remains to be investigated whether these issues would affect use cases for research purposes with data that is not distributed. Nevertheless, in the context of implicit VGI, these aspects of real estate data are not necessarily much different from other VGI instances such as Flickr, which are well established in the community.

[1] https://singaporelegaladvice.com/law-articles/legal-scrape-crawl-websites-data-singapore/. Last accessed: 2022-10-07

[2] https://www.mlaw.gov.sg/news/press-releases/2021-11-19-commencement-of-copyright-act. Last accessed: 2022-10-07

6 Conclusion

Real estate data sources provide a wealth of information that have been underused outside their primary purposes in the real estate domain, and may be used to augment their value by extending their application in other domains, transcending their value beyond their primary purposes. In this paper, we bring attention to the spatial aspect of various forms of real estate data, and we put forward a new idea — exploiting real estate data as an unexplored and underused form of crowdsourced data and volunteered geographic information, which among other applications, can help to provide a rapid and efficient method to keep spatial databases updated and correct. We provided a proof of concept — we have demonstrated that data obtained from real estate ads and property transactions may be a new source for collecting building data for the geospatial domain. The method is simple and powerful, and we demonstrate its feasibility by performing an experiment in Singapore. The results suggest that we are able to retrieve several sets of building information for a large number of buildings in the country, many of which have not been available in OSM in the study area (and are rarely available elsewhere), and are key ingredients in a breadth of spatial analyses that require semantic data on buildings, such as in planning, and in energy and microclimate simulations. By considering real estate data as reference (ground truth) data, the method doubles as an instrument to carry out data quality assessment studies, and our paper essentially contributes to the field by introducing a new method for spatial data quality assurance. While this method will unlikely be able to provide information on all buildings in a study area (some buildings will take time to be advertised or transacted, if ever), its heterogeneity is in line with other forms of VGI such as OpenStreetMap and Mapillary (Juhász & Hochmair 2016; Quinn & León 2019), and the amount of buildings that is enriched with this method may be considered as a significant advancement given that this data source was overlooked hitherto. In fact, our implementation suggests that just one month of listings uncovers information of a large portion of a city's building stock, which is ahead of some other acquisition techniques. New buildings, which are yet to be constructed, can also be detected, but much of the method is burdened by errors in the listings. Future buildings can be identified also from property transactions by extracting those with completion year in the future, and the same goes for historical data of demolished ones, possibly allowing analyses of the evolution of the urban form.

In the work, we have taken advantage of amenities that are being advertised as part of real estate. The mechanism is reliable in identifying the approximate locations of amenities missing in databases, providing both a mechanism to assess completeness and one that signals omissions in the database to mappers and other parties managing a spatial source.

As this work presents a new channel to collect spatial data, and it argues that real estate may be a form of user-generated geospatial data that was previously not considered in the field, it provides prospects for multiple directions for future work. For example, it would be beneficial to investigate whether we could use photographs provided in listings for further applications, such as 3D model reconstruction and to infer footprints. While points extracted from real estate data have been sufficient to serve the purpose of this research, reconstructing footprints may bring the research forward. Next, regarding implementation, it would be useful to provide a system that would continuously scrape ads and download property transactions for always-on sensing of building information and translating it to a building database.

Acknowledgements
The authors appreciate the input data used in this research. The help of Jonas Teuwen and Yingwei Yan are gratefully acknowledged.

Authors' contributions
Xinyu Chen — Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Visualization, Project administration; Filip Biljecki — Conceptualization, Investigation, Writing - Review & Editing, Supervision, Project administration, Funding acquisition. Both authors read and approved the final manuscript.

Funding
This research is part of the project Large-scale 3D Geospatial Data for Urban Analytics, which is supported by the National University of Singapore under the Start Up Grant R-295-000-171-133 and the Google Cloud Academic Research Grant.

Availability of data and materials
The data on public housing transactions is available openly at https://data.gov.sg/.

Declarations

Ethics approval and consent to participate
Not applicable. No human participants have been part of the study.

Consent for publication
All authors agreed with the content and gave explicit consent to submit.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Geography, National University of Singapore, Singapore, Singapore. [2]Department of Architecture, National University of Singapore, Singapore, Singapore. [3]Department of Real Estate, National University of Singapore, Singapore, Singapore.

Received: 1 August 2022 Revised: 10 October 2022 Accepted: 28 October 2022
Published online: 23 November 2022

References

Abdelrahman, M. M., Zhan, S., Miller, C., & Chong, A. (2021). Data science for building energy efficiency: A comprehensive text-mining driven review of scientific literature. *Energy and Buildings, 242*, 110885. https://doi.org/10.1016/j.enbuild.2021.110885

Agugiaro, G., González, F., & Cavallo, R. (2020). The city of tomorrow from... the data of today. *ISPRS International Journal of Geo-Information, 9*(9), 554. https://doi.org/10.3390/ijgi9090554

Biljecki, F. (2020). Exploration of open data in Southeast Asia to generate 3D building models. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, VI–4/W1–2020*, 37–44. https://doi.org/10.5194/isprs-annals-vi-4-w1-2020-37-2020

Biljecki, F., & Ito, K. (2021). Street view imagery in urban analytics and GIS: A review. *Landscape and Urban Planning, 215*, 104217. https://doi.org/10.1016/j.landurbplan.2021.104217

Boeing, G. (2020). Online rental housing market representation and the digital reproduction of urban inequality. *Environment and Planning A: Economy and Space, 52*(2), 449–468. https://doi.org/10.1177/0308518x19869678

Boeing, G. & Waddell, P. (2016). New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings. *Journal of Planning Education and Research*. https://doi.org/10.1177/0739456X16664789

Bourdeau, M., Qiang Zhai, X., Nefzaoui, E., Guo, X., & Chatellier, P. (2019). Modeling and forecasting building energy consumption: A review of data-driven techniques. *Sustainable Cities and Society, 48*, 101533. https://doi.org/10.1016/j.scs.2019.101533

Braun, R., Padsala, R., Malmir, T., Mohammadi, S., & Eicker, U. (2021). Using 3D CityGML for the Modeling of the Food Waste and Wastewater Generation–A Case Study for the City of Montréal. *Frontiers in Big Data, 4*, 662011. https://doi.org/10.3389/fdata.2021.662011

Brovelli, M. A., Minghini, M., Molinari, M., & Mooney, P. (2016). Towards an automated comparison of OpenStreetMap with authoritative road datasets. *Transactions in GIS, 21*(2), 191–206. https://doi.org/10.1111/tgis.12182

Brovelli, M. A., & Zamboni, G. (2018). A New Method for the Assessment of Spatial Accuracy and Completeness of OpenStreetMap Building Footprints. *ISPRS International Journal of Geo-Information, 7*(8), 289. https://doi.org/10.3390/ijgi7080289

Bshouty, E., Shafir, A., & Dalyot, S. (2019). *Towards the generation of 3D OpenStreetMap building models from single contributed photographs* (p. 101421). Environment and Urban Systems: Computers. https://doi.org/10.1016/j.compenvurbsys.2019.101421

Cerri, M., Steinhausen, M., Kreibich, H., & Schröter, K. (2020). Are OpenStreetMap building data useful for flood vulnerability modelling? *Natural Hazards and Earth System Sciences, 21*(2), 643–662. https://doi.org/10.5194/nhess-21-643-2021

Chen, H.-C., Han, Q., & de Vries, B. (2020). Urban morphology indicator analyzes for urban energy modeling. *Sustainable Cities and Society, 52*, 101863. https://doi.org/10.1016/j.scs.2019.101863

Chen, J., Stouffs, R., & Biljecki, F. (2021). Hierarchical (Multi-Label) Architectural Image Recognition and Classification. In *Proceedings of the 26th International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA) 2021*, 161–170

Chen, N., Zhang, Y., Du, W., Li, Y., Chen, M., & Zheng, X. (2021). KE-CNN: A new social sensing method for extracting geographical attributes from text semantic features and its application in wuhan, china. *Computers, Environment and Urban Systems, 88*, 101629. https://doi.org/10.1016/j.compenvurbsys.2021.101629

Chen, W. H. (2020). Assessing the quality of OpenStreetMap building data in Singapore. Master's thesis, National University of Singapore

Cheng, H., Li, Y., & Lin, Y. (2008). A study on multi-agent spatial database update mechanism based on Wiki idea. In *Geoinformatics 2008 and Joint Conference on GIS and Built Environment: Geo-Simulation and Virtual GIS Environments, volume 7143* (pp. 71431K). International Society for Optics and Photonics. https://doi.org/10.1117/12.812585

Chu, H., Wang, S., Urtasun, R., & Fidler, S. (2016). HouseCraft: Building houses from rental ads and street views. In *Computer Vision – ECCV 2016*, (pp. 500–516). Springer International Publishing

Craglia, M., Ostermann, F., & Spinsanti, L. (2012). Digital Earth from vision to practice: making sense of citizen-generated content. *International Journal of Digital Earth, 5*(5), 398–416. https://doi.org/10.1080/17538947.2012.712273

Delmelle, E. C., & Nilsson, I. (2021). The language of neighborhoods: A predictive-analytical framework based on property advertisement text and mortgage lending data. *Computers, Environment and Urban Systems, 88*, 101658. https://doi.org/10.1016/j.compenvurbsys.2021.101658

Feng, X., & Humphreys, B. R. (2012). The impact of professional sports facilities on housing values: Evidence from census block group data. *City, Culture and Society, 3*(3), 189–200. https://doi.org/10.1016/j.ccs.2012.06.017

Fesselmeyer, E., & Seah, K. Y. S. (2018). The effect of localized density on housing prices in singapore. *Regional Science and Urban Economics, 68*, 304–315. https://doi.org/10.1016/j.regsciurbeco.2017.10.1016/j.regsciurbeco.2017.

Florio, P., Peronato, G., Perera, A., Blasi, A. D., Poon, K. H., & Kämpf, J. H. (2021). Designing and assessing solar energy neighborhoods from visual impact. *Sustainable Cities and Society, 71*, 102959. https://doi.org/10.1016/j.scs.2021.102959

Frantz, D., Schug, F., Okujeni, A., Navacchi, C., Wagner, W., Linden, S. V. D., & Hostert, P. (2021). National-scale mapping of building height using Sentinel-1 and Sentinel-2 time series. *Remote Sensing of Environment, 252*, 112128. https://doi.org/10.1016/j.rse.2020.112128

Ghermandi, A., & Sinclair, M. (2019). Passive crowdsourcing of social media in environmental research: A systematic map. *Global Environmental Change, 55*, 36–47. https://doi.org/10.1016/j.gloenvcha.2019.02.003

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal, 69*(4), 211–221. https://doi.org/10.1007/s10708-007-9111-y

Gui, S., & Qin, R. (2021). Automated LoD-2 model reconstruction from very-high-resolution satellite-derived digital surface model and orthophoto. *ISPRS Journal of Photogrammetry and Remote Sensing, 181*, 1–19. https://doi.org/10.1016/j.isprsjprs.2021.08.025

Guo, H., Li, X., Wang, W., Lv, Z., Wu, C., & Xu, W. (2016). An event-driven dynamic updating method for 3D geo-databases. *Geo-spatial Information Science, 19*, 1–8. https://doi.org/10.1080/10095020.2016.1182808

Hopf, K. (2018). Mining volunteered geographic information for predictive energy data analytics. *Energy Informatics, 1*(1), 4. https://doi.org/10.1186/s42162-018-0009-3

Huang, X., & Wang, C. (2020). Estimates of exposure to the 100-year floods in the conterminous United States using national building footprints. *International Journal of Disaster Risk Reduction, 50*, 101731. https://doi.org/10.1016/j.ijdrr.2020.101731

Huang, X., Wang, C., Li, Z., & Ning, H. (2020). A 100 m population grid in the CONUS by disaggregating census data with open-source microsoft building footprints. *Big Earth Data*, 1–22. https://doi.org/10.1080/20964471.2020.1776200

Juhász, L., & Hochmair, H. H. (2016). User contribution patterns and completeness evaluation of mapillary, a crowdsourced street level photo service. *Transactions in GIS, 20*(6), 925–947. https://doi.org/10.1111/tgis.12190

Juhász, L., & Hochmair, H. H. (2018). OSM Data Import as an Outreach Tool to Trigger Community Growth? A Case Study in Miami. *ISPRS International Journal of Geo-Information, 7*(3), 113. https://doi.org/10.3390/ijgi7030113

Kang, Y., Zhang, F., Gao, S., Peng, W., & Ratti, C. (2021). Human settlement value assessment from a place perspective: Considering human dynamics and perceptions in house price modeling. *Cities, 118*, 103333. https://doi.org/10.1016%2Fj.cities.2021.103333

Kang, Y., Zhang, F., Peng, W., Gao, S., Rao, J., Duarte, F., & Ratti, C. (2020). Understanding house price appreciation using multi-source big geo-data and machine learning. *Land Use Policy*, 104919. https://doi.org/10.1016/j.landusepol.2020.104919

Koeva, M., Humayun, M. I., Timm, C., Stöcker, C., Crommelinck, S., Chipofya, M., Bennett, R., & Zevenbergen, J. (2021). Geospatial tool and geocloud platform innovations: A fit-for-purpose land administration assessment. *Land, 10*(6), 557. https://doi.org/10.3390/land10060557

Komadina, A. & Mihajlovic, Z. (2022). Automated 3D Urban Landscapes Visualization Using Open Data Sources on the Example of the City of Zagreb. *KN - Journal of Cartography and Geographic Information*, 1–14. https://doi.org/10.1007/s42489-022-00102-w

Krotov, V., Johnson, L., & Silva, L. (2020). Tutorial: Legality and Ethics of Web Scraping. *Communications of the Association for Information Systems, 47*, 22. https://doi.org/10.17705/1CAIS.04724

Ledoux, H., Biljecki, F., Dukai, B., Kumar, K., Peters, R., Stoter, J., & Commandeur, T. (2021). 3dfier: automatic reconstruction of 3D city models. *Journal of Open Source Software, 6*(57), 2866. https://doi.org/10.21105/joss.02866

Lee, K. O., & Ooi, J. T. (2018). Property rights restrictions and housing prices. *The Journal of Law and Economics, 61*(2), 335–360. https://doi.org/10.1086/698747

León-Sánchez, C., Giannelli, D., Agugiaro, G., & Stoter, J. (2021). Testing the new 3D BAG dataset for energy demand estimation of residential buildings. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLVI-4/W1-2021*, 69–76. https://doi.org/10.5194/isprs-archives-xlvi-4-w1-2021-69-2021

Li, H., Liu, Y., Zhang, H., Xue, B., & Li, W. (2021). Urban morphology in China: dataset development and spatial pattern characterization. *Sustainable Cities and Society*, 102981. https://doi.org/10.1016/j.scs.2021.102981

Li, J. & Biljecki, F. (2019). The Implementation of Big Data Analysis in Regulating Online Short-term Rental Business: A Case of Airbnb in Beijing. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, IV-4/W9*, 79–86. https://doi.org/10.5194/isprs-annals-IV-4-W9-79-2019

Li, M., Koks, E., Taubenböck, H., & van Vliet, J. (2020). Continental-scale mapping and analysis of 3D building structure. *Remote Sensing of Environment, 245*, 111859. https://doi.org/10.1016/j.rse.2020.111859

Li, X., Cheng, S., Lv, Z., Song, H., Jia, T., & Lu, N. (2020). Data analytics of urban fabric metrics for smart cities. *Future Generation Computer Systems, 107*, 871–882. https://doi.org/10.1016/j.future.2018.02.017

Liang, C., Yeung, M. C. H., & Au, A. K. M. (2021). *The impact of airbnb on housing affordability: Evidence from hong kong* (p. 239980832110431). Environment and Planning B: Urban Analytics and City Science. https://doi.org/10.1177/23998083211043123

Lines, T., & Basiri, A. (2021). 3D map creation using crowdsourced GNSS data. *Computers, Environment and Urban Systems, 89*, 101671. https://doi.org/10.1016/j.compenvurbsys.2021.101671

Liu, X., Andris, C., Huang, Z., & Rahimi, S. (2019). Inside 50,000 living rooms: an assessment of global residential ornamentation using transfer learning. *EPJ Data Science, 8*(1), 1–18. https://doi.org/10.1140/epjds/s13688-019-0182-z

Liu, Y., Chen, C., Li, J., & Chen, W.-Q. (2020). Characterizing three dimensional (3-d) morphology of residential buildings by landscape metrics. *Landscape Ecology, 35*(11), 2587–2599. https://doi.org/10.1007/s10980-020-01084-8

Ma, R., Wang, T., Wang, Y., & Chen, J. (2022). Tuning urban microclimate: A morpho-patch approach for multi-scale building group energy simulation. *Sustainable Cities and Society, 76*, 103516. https://doi.org/10.1016/j.scs.2021.103516

Majic, I., Naghizade, E., Winter, S., & Tomko, M. (2021). There is no way! Ternary qualitative spatial reasoning for error detection in map data. *Transactions in GIS*. https://doi.org/10.1111/tgis.12765

Manoli, G., Fatichi, S., Schläpfer, M., Yu, K., Crowther, T. W., Meili, N., Burlando, P., Katul, G. G., & Bou-Zeid, E. (2019). Magnitude of urban heat islands largely explained by climate and population. *Nature, 573*(7772), 55–60. https://doi.org/10.1038/s41586-019-1512-9

Milojevic-Dupont, N., & Creutzig, F. (2021). Machine learning for geographically differentiated climate change mitigation in urban areas. *Sustainable Cities and Society, 64*, 102526. https://doi.org/10.1016/j.scs.2020.102526

Milojevic-Dupont, N., Hans, N., Kaack, L. H., Zumwald, M., Andrieux, F., Soares, D. D. B., Lohrey, S., Pichler, P.-P., & Creutzig, F. (2020). Learning from urban form to predict building heights. *PLoS ONE, 15*(12), e0242010. https://doi.org/10.1371/journal.pone.0242010

Mirkatouli, J., Samadi, R., & Hosseini, A. (2018). Evaluating and analysis of socio-economic variables on land and housing prices in mashhad, iran. *Sustainable Cities and Society, 41*, 695–705. https://doi.org/10.1016/j.scs.2018.06.022

Ning, H., Li, Z., Ye, X., Wang, S., Wang, W., & Huang, X. (2021). Exploring the vertical dimension of street view image based on deep learning: a case study on lowest floor elevation estimation. *International Journal of Geographical Information Science*, 1–26. https://doi.org/10.1080/13658816.2021

Nitoslawski, S. A., Galle, N. J., Bosch, C. K. V. D., & Steenberg, J. W. (2019). Smarter ecosystems for smarter cities? a review of trends, technologies, and turning points for smart urban forestry. *Sustainable Cities and Society, 51*, 101770. https://doi.org/10.1016/j.scs.2019.101770

Nowak, A., & Smith, P. (2016). Textual analysis in real estate. *Journal of Applied Econometrics, 32*(4), 896–918. https://doi.org/10.1002/jae.2550

Palliwal, A., Song, S., Tan, H. T. W., & Biljecki, F. (2021). 3D city models for urban farming site identification in buildings. *Computers, Environment and Urban Systems, 86*, 101584. https://doi.org/10.1016/j.compenvurbsys.2020.101584

Park, Y., & Guldmann, J.-M. (2019). Creating 3D city models with building footprints and LIDAR point cloud classification: A machine learning approach. *Computers, Environment and Urban Systems, 75*, 76–89. https://doi.org/10.1016/j.compenvurbsys.2019.01.004

Quinn, S., & León, L. A. (2019). Every single street? rethinking full coverage across street-level imagery platforms. *Transactions in GIS, 23*(6), 1251–1272. https://doi.org/10.1111/tgis.12571

Rahimi, S., Liu, X., & Andris, C. (2016). Hidden style in the city: an analysis of geolocated airbnb rental images in ten major cities. In *Proceedings of the 2nd ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics, UrbanGIS '16* (pp. 1–7). Association for Computing Machinery, New York. https://doi.org/10.1145/3007540.3007547

Schilling, J., & Tränckner, J. (2020). Estimation of Wastewater Discharges by Means of OpenStreetMap Data. *Water, 12*(3), 628. https://doi.org/10.3390/w12030628

See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., Fritz, S., Kerle, N., Jiang, B., Laakso, M., Liu, H.-Y., Milčinski, G., Nikšič, M., Painho, M., Pődör, A., Olteanu-Raimond, A.-M., & Rutzinger, M. (2016). Crowdsourcing, citizen science or volunteered geographic information? the current state of crowdsourced geographic information. *ISPRS International Journal of Geo-Information, 5*(5), 55. https://doi.org/10.3390/ijgi5050055

Shi, W., Zhang, M., Zhang, R., Chen, S., & Zhan, Z. (2020). Change Detection Based on Artificial Intelligence: State-of-the-Art and Challenges. *Remote Sensing, 12*(10), 1688. https://doi.org/10.3390/rs12101688

Sirko, W., Kashubin, S., Ritter, M., Annkah, A., Bouchareb, Y. S. E., Dauphin, Y., Keysers, D., Neumann, M., Cisse, M., & Quinn, J. (2021). Continental-Scale Building Detection from High Resolution Satellite Imagery. arXiv:2107.12283

So, W., & Duarte, F. (2020). Cartographers of North Korea: Who are they and what are the technical, political, and social issues involved in mapping North Korea. *Geoforum, 110*, 147–156. https://doi.org/10.1016/j.geoforum.2020.02.008

Su, S., He, S., Sun, C., Zhang, H., Hu, L., & Kang, M. (2021). Do landscape amenities impact private housing rental prices? A hierarchical hedonic modeling approach based on semantic and sentimental analysis of online housing advertisements across five Chinese megacities. *Urban Forestry & Urban Greening, 58*, 126968. https://doi.org/10.1016/j.ufug.2020.126968

Szarka, N., & Biljecki, F. (2022). Population estimation beyond counts–inferring demographic characteristics. *PLoS ONE, 17*(4), e0266484. https://doi.org/10.1371/journal.pone.0266484

Tian, W., Zhu, X., & Liu, Y. (2012). A Bottom-up Geospatial Data Update Mechanism for Spatial Data Infrastructure Updating. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XXXIX-B4. https://doi.org/10.5194/isprsarchives-XXXIX-B4-445-2012

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*, 102998. https://doi.org/10.1016/j.scs.2021.102998

Westrope, C., Banick, R., & Levine, M. (2014). Groundtruthing OpenStreetMap Building Damage Assessment. *Procedia Engineering, 78*, 29–39. https://doi.org/10.1016/j.proeng.2014.07.035

Witt, R., Loos, L., & Zipf, A. (2021). Analysing the Impact of Large Data Imports in OpenStreetMap. *ISPRS International Journal of Geo-Information, 10*(8), 528. https://doi.org/10.3390/ijgi10080528

Wu, A. N., & Biljecki, F. (2021). Roofpedia: Automatic mapping of green and solar roofs for an open roofscape registry and evaluation of urban sustainability. *Landscape and Urban Planning, 214*, 104167. https://doi.org/10.1016/j.landurbplan.2021.104167

Wu, A. N., & Biljecki, F. (2022). GANmapper: geographical data translation. *International Journal of Geographical Information Science, 36*, 1394–1422. https://doi.org/10.1080/13658816.2022.2041643

Xie, Y., Cai, J., Bhojwani, R., Shekhar, S., & Knight, J. (2019). A locally-constrained YOLO framework for detecting small and densely-distributed building footprints. *International Journal of Geographical Information Science, 34*(4), 1–25. https://doi.org/10.1080/13658816.2019.1624761

Yan, Y., Eckle, M., Kuo, C.-L., Herfort, B., Fan, H., & Zipf, A. (2017). Monitoring and assessing post-disaster tourism recovery using geotagged social media data. *ISPRS International Journal of Geo-Information, 6*(5), 144. https://doi.org/10.3390/ijgi6050144

Yan, Y., Feng, C.-C., Huang, W., Fan, H., Wang, Y.-C., & Zipf, A. (2020). Volunteered geographic information research in the first decade: a narrative review of selected journal articles in GIScience. *International Journal of Geographical Information Science, 34*(9), 1–27. https://doi.org/10.1080/13658816.2020.1730848

Yan, Y., Kuo, C.-L., Feng, C.-C., Huang, W., Fan, H., & Zipf, A. (2018). Coupling maximum entropy modeling with geotagged social media data to determine the geographic distribution of tourists. *International Journal of Geographical Information Science, 32*(9), 1699–1736. https://doi.org/10.1080/13658816.2018.1458989

Yang, J., Rong, H., Kang, Y., Zhang, F., & Chegut, A. (2021). The financial impact of street-level greenery on new york commercial buildings. *Landscape and Urban Planning, 214*, 104162. https://doi.org/10.1016/j.landurbplan.2021.104162

Yin, Y., Varadarajan, J., Wang, G., Wang, X., Sahrawat, D., Zimmermann, R., & Ng, S.-K. (2020). A Multi-task Learning Framework for Road Attribute Updating via Joint Analysis of Map Data and GPS Traces. In *Proceedings of The Web Conference 2020* (pp. 2662–2668). https://doi.org/10.1145/3366423.3380021

You, Q., Pang, R., Cao, L., & Luo, J. (2017). Image-Based Appraisal of Real Estate Properties. *IEEE Transactions on Multimedia, 19*(12), 2751–2759. https://doi.org/10.1109/TMM.2017.2710804

Yuan, C., Adelia, A. S., Mei, S., He, W., Li, X.-X., & Norford, L. (2020). Mitigating intensity of urban heat island by better understanding on urban morphology and anthropogenic heat dispersion. *Building and Environment, 176*, 106876. https://doi.org/10.1016/j.buildenv.2020.106876

Zhang, C., Fan, H., & Kong, G. (2021). VGI3D: an Interactive and Low-Cost Solution for 3D Building Modelling from Street-Level VGI Images. *Journal of Geovisualization and Spatial Analysis, 5*(2), 18. https://doi.org/10.1007/s41651-021-00086-7

Zhang, N., Luo, Z., Liu, Y., Feng, W., Zhou, N., & Yang, L. (2022). Towards low-carbon cities through building-stock-level carbon emission analysis: a calculating and mapping method. *Sustainable Cities and Society, 78*, 103633. https://doi.org/10.1016/j.scs.2021.103633

Zhang, X., Yin, W., Yang, M., Ai, T., & Stoter, J. (2018). Updating authoritative spatial data from timely sources: A multiple representation approach. *International Journal of Applied Earth Observation and Geoinformation, 72*, 42–56. https://doi.org/10.1016/j.jag.2018.05.022

Zhang, Y., Chen, N., Du, W., Li, Y., & Zheng, X. (2021). Multi-source sensor based urban habitat and resident health sensing: A case study of Wuhan. *China. Building and Environment, 198*, 107883. https://doi.org/10.1016/j.buildenv.2021.107883

Zheng, S., & Zheng, J. (2014). Assessing the Completeness and Positional Accuracy of OpenStreetMap in China. Lecture Notes in Geoinformation and CartographyIn T. Bandrova, M. Konecny, & S. Zlatanova (Eds.), *Thematic Cartography for the Society* (pp. 171–189). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-08180-9_14

Zhou, Q. (2018). Exploring the relationship between density and completeness of urban building data in OpenStreetMap for quality estimation. *International Journal of Geographical Information Science, 32*(2), 257–281. https://doi.org/10.1080/13658816.2017.1395883

Zielstra, D., Hochmair, H. H., & Neis, P. (2013). Assessing the Effect of Data Imports on the Completeness of OpenStreetMap - A United States Case Study. *Transactions in GIS, 17*(3), 315–334. https://doi.org/10.1111/tgis.12037

Publisher's Note