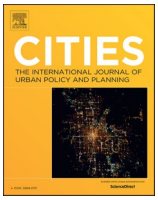
Cities 152 (2024) 105223

Contents lists available at ScienceDirect

Cities

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



Enhancing population data granularity: A comprehensive approach using LiDAR, POI, and quadratic programming

Xinyue Ye [a,b,*], Weishan Bai [a], Wenyu Wang [c], Xiao Huang [d]

[a] Department of Landscape Architecture and Urban Planning & Center for Geospatial Sciences, Applications and Technology, Texas A&M University, College Station, TX 77840, United States of America
[b] Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77840, United States of America
[c] Department of City and Regional Planning, Ohio State University, Columbus, OH 43210, United States of America
[d] Department of Environmental Sciences, Emory University, Atlanta, GA 30322, United States of America

ARTICLE INFO

Keywords:
Population downscaling
LiDAR remote sensing
Microsoft building footprint
Quadratic programming
Monte Carlo simulation

ABSTRACT

This research presents a sophisticated framework for the precise downscaling of population data from census blocks to individual residential units, employing an integration of housing unit characteristics. The aim was to devise and substantiate a thorough methodology for the distribution of households within specific residential buildings. Utilizing the Microsoft Building Footprint dataset, LiDAR remote sensing, and Point of Interest (POI) data, a detailed inventory of residential structures was compiled. A quadratic programming model and Monte Carlo Simulation techniques were applied independently for the strategic allocation of households to these buildings. For validation, this study conducted a comparative analysis between the two methods. The outcomes revealed that the quadratic programming model provided superior precision and detail in population data compared to the Monte Carlo Simulation technique. Consequently, the quadratic programming model significantly enhances the granularity of population distribution data, offering a valuable tool for more informed decision-making.

1. Introduction

The spatial distribution of populations is a critical area of study across various academic fields, such as demography, economics, and environmental science. These disciplines require detailed population data at a high resolution, including at the building level, to effectively conduct their research (Bakillah et al., 2014; Liu et al., 2008; Niu et al., 2017; Wan et al., 2022). Additionally, the accelerating dynamics of climate change and urbanization underscore the importance of examining vulnerability and sustainability at the building scale within urban design and landscape sustainability science (Elmer & Fraker, 2011; Rode et al., 2018; Ye et al., 2023). Understanding the demographic and socioeconomic characteristics at the building level is vital for gaining insights into residential behavior patterns and routines. These insights are instrumental in shaping strategies related to human mobility, public health, climate change adaptation, urban planning, and disaster mitigation (Watthanasutthi, 2016; Yao et al., 2017).

The disaggregation of census data raises significant privacy concerns due to the potential for demographic and socioeconomic information to inadvertently reveal individual identities, rendering such detailed data unavailable at the building level (Ural et al., 2011). In the absence of comprehensive datasets, alternative sources such as micro samples or Public Use Microdata Samples (PUMS) and marginal statistics provide key socio-demographic attributes. However, existing methodologies often lack spatial precision, resulting in broadly generalized regional data. To achieve a more spatially precise population allocation, it is essential to expand the synthetic population representation to include detailed aspects of the built environment, such as housing units, buildings, and other establishments. The integration of precise spatial data can significantly enhance the accuracy and applicability of these analyses, thereby informing more effective urban planning and transportation strategies.

This paper introduces a novel framework for the downscaling of population data from the census block level to the individual building

* Corresponding author at: Department of Landscape Architecture and Urban Planning & Center for Geospatial Sciences, Applications and Technology, Texas A&M University, College Station, TX 77840, United States of America.
E-mail addresses: xinyue.ye@tamu.edu (X. Ye), weishanb@tamu.edu (W. Bai), wang.17238@buckeyemail.osu.edu (W. Wang), xiao.huang2@emory.edu (X. Huang).

https://doi.org/10.1016/j.cities.2024.105223
Received 4 December 2023; Received in revised form 10 June 2024; Accepted 14 June 2024
Available online 1 July 2024

level. It utilizes open-source data to compile a comprehensive inventory of housing units. Additionally, the Microsoft Building Footprint dataset, LiDAR remote sensing data, and Point of Interest (POI) data are meticulously combined using dasymetric mapping techniques to create a detailed building inventory. A quadratic programming model is subsequently developed for the precise allocation of individual housing units to specific residential buildings. The methodology's effectiveness is demonstrated through its application in Galveston Island, TX. Validation is achieved through a comparative analysis with Monte Carlo Simulation, which highlights the method's enhanced accuracy and detail. The key contributions of this study include:

- Developing a method for fine-grained population downscaling from the census block level to the building level.
- Employing the Microsoft Building Footprint dataset, LiDAR remote sensing data, and POI data to generate comprehensive building inventories.
- Implementing a Mathematical Formulation approach to optimize the allocation of households to individual buildings using a static Quadratic Programming model.

The paper is organized as follows: The next section provides a review of relevant literature, identifying key gaps and contributions. Section 3 describes the process of compiling housing unit and building inventories for Galveston Island, TX. Section 4 introduces the quadratic programming model and the Monte Carlo Simulation methodology. Section 5 presents the results and compares the two methods. Finally, Section 6 concludes the paper with discussions and perspectives for future research directions.

2. Relevant works

2.1. Data utilization in population allocation

Interpolation and dasymetric methods are developed for the refined disaggregation of census data into grid cells, incorporating a variety of ancillary data sources (Briggs et al., 2007; Gallego et al., 2011; Leyk et al., 2019; Li & Zhou, 2018). A primary technique in this realm is a real interpolation, essential for transposing socioeconomic data across varied spatial units (Wu et al., 2005). Among these techniques, the direct areal weighting method is fundamental, distributing population data from a source zone to target zones in proportion to each target zone's area (Goodchild & Lam, 1980). However, this method is limited by its assumption of uniform population density within the source zone, which is rarely the case in real-world scenarios. Dasymetric mapping, by contrast, has emerged as a more effective approach for spatial downscaling (Mennis, 2009; Li & Zhou, 2018). When specific population density data is not available, binary dasymetric mapping is often used. This method divides target areas into two distinct zones, each characterized by unique population features derived from additional datasets (Su et al., 2010).

Satellite-based remote sensing products are commonly used as supplementary data sources. These include land use (Tan et al., 2018; Weber et al., 2018) and nighttime light imagery (Chen et al., 2019; Li & Zhou, 2018), along with Point of Interest (POI) data (Yang et al., 2019; Ye et al., 2019). These datasets act as weighting surfaces that help delineate the uneven distribution of populations. However, nighttime light data can be less effective for smaller areas due to its relatively lower spatial resolution (Stathakis & Baltas, 2018). In a novel approach, Huang, Wang, et al. (2021) successfully disaggregated population data from the census tract level to a 100-m grid within the CONUS region, utilizing the open-source Microsoft building footprint data. Launched in June 2018, the Microsoft building footprint dataset, with its 125 million building footprints, represents one of the most comprehensive collections available. A significant limitation, however, is its lack of building height data, leading to potential inaccuracies in population estimates, particularly in urban areas with high-rise structures.

Incorporating building height data can significantly enhance the accuracy of building volume calculations, thereby improving population estimations, especially in densely populated urban settings. Research by Stal et al. (2013) demonstrated the potential of extracting building heights using aerial photogrammetry techniques. In this study, we integrate LiDAR data, which provides building height information, with the Microsoft building footprints dataset and POI data. The inclusion of POI data is instrumental in differentiating residential from non-residential buildings. Ultimately, we apply dasymetric mapping techniques to effectively combine these diverse datasets for more precise population distribution analysis.

2.2. Population allocation methods

Iterative Proportional Fitting (IPF) has gained prominence in demographic research, particularly for its effectiveness in addressing complex population synthesis challenges. Initially conceptualized as a numerical method for analyzing contingency tables (Deming & Stephan, 1940), IPF's fundamental operation involves aligning a contingency table, derived from microdata, with marginal constraints obtained from more extensive aggregated census data (Beckman et al., 1996).

Despite its widespread application due to its intuitive concept, IPF encounters several notable challenges. In response to these challenges, researchers have sought to refine the IPF process. A significant development is the Iterative Proportional Updating (IPU) algorithm introduced by Ye et al. (2009), designed to reconcile both household and individual-level data distributions. This algorithm has been implemented in PopGen, a leading-edge open-source synthetic population generator (Konduri et al., 2016). Moving towards probabilistic approaches, some methods have deviated from traditional IPF techniques, creating more diverse agents that are not simply replicas of the microdata sample (Saadi et al., 2016; Sun et al., 2018; Zhang et al., 2019). Innovative strategies, including the application of Markov Chains for probabilistic population synthesis and the use of data-driven inferential methods like Bayesian Networks, have also been explored (Sun & Erath, 2015; Zhang et al., 2019).

While these methods excel in synthesizing populations at larger geographic scales, achieving high spatial resolution requires more complex allocation strategies. Some approaches, like Rosenheim et al.'s (2021) method, rely on a randomized selection process, using Monte Carlo simulations to distribute socio-demographic data to individual housing units. Fereshtehnejad et al. (2021) extended this approach by predicting housing unit occupancy status and, when occupied, determining household characteristics. Other research has aimed to reduce the randomness of these assignments. Harada and Murata (2017) developed a technique to project synthetic households onto geographic maps, utilizing critical geospatial data to allocate households according to building specifications. Chapuis et al. (2018) combined satellite imagery and building geometry data to estimate population density at the micro-level, applying areal interpolation for allocation while controlling distribution.

In conclusion, significant advancements have been made in the field of population synthesis, yet there remains a discernible gap in fully integrating spatial details and the built environment's representation. To authentically capture the spatial dynamics of populations, a comprehensive and accurate integration of spatial information, including precise population allocation to buildings, is crucial.

3. Study area and data

3.1. Testing site and method overview

Galveston Island, located about 50 miles south of Houston at the juncture of Galveston Bay, is inherently prone to natural disasters, particularly hurricanes, due to its strategic geographic location. This

susceptibility highlights the island's importance as a critical site for urban research, as noted in Huang, Ye, et al. (2021). The urban design and architectural features of Galveston Island are representative of typical U.S. urban planning models, as discussed in Beasley (2006). This alignment makes the island an ideal case study for examining urban planning principles and understanding their wider implications. Insights derived from Galveston Island are of considerable value and can be applied to similar urban environments, thus contributing significantly to the field of urban planning research.

This study introduces a sophisticated framework to enhance the resolution of population data, shifting from the traditional census block approach to a more detailed, building-focused analysis, incorporating relevant housing unit data (illustrated in Fig. 1). The methodology employs the Microsoft Building Footprint dataset, LiDAR remote sensing data, and Point of Interest (POI) datasets to create a comprehensive building inventory. The next phase combines housing unit and building data, using a two-pronged approach of Mathematical Formulation and Monte Carlo Simulation to assign households to specific buildings. The Mathematical Formulation, based on optimizing population allocation within household number constraints, results in a static Quadratic Programming model. Conversely, the Monte Carlo Simulation, guided by population density metrics, performs allocation through a probabilistic process. Together, this integrated approach significantly enhances the accuracy of population distribution datasets, making it a valuable tool for informed decision-making in various academic fields.

3.2. Datasets

This study endeavors to elucidate the intricate interrelationships between distinct social entities, notably households characterized by diverse attributes, and architectural structures, such as buildings. Central to this exploration is the compilation of both housing unit and building inventories. These inventories serve as instrumental conduits, enabling the transformation of aggregated household and housing unit data into discrete housing units. Each of these units is imbued with particular attributes, resonating with a specific housing unit typology. These attributes can then be seamlessly integrated with the overarching building inventory.

3.2.1. Housing unit inventory

This research draws upon the housing unit inventory methodology delineated by Rosenheim et al. (2021) to meticulously catalog the nuanced characteristics of residential edifices located on Galveston Island. The exhaustive dataset derived from this housing unit inventory offers intricate specifics pertaining to each individual housing unit, with representative examples presented in Table 1.where the *huid* functions as a distinct identifier, uniquely demarcating each housing unit. The *blockid* aligns with the 2010 Census Block ID and acts as the pivotal reference, facilitating the association of housing units with the corresponding buildings in the inventory. This association is not a direct one-to-one linkage but aligns each housing unit with potentially multiple buildings within the same block. Concurrently, *numprec* denotes the population tally associated with the specific *huid*. Expanding upon this, the housing unit inventory encapsulates a plethora of attributes, including but not limited to tenure status, racial categorization, vacancy typology, poverty indicators, and other salient characteristics.3.2.2 Building Inventory

A nuanced comprehension of both the volume and spatial dispersion of architectural structures is intrinsically linked to the availability and accuracy of rooftop data. In this context, the Microsoft Building Footprint dataset stands out as an indispensable asset, proffering meticulous building contours that encapsulate both positional and morphological facets of building rooftops. These facets are instrumental in charting the spatial orientation and unique attributes of buildings (Heris et al., 2020). The Microsoft Building Footprint dataset used in this research is a public dataset released by Microsoft.[1] It is generated through a DNN model based on Bing Imagery (with an average shooting year of 2012) and has extracted polygons for 10,678,921 buildings in the Texas area. Polygon evaluation metrics calculated from three dimensions: Intersection over Union, Shape distance, and Dominant angle rotation error, indicate that the dataset's precision reached 98.5 %, and recall reached 92.4 %.

Concurrently, LiDAR remote sensing data, distinguished by its unparalleled precision and high-resolution capabilities, furnishes an intricate point cloud depiction of the urban landscape (Zhao et al., 2017). This representation ensures the accurate delineation of building features, encompassing geographical coordinates, stature, and spatial dimensions. The LiDAR data used in this study originates from Houston-Galveston Area Council(H-GAC), Texas Natural Resources Information System (TNRIS), and United States Geological Survey (USGS), and was ultimately published through Environmental Systems Research Institute (ESRI).[2] The data includes radar point cloud files related to the shape, location, and height of buildings, with the collection year being 2018. The data covers the Harris County and Galveston County areas. The accuracy of this LiDAR data has reached an RMSE of 10 cm (non-vegetated). The amalgamation of insights from the Microsoft Building Footprint dataset and LiDAR data results in the extraction of 29,148 building rooftops, rendered in polygonal format, as depicted in Fig. 2b.

The POI dataset serves as an instrumental criterion for discerning residential edifices from their non-residential counterparts (Ye et al., 2019). The POI data used in this study was collected from the SafeGraph Places POI data collection in December 2022. This data set covers the attribute information of a total of 2086 POIs in the study area, including latitude and longitude, address, business information, POI category, etc. By comparing with Google Maps, the error of the data in latitude and longitude is between 0 and 5 m, the accuracy of matching with buildings exceeds 70 %, and the accuracy of POI attributes exceeds 99 %.[3]

By applying specific spatial analysis tools within ArcGIS, our methodology involves a detailed spatial intersect analysis between building polygons and POI obtained from Safegraph. This process utilizes ArcGIS's 'Intersect' tool to pinpoint buildings that occupy the same geographical space as any POI. This analysis identifies buildings that share spatial coordinates with any POIs. Subsequently, buildings with such overlaps are designated as non-residential, reflecting the assumption that their intersection with varied POIs suggests uses beyond residential purposes. When dealing with mixed types of buildings, a simplifying assumption is adopted: a building is classified as nonresidential if it spatially coincides with one or more POIs for nonresidential uses.

This methodological approach effectively omits commercial, industrial, and other non-residential infrastructures from the analytical purview since the dataset of POIs encompasses attributes of locations, incorporating names (for instance, McDonald's, Fresenius Kidney Care, Shoe Dept. Encore, among others) as well as categories (such as Automotive Repair and Maintenance, Grocery Stores, etc.). The view of the POI dataset is shown in Fig. 2c.

By transmuting the information gleaned from both the Microsoft

[1] This Microsoft Building Footprint is made available under the Open Database License: http://opendatacommons.org/licenses/odbl/1.0/. Any rights in individual contents of the database are licensed under the Database Contents License: http://opendatacommons.org/licenses/dbcl/1.0/

[2] https://tiles.arcgis.com/tiles/lqRTrQp2HrfnJt8U/arcgis/rest/services/Building_Footprints_2018/MapServer

[3] https://docs.safegraph.com/docs/places-data-evaluation

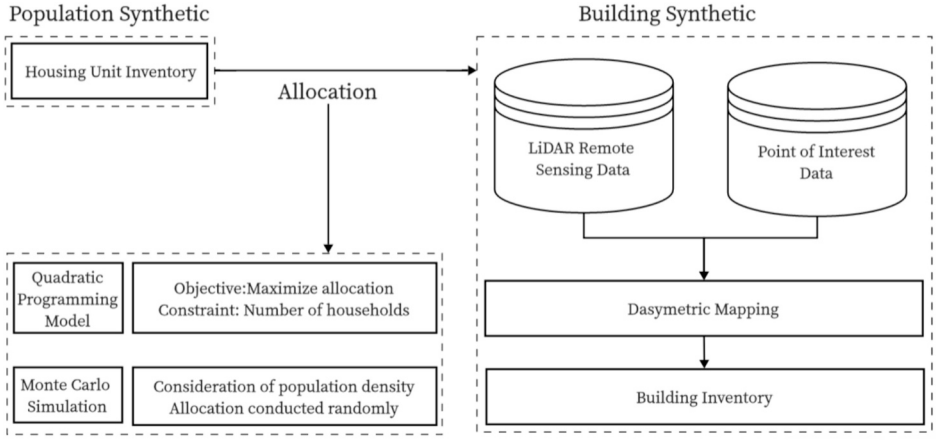


Fig. 1. The proposed framework in this study.

Table 1
A sample of housing unit inventory.

huid	*blockid*	*numprec*
B481677240001047H006	481,677,240,001,047	1
B481677240002091H001	481,677,240,002,091	1
B481677241011038H001	481,677,241,011,038	1
B481677241011053H001	481,677,241,011,053	1
B481677241011056H001	481,677,241,011,056	1

Building Footprint dataset and the LiDAR point cloud into vectorial data, and subsequently excising non-residential edifices, ArcGIS shapefiles are synthesized. These files encapsulate the exact geographical coordinates, elevation metrics, and spatial dimensions of each edifice, each distinctly demarcated by the *buid*. A representative excerpt of the residential building inventory, rendered in shapefile format, is depicted in Fig. 2d.

Dasymetric mapping employs a systematic series of spatial subdivisions to augment the granularity of datasets originally amassed at a broader scale (Eicher & Brewer, 2001). Drawing upon the methodological framework delineated by Huang, Ye, et al. (2021), with a particular emphasis on the disaggregation of potential household numbers and population metrics, the process unfolds as follows, given the established number of households (H) and population (P) within the designated study block:

$$H_b^e = H \times \frac{D_b \times H_b}{\sum_b D_b \times H_b} \tag{1}$$

$$P_b^e = P \times \frac{D_b \times H_b}{\sum_b D_b \times H_b} \tag{2}$$

whereH_b^e and P_b^e refer to the estimated number of households and population respectively in the target building b. D_b and H_b denote the dimensions of the allocated building and the corresponding height assigned to this building respectively. The selected building b is within the study block spatially.

Eventually, the comprehensive building inventory dataset is exemplified in Table 2.

In Table 2, *buid* serves as a distinct identifier, uniquely demarcating each architectural structure. Concurrently, the *blockid* is utilized to forge a linkage with the housing unit inventory. Furthermore, the variables hh and pop denote the tallies of households and populations, respectively, corresponding to the specific buid. Within a given *blockid*, multiple *buid* entities can be identified. Each *buid* is, in turn, associated with several huid entities, all unified under the unique identifier *blockid*. Conceptually, a block encompasses an array of buildings, with each building housing multiple individual units.

4. Population synthesis algorithm

4.1. Mathematical formulation

Table 3 summarizes all the variables and parameters used in formulating the model.

The objective function is defined as the total number of households allocated to buildings, aiming to maximize the allocation of households to buildings as much as possible, which can be formulated as

$$max\ \text{obj} = \sum_{h\in\mathscr{H}}\sum_{b\in\mathscr{B}} x_{h,b} \tag{3}$$

To assign each household to a building, which can be formulated as

$$\sum_{b\in\mathscr{B}} x_{h,b} \le 1 \forall h \in \mathscr{H} \tag{4}$$

The assignment of each household to a single building ensures the assignment's uniqueness. Additionally, allocating household to a building is contingent upon meeting the building's maximum number of households and total population.

$$\sum_{h\in\mathscr{H}} x_{h,b} \le H_b^e \forall b \in \mathscr{B} \tag{5}$$

$$\sum_{h\in\mathscr{H}} P_h x_{h,b} \le P_b^e \forall b \in \mathscr{B} \tag{6}$$

The stated formulas indicate that the assignment of households to a building is subject to a constraint ensuring that the number of households does not exceed the building's designated maximum capacity. Likewise, the total population residing within a building must adhere to the building's maximum occupancy threshold. This ensures that both household and population allocations remain within feasible and predefined limits for each building.

Therefore, the household allocation problem can be formulated as the following programming model:

$$(\text{HUAP})\ max\left\{\sum_{h\in\mathscr{H}}\sum_{b\in\mathscr{B}} x_{h,b} : (4) \sim (6)\right\} \#(7) \tag{7}$$

Several modifications are made to improve the robustness of the model. Firstly, the original eq. (4) is altered to ensure all households can be assigned to buildings, i.e., eq. (8). Eq. (5) is adjusted to ensure that the number of households corresponds to the buildings, i.e., eq. (9). Additionally, eq. (6) undergoes a relaxation operation. Finally, the objective function is updated to minimize the difference between the relative total population of households assigned to each building, i.e., eq. (10). The improved formulas are as follows.

Fig. 2. Data set in the study area. a) Study area; b) Zoom in view of building's roof data; c) Zoom in view of POI data; d) Zoom in view of residential building inventory.

Table 2
A sample of building inventory.

buid	*blockid*	*hh*	*pop*
10,028	481,677,240,001,047	5	9
307	481,677,240,002,091	1	5
5714	481,677,241,011,038	10	11
10,903	481,677,241,011,053	37	89
8212	481,677,241,011,056	6	9

Table 3
Notations of sets and parameters.

Notation	Detailed Definition
Sets	
$\mathscr{H}$	Set of households
$\mathscr{B}$	Set of buildings
Parameters	
P_h	The population for each household h, $h \in \mathscr{H}$
H_b^e	The estimated number of households for each building b, $b \in \mathscr{B}$
P_b^e	The estimated population for each building b, $b \in \mathscr{B}$
Decision variables	
$x_{h,b}$	Binary variables, $= 1$ if household h is allocated to building b, $=$ 0 otherwise. $h \in \mathscr{H}, b \in \mathscr{B}$.

$$\sum_{b\in\mathscr{B}} x_{h,b} = 1 \forall h \in \mathscr{H} \tag{8}$$

$$\sum_{h\in\mathscr{H}} x_{h,b} = H_b^e \forall b \in \mathscr{B} \tag{9}$$

$$\text{minax obj} = \sum_{h\in\mathscr{H}} \left| \sum_{b\in\mathscr{B}} P_h x_{h,b} - P_b^e \right| \tag{10}$$

The household allocation problem can be formulated as the improved quadratic programming model:

$$(\text{HUAP\_improved})\ min \left\{ \sum_{h\in\mathscr{H}} \left| \sum_{b\in\mathscr{B}} P_h x_{h,b} - P_b^e \right| : (8) \sim (9) \right\} \tag{11}$$

4.2. Monte Carlo simulation

The Monte Carlo Simulation process, as detailed by Rosenheim et al. (2021), entails the allocation of households to buildings within specific census blocks. This allocation is conducted probabilistically to predict the likelihood that a building will be occupied by its owner.

The initial step involves the calculation of the cumulative size of all buildings within a specified block. The proportionate size of each building is determined by dividing the dimensions of all buildings within the given block.

$$Pr_b = \frac{D_b}{\sum_b D_b} \forall b \in \mathscr{B} \tag{12}$$

where Pr_b is the relative size of each building as a proportion of the total size.

Then, the predetermined number of iterations is defined, and in each iteration, a random number of households are assigned to each building according to its size proportionally.

$$H_b = \text{round}\left(Pr_b \times \sum_b H_b^e\right) \forall b \in \mathscr{B} \tag{13}$$

where H_b represents the number of households that need to be allocated to the building and using 'round' ensures that H_b becomes an integer.

$$P_b = \sum_h P_h \forall h \in \mathscr{H}_b \tag{14}$$

where P_b denotes the population that needs to be allocated to a building, and $\mathscr{H}_b$ represents the assigned households in a given building. When P_b is at least equal to the estimated population stated in eq. (15), households will be given to the building. Otherwise, households are randomly assigned to other buildings while adhering to the population constraint.

$$P_b >= P_b^e \forall b \in \mathscr{B} \tag{15}$$

Household counts and population figures are iteratively refreshed until all households have been allocated to buildings.

5. Results

The study encompassed 1641 blocks and an allocated population of 22,530 households. In the application of the quadratic programming model, Gurobi 9.5.2 was adopted to address the intricacies of the quadratic programming paradigm. A temporal constraint, capped at 7200 s, was instituted to ensure timely convergence of the solution. In this study, residual plots and *Mean Squared Error* (*MSE*) are employed to scrutinize the effectiveness of the two methods.

Residuals, representing the disparities between observed values and model predictions, are crucial for evaluating predictive model efficacy (Cox & Snell, 1968). We utilize residual plots to assess the effectiveness of population allocation methods for 22,520 households, as shown in Fig. 3. The $x-axis$ in our graphical representation denotes the allocated population for each household. At the same time, the $y-axis$ captures the residual values—depicting the differences between the actual and given population. In Fig. 3, the x-axis shows the allocated population for each household. The extremely large populations for some households are a result of the way data is recorded for certain census blocks that include institutions or organizations. For example, a census block may contain a single building but be associated with the entire population of an institution that spans multiple blocks. In the case of the university campus, the census block includes only the residential life building but registers the total population of 328 individuals, reflecting the entire school. This allocation method leads to unusually high population counts for that.

As left of the Fig. 3 corresponding to the *HUAP_improved* method, the residual plot reveals a concentration of residuals within the −10 to +10 range. However, notable deviations emerge, with four outliers exhibiting errors of around 20 and a single point demonstrating an error of approximately 70. It is that this specific block encompasses only one building. Due to a necessity for the total household count to align, this condition contributes to an increased discrepancy in population allocation. In contrast, the residuals associated with the *MCS* method predominantly cluster about 20. As a result, the *HUAP_improved* method demonstrates superior performance, evident in a more confined dispersion of residuals. This characteristic suggests heightened accuracy in population allocation, underscoring its efficacy compared to the *MCS* method.

The *MSE* is a statistical metric utilized to quantify the average squared differences between observed and predicted values (Das et al., 2004). *MSE* is a pivotal measure for assessing model performance, as it emphasizes systematic and random errors, as follows:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(P_h^e - \widehat{P_h^e}\right)^2 \tag{16}$$

Where n represents the number of observations, totaling 22,520 households, P_h^e signifies the estimated population for each household, and $\widehat{P_h^e}$ denotes the allocated population. Delving into the outcomes, the *Mean Squared Error* (*MSE*) for the *HUAP_improved* method computes 0.6522. This value signifies a diminished average squared difference between observed and allocated values. In contrast, the *MCS* method presents a substantially elevated *MSE* of 28.4717, indicative of a more pronounced discrepancy between assigned and actual values. Consequently, the *HUAP_improved* method's superiority is accentuated by its notably lower *MSE*, underscoring its efficacy in minimizing errors compared to the *MCS* method.

For illustrative clarity, a census block identified by the *blockid* 481,677,246,002,024 was selected. In this census block, 15 buildings

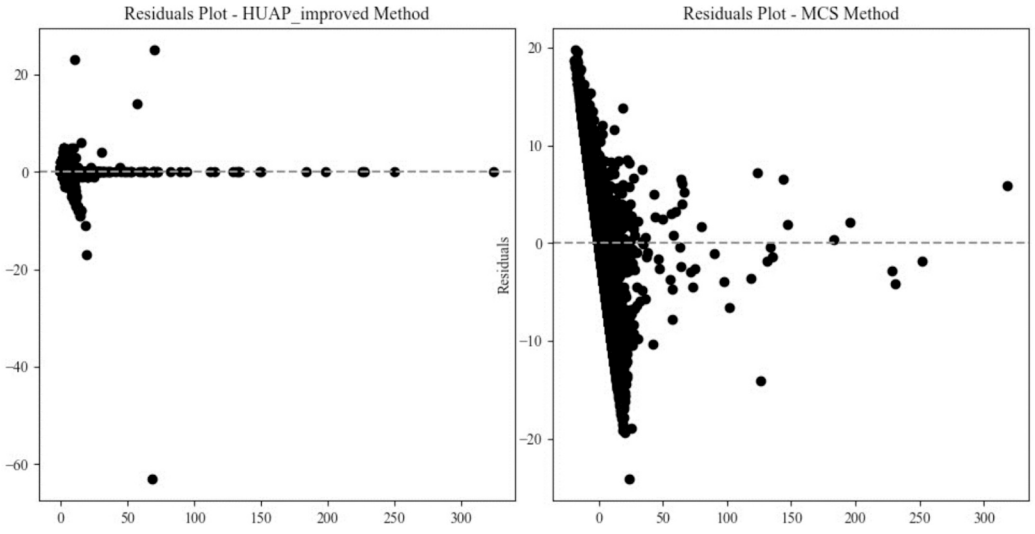


Fig. 3. Residual Plot for the two allocation methods

accommodate a population of 601 individuals and consist of 184 households. As demonstrated in Table 4, P_b^e and H_b^e denote the estimated population and households assigned to building b, respectively, while P_b and H_b represent the actual allocation to building b. The errors in allocating population and households to buildings are computed to compare the quadratic programming model (*HUAP_improved*) and Monte Carlo Simulation (*MCS*) methodologies.

As Table 4, when utilizing the *HUAP_improved* methodology, allocating households and the population count to each building is accomplished with remarkable precision, exhibiting a negligible 0 % error rate. In contrast, when compared to the *MCS* method, which shows an almost imperceptible 5.49 % error rate.

In summary, the *HUAP_improved* method substantially reduces assignment errors compared to the *MCS* method by employing a Quadratic Programming model. It systematically allocates population and households to individual buildings, focusing on optimizing allocation accuracy, which significantly outperforms random assignment in terms of precision and reliability.

5 Discussion

5.1. Transferability and reproducibility of this research

The datasets used in this study primarily include data from the United States Census, Microsoft Building Footprints, SafeGraph Points of Interest data, and LiDAR remote sensing data. The availability of these datasets poses certain challenges for the transferability and reproducibility of the study. The first two datasets are publicly accessible and downloadable, covering the entire United States with stable updates (usually annually, while the census block scale data is updated every ten years). However, SafeGraph POI data does not offer free access or downloads, but it provides good accuracy, covers >220 countries & territories, and maintains a good update frequency (every 3–6 months). As for the LiDAR data, it does not provide free public access or clear information on update frequency, but this is only for the current study area. Availability of data can vary for different research areas. For the two datasets that are not publicly available, local government departments or academic institutions often provide corresponding POI or LiDAR data depending on the study area. Even without these two types of data, the method proposed in this study for population disaggregation remains valid. This is because the main purpose of these two types of data is to determine the physical information and usage type of buildings, and many studies have proposed different data sources and methods to achieve this goal. For example, Atwal et al. (2022) used the publicly available dataset OpenStreetMap to predict building types. Of course, it must be acknowledged that the quality of data, especially concerning the physical and usage information of buildings, significantly affects the outcomes of this study since they are used to ascertain the number, location, and volume of residential buildings.

Table 4
Comparison of results between *HUAP_improved* and *MCAS*.

buid	P_b^e	P_b		H_b^e	H_b	
		HUAP_improved	*MCS*		*HUAP_improved*	*MCS*
11,985	24	24	24	7	7	7
12,000	33	33	34	10	10	10
12,014	2	2	4	1	1	1
12,027	60	60	59	18	18	18
12,043	13	13	12	4	4	4
12,050	25	25	23	8	8	8
12,056	50	50	50	15	15	16
12,091	48	48	27	15	15	8
12,101	25	25	23	8	8	8
12,105	47	47	43	14	14	16
12,113	18	18	19	6	6	8
13,874	68	68	66	21	21	21
13,878	70	70	73	22	22	23
14,444	94	94	90	28	28	29
28,128	24	24	21	7	7	8
Sum	601	601	568	184	184	185
Error	–	0	5.49 %	–	0	0.54 %

5.2. Limitation

In this study, the accuracy assessment of the HUAP-improved and MCS methods was based on estimated population data derived from Microsoft building footprints augmented with LiDAR and POI data. While this approach allowed us to conduct a comprehensive analysis within the constraints of available resources and data, it is essential to acknowledge that these estimates do not represent ground-truth data. Consequently, the accuracy metrics derived from this comparison must be interpreted cautiously as they do not reflect accurate ground-truth accuracies. This limitation is significant as it could potentially influence the perceived effectiveness of the population downscaling results produced by both algorithms. A more robust validation method would involve downscaling from the census block group level to the building level, followed by aggregation to the census block level. This aggregated data could then be compared against actual ground-truth survey data to provide a more definitive assessment of each method's accuracy. Implementing such a method would allow for a more transparent and precise evaluation of how these algorithms replicate known population distributions, thereby offering more substantial contributions to the population downscaling literature. Here are the specific limitations identified in the approach used for the accuracy assessment in the population downscaling study.

5.2.1. The assumption of building's maximum occupancy threshold

We constrain the allocation algorithm by setting a maximum number of people/households per building. This is based on the concept that there is a positive correlation between building volume and population density. Several studies have reached this conclusion. For example, in a study by Zhao et al. (2017), they tested the relationship between population and built volume in four different cities in Texas, and the results showed that in family-oriented mixed-use communities, there is a positive correlation between the relationship between building volume and population size. In the study of Shahfahad et al. (2021), using satellite image data and census data, they found through statistical analysis that there is a strong positive correlation between population density and built-up area area. It is true that we must admit that such a concept has limitations and does not hold true in many cases. For example, in high-income neighborhoods, a small number of people/households live in large buildings. Zięba-Kulawik et al. (2020) used aerial orthophotos and airborne laser scanning (ALS) point clouds to calculate the building volume of Luxembourg City from 2001 to 2019, and conducted correlation analysis with population data, and found that population and the correlation of building volumes is dynamic. However, we were unable to obtain detailed information on the population living in a specific building from publicly available data to analyze the specific situation. Therefore, our study is based on the consensus of many population studies: population density is directly proportional to residential building volume.

5.2.2. POI and mixed type buildings

We faced the challenge of accurately handling and classifying mixed-use buildings, which in reality may serve both residential and non-residential functions. Given the nature of the non-residential POI dataset on which we relied primarily, we adopted a methodological strategy that was both practical and consistent with data availability to classify possible mixed-use buildings.

Specifically, when a building geographically coincides with one or more POI identified as non-residential uses, we classify it as a non-residential building based on existing data limitations and simplified analysis needs. This decision reflects the trade-offs researchers face during the data collection and processing stages when faced with

common situations of incomplete information and limited processing capabilities. We recognize that this approach may not accurately reflect the characteristics of all uses of a building, and particularly for buildings with commercial space on the ground floor and residential above, this classification may ignore the residential component of the building.

Although adopting this classification strategy increases the simplicity and consistency of research operations, it may also have a specific impact on the research results. For example, our analysis may overemphasize buildings with non-residential uses, thereby affecting to some extent the understanding of the distribution of uses in urban space. In this regard, our study reveals an important data and methodological limitation of current urban research and points to directions for further exploration and improvement in future work.

Future research can improve the accuracy of building use classification by introducing more detailed and comprehensive data sets, such as combining real estate registration information, resident survey results, or using more advanced spatial data analysis techniques. This will not only provide a more accurate classification of mixed-use buildings, but also help researchers better understand the role and function of these buildings in urban space, thereby providing richer insights for urban planning and development.

5.2.3. The impact of dynamic population

We adopted a method focused on residential buildings to enhance the granularity of population data. This method aims to simulate population distribution by identifying and analyzing residential buildings. We recognize that while this approach has advantages in improving the accuracy of population data in residential areas, it also has certain limitations. Specifically, this method may not fully capture the living situations in non-traditional living spaces (such as mixed-use buildings with both commercial and residential functions), thereby affecting the understanding of the complete picture of urban population distribution. Additionally, the treatment and analysis of population data in this study are based on static data benchmarked against census data, meaning that our study lacks the ability to capture population dynamics. That is, the population we study refers to people who reside permanently in their own residential buildings and does not include those who live elsewhere for extended periods or are frequently in a state of mobility.

Future research could start with data sources closer to population mobility trends, such as mobile devices signal and social media, and employ more advanced spatial analysis methods and population simulation technologies to more accurately simulate population distribution in urban area. This would enhance the study of dynamic changes in the population within urban area, especially considering the emergence of new living forms during urban development and their impact on population distribution.

5.2.4. Building height data and dynamic allocation methods

The inclusion of building height data represents a significant enhancement in the accuracy of population distribution models. Urban areas, particularly those with high-density housing, such as apartments and high-rises, benefit substantially from this approach. Traditional population estimation methods often overlook vertical spatial variations, leading to significant inaccuracies in densely populated areas. By incorporating building height, we can assign population figures more accurately across different floors and better reflect the actual distribution within urban landscapes. This approach improves the granularity of population data and aids in urban planning and infrastructure development, ensuring resources are appropriately allocated based on actual population densities.

Our study aimed to use a static quadratic programming model to optimize the allocation of resources based on predefined constraints and objectives. This model is particularly beneficial for its efficiency and stability in producing optimal and feasible solutions within the set parameters. However, one limitation of the static model is its inability to dynamically adjust to real-time changes in input data, such as shifts in population dynamics or urban development patterns. In future work, exploring dynamic models that adjust allocations based on ongoing data updates could be especially useful in rapidly changing urban environments.

6. Conclusion

This paper introduces a novel framework aimed at enhancing the resolution of population data, transitioning from the traditional census block level to a more detailed scale—individual buildings—while integrating relevant housing unit data. Utilizing the advanced Microsoft Building Footprint dataset, LiDAR remote sensing techniques, and Point of Interest (POI) data, this methodology meticulously compiles a comprehensive and accurate building inventory.

Further, the study employs sophisticated Mathematical Formulation and Monte Carlo Simulation methods to assign households to specific buildings accurately. A comparative analysis of these techniques demonstrates significant improvements in the accuracy and detail of the resulting population distribution models, owing to the refined approach of the former. The research presents an innovative method for fine-grained population downscaling to the building level, offering substantial contributions to demographic research. This detailed perspective enables a more nuanced analysis of population dynamics and their relationship with spatial distribution. It is particularly beneficial in evaluating the effects of population growth and distribution on urban infrastructure, services, and disaster resilience, thereby informing more strategic urban development plans. Additionally, this integration is crucial for optimizing resource allocation and enhancing risk management, leading to more resilient and sustainable urban environments. Such downscaled population data are invaluable for urban planners in making informed decisions to improve urban livability and adaptability.

Regarding the allocation methodology, the study currently focuses on population and household counts, acknowledging limitations in attribute diversity. Future research directions include exploring scenarios that utilize a broader range of attributes for household and population allocation, representing a multi-objective optimization challenge that adds depth and complexity to the optimization process. The density of residential structures within a census block significantly influences the distribution of housing units and resident populations. Data accuracy issues, such as missing building clusters within a block group, can lead to population underestimation in those areas and overestimation in others within the same census tract. Future research should prioritize enhancing data accuracy in building inventories, potentially leveraging advanced remote sensing techniques to achieve this goal.

CRediT authorship contribution statement

Xinyue Ye: Supervision, Project administration, Methodology, Funding acquisition, Conceptualization. **Weishan Bai:** Software, Methodology, Formal analysis, Data curation, Writing – original draft. **Wenyu Wang:** Software, Methodology, Formal analysis, Data curation, Writing – review & editing, Writing – original draft. **Xiao Huang:** Methodology, Conceptualization, Writing – review & editing. **xinyue ye:** Writing – review & editing, Writing – original draft.

Declaration of interests

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

Data availability

Data will be made available on request.

Acknowledgements

We greatly appreciate the helpful comments and suggestions from the editor and anonymous reviewers. The research was partially supported by National Science Foundation (NSF) under grants 2122054, 2112356, 2232533, and 2235678, as well as Texas A&M University Harold Adams Interdisciplinary Professorship Research Fund, and College of Architecture Faculty Startup Fund. The funders had no role in the study design, data collection, analysis, or preparation of this article.

References

Atwal, K. S., Anderson, T., Pfoser, D., et al. (2022). Predicting building types using OpenStreetMap[J]. *Scientific Reports, 12*(1), 19976.

Bakillah, M., Liang, S., Mobasheri, A., et al. (2014). Fine-resolution population mapping using OpenStreetMap points-of-interest. *International Journal of Geographical Information Science, 28*(9), 1940–1963.

Beasley, E. (2006). *The alleys and back buildings of Galveston: An architectural and social history*. Texas A&M University Press.

Beckman, R. J., Baggerly, K. A., & McKay, M. D. (1996). Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice, 30*(6), 415–429.

Briggs, D. J., Gulliver, J., Fecht, D., et al. (2007). Dasymetric modelling of small-area population distribution using land cover and light emissions data. *Remote sensing of Environ- ment, 108*(4), 451–466.

Chapuis, K., Taillandier, P., Renaud, M., et al. (2018). Gen*: A generic toolkit to generate spatially explicit synthetic populations. *International Journal of Geographical Information Science, 32*(6), 1194–1210.

Chen, J., Fan, W., Li, K., et al. (2019). Fitting Chinese cities' population distributions using remote sensing satellite data. *Ecological Indicators, 98*, 327–333.

Cox, D. R., & Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society: Series B: Methodological, 30*(2), 248–265.

Das, K., Jiang, J., & Rao, J. N. K. (2004). *Mean squared error of empirical predictor*.

Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics, 11*(4), 427–444.

Eicher, C. L., & Brewer, C. A. (2001). Dasymetric mapping and areal interpolation: Implementa- tion and evaluation. *Cartography and Geographic Information Science, 28*(2), 125–138.

Elmer, V., & Fraker, H. (2011). Water, neighborhoods and urban design: micro-utilities and the fifth infrastructure. Water sensitive. *Cities*, 193.

Fereshtehnejad, E., Gidaris, I., Rosenheim, N., et al. (2021). Probabilistic risk assessment of coupled natural-physical-social systems: Cascading impact of hurricane-induced damages to civil infrastructure in Galveston, Texas. *Natural Hazards Review, 22*(3), 04021013.

Gallego, F. J., Batista, F., Rocha, C., et al. (2011). Disaggregating population density of the European Union with CORINE land cover. *International Journal of Geographical Information Science, 25*(12), 2051–2069.

Goodchild, M. F., & Lam, N. S. N. (1980). Areal interpolation: A variant of the traditional spatial problem. *Geo-processing, 1*(3), 297–312.

Harada, T., & Murata, T. (2017). Projecting households of synthetic population on buildings using fundamental geospatial data. *SICE Journal of Control, Measurement, and System Integration, 10*(6), 505–512.

Heris, M. P., Foks, N. L., Bagstad, K. J., et al. (2020). A rasterized building footprint dataset for the United States. *Scientific Data, 7*(1), 207.

Huang, W., Ye, F., Zhang, Y. J., et al. (2021). Compounding factors for extreme flooding around Galveston Bay during hurricane Harvey. *Ocean Modelling, 158*, Article 101735.

Huang, X., Wang, C., Li, Z., et al. (2021). A 100 m population grid in the CONUS by disaggre- gating census data with open-source Microsoft building footprints. *Big Earth Data, 5*, 112–133.

Konduri K C, You D, Garikapati V M, et al. Application of an enhanced population synthesis model that accommodates controls at multiple geographic resolutions. Proceedings of the 95th annual meeting of the transportation research board, Washington, DC, USA. 2016: 10–14.

Leyk, S., Gaughan, A. E., Adamo, S. B., de Sherbinin, A., Balk, D., Freire, S., … Comenetz, J. (2019). The spatial allocation of population: A review of large-scale gridded population data products and their fitness for use. *Earth System Science Data, 11*(3), 1385–1409. Sep 11.

Li, X., & Zhou, W. (2018). Dasymetric mapping of urban population in China based on radiance corrected DMSP-OLS nighttime light and land cover data. *Science of the Total Environment, 643*, 1248–1256.

Liu, X. H., Kyriakidis, P. C., & Goodchild, M. F. (2008). Population-density estimation using regression and area-to-point residual kriging. *International Journal of Geographical Information Science, 22*(4), 431–447.

Mennis, J. (2009). Dasymetric mapping for estimating population in small areas. *Geography Compass, 3*(2), 727–745.

Niu, N., Liu, X., Jin, H., Ye, X., Liu, Y., Li, X., … Li, S. (2017). Integrating multi-source big data to infer building functions. *International Journal of Geographical Information Science, 31*(9), 1871–1890. Sep 2.

Rode, S., Guevara, S., & Bonnefond, M. (2018). Resilience in urban development projects in flood-prone areas: A challenge to urban design professionals. *Town Planning Review, 89*(2), 167–190.

Rosenheim, N., Guidotti, R., Gardoni, P., et al. (2021). Integration of detailed household and housing unit characteristic data with critical infrastructure for post-hazard resilience modeling. *Sustainable and Resilient Infrastructure, 6*(6), 385–401.

Saadi, I., Mustafa, A., Teller, J., et al. (2016). Hidden Markov model-based population synthesis. *Transportation Research Part B: Methodological, 90*, 1–21.

Shahfahad, M. M., Kumari, B., et al. (2021). Indices based assessment of built-up density and urban expansion of fast growing Surat city using multi-temporal Landsat data sets[J]. *GeoJournal, 86*, 1607–1623.

Stal, C., Tack, F., De Maeyer, P., et al. (2013). Airborne photogrammetry and lidar for DSM extraction and 3D change detection over an urban area – A comparative study. *International Journal of Remote Sensing, 34*(4), 1087–1110.

Stathakis, D., & Baltas, P. (2018). Seasonal population estimates based on night-time lights. *Computers, Environment and Urban Systems, 68*, 133–141.

Su, M. D., Lin, M. C., Hsieh, H. I., et al. (2010). Multi-layer multi-class dasymetric mapping to estimate population distribution. *Science of the Total Environment, 408*(20), 4807–4816.

Sun, L., & Erath, A. (2015). A Bayesian network approach for population synthesis. *Transportation Research Part C: Emerging Technologies, 61*, 49–62.

Sun, L., Erath, A., & Cai, M. (2018). A hierarchical mixture modeling framework for population synthesis. *Transportation Research Part B: Methodological, 114*, 199–212.

Tan, M., Li, X., Li, S., et al. (2018). Modeling population density based on nighttime light images and land use data in China. *Applied Geography, 90*, 239–247.

Ural, S., Hussain, E., & Shan, J. (2011). Building population mapping with aerial imagery and GIS data. *International Journal of Applied Earth Observation and Geoinformation, 13*(6), 841–852.

Wan, H., Yoon, J., Srikrishnan, V., et al. (2022). Population downscaling using high-resolution, temporally-rich US property data. *Cartography and Geographic Information Science, 49*(1), 18–31.

Watthanasutthi N, Muangsin V. Generating synthetic population at individual and household levels with aggregate data. 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE). IEEE, 2016: 1–6.

Weber, E. M., Seaman, V. Y., Stewart, R. N., et al. (2018). Census-independent population mapping in northern Nigeria. *Remote Sensing of Environment, 204*, 786–798.

Wu, S., Qiu, X., & Wang, L. (2005). Population estimation methods in GIS and remote sensing: A review. *GIScience & Remote Sensing, 42*(1), 80–96.

Yao, Y., Liu, X., Li, X., et al. (2017). Mapping fine-scale population distributions at the building level by integrating multisource geospatial big data. *International Journal of Geogra- phical Information Science, 31*(6), 1220–1244.

Yang, X., Ye, T., Zhao, N., et al. (2019). Population mapping with multisensor remote sensing images and point-of-interest data. *Remote Sensing, 11*(5), 574.

Ye, X., Konduri, K., Pendyala, R. M., Sana, B., & Waddell, P. (2009). A methodology to match distributions of both household and person attributes in the generation of synthetic populations. In *88th Annual Meeting of the Transportation Research Board*. Washington, DC.

Ye, T., Zhao, N., Yang, X., et al. (2019). Improved population mapping for China using remotely sensed and points-of-interest data within a random forests model. *Science of the Total Environment, 658*, 936–946.

Ye, X., Du, J., Han, Y., Newman, G., Retchless, D., Zou, L., … Cai, Z. (2023). Developing human-centered urban digital twins for community infrastructure resilience: A research agenda. *Journal of Planning Literature, 38*(2), 187–199. May.

Zhang, D., Cao, J., Feygin, S., et al. (2019). Connected population synthesis for transportation simulation. *Transportation Research Part C Emerging Technologies, 103*, 1–16.

Zhao, Y., Ovando-Montejo, G. A., Frazier, A. E., et al. (2017). Estimating work and home population using lidar-derived building volumes[J]. *International Journal of Remote Sensing, 38*(4), 1180–1196.

Zięba-Kulawik, K., Skoczylas, K., Mustafa, A., et al. (2020). Spatiotemporal changes in 3D building density with LiDAR and GEOBIA: A city-level analysis[J]. *Remote Sensing, 12*(21), 3668.