

A three-weight surface modeling approach for optimizing small-scale population disaggregation

Yucheng ZHOU¹, Ling QIN¹, Yirun CHEN¹, Le WANG¹, Xinyan HUANG², Liangfeng ZHU (✉)¹

¹ Key Laboratory of GIScience (Ministry of Education), School of Geographic Sciences, East China Normal University, Shanghai 200241, China
² School of Architecture and Urban Planning, Shenzhen University, Shenzhen 518060, China

© Higher Education Press 2025

Abstract In recent years, fine-scale gridded population data has been widely adopted for assessing and monitoring the Sustainable Development Goals (SDGs). However, the existing population disaggregation techniques struggle to generate precise population grids for small areas with scarce data. To address this, we have introduced a novel, lightweight population gridding technique that integrates dasymetric mapping and point-based surface modeling, titled three-weight surface modeling. This method comprises three weights, each offering a unique perspective on population spatial heterogeneity. The first weight, termed building-volume weight, is equivalent to the preliminary results of assigning population based on building volume data. The second weight, termed POI-center weight, comprises POI (Point of Interest) categories and aggregation patterns, aiming to articulate high-density population centers. It is computed using the neighborhood accumulation rule of Spearman's correlation coefficients between POIs and population size. The third weight, termed POI-distance weight, represents varying decay rates of population with distance from high-density centers. This three-weight surface model facilitates dynamic adjustment of parameters to refine the building-volume weight according to the remaining POI-related weights, thereby generating a more precise population surface. Our analysis of the census population and the disaggregation outcomes from 544 villages in three counties of southern Guizhou Province, China (namely, Huishui, Luodian, and Pingtang) revealed that the three-weight surface model using local parameter groups outperformed individual dasymetric mapping or point-based surface modeling in terms of accuracy. Also, the 10 m population grid generated by this local parameter model (LPTW-POP) presented greater resolution and fewer errors (RMSE of 1109, MAE of 422, and MRE of 0.2630) compared to commonly use gridded

population datasets like LandScan, WorldPop, and GHS-POP.

Keywords population disaggregation, dasymetric mapping, surface modeling, Points-of-Interest (POIs)

1 Introduction

According to the United Nations projection, there is an unprecedented growth in the global population, estimated to reach 8.5 billion by 2030 and expected to increase further to 9.7 billion by 2050 (United Nations, 2021). This dramatic rise in population, coupled with unsustainable consumption and production patterns, has triggered various forms of environmental degradation, including global warming, climate change, deforestation, and biodiversity loss (Zeifman et al., 2022). Hence, an unequivocal comprehension of the present global population situation and trustworthy predictions of future population fluctuations, inclusive of population size, structure, and spatial distribution, are imperative for steering nations toward sustainable development (Weber et al., 2018). Leveraging advancements in earth observation technologies, the production of population distribution datasets has become more efficient, thereby serving as a distinctive data source for computation of regional development indicators, such as the Sustainable Development Goals (SDGs) (Thomson et al., 2022).

The practice of downscaling, also referred to as population disaggregation, entails transforming census data into a more refined population grid. This methodology is extensively employed to recreate a precise representation of the actual population distribution (Eicher and Brewer, 2001; Mei et al., 2022). The core idea of population disaggregation is to establish a correlation between census data and ancillary data, allowing population in large units to be mapped to small grids (Sinha et al., 2019). Due to the availability of

diverse data sources, numerous top-down population disaggregation methodologies have been utilized to construct global gridded population datasets, including WorldPop (Tatem, 2017), LandScan Global Population (Rose et al., 2021), Gridded Population of the World version 4 (GPWv4) (CIESIN, 2018), and GHS-POP (European Commission, 2023).

In recent decades, the development of nearly all population disaggregation tactics has been driven by areal interpolation and its derivatives. This technique facilitates the transfer of socioeconomic data from one source zone to multiple target zones (Goodchild and Lam, 1980). Due to the necessity of preserving volume in the process of population disaggregation, areal interpolation is widely employed directly on source zone units (Lam, 1983). Two primary strategies within area-based areal interpolation are area weighting and pycnophylactic interpolation (Qiu et al., 2022). Area weighting, which presumes a uniform population distribution within the source zone, leverages the intersecting area of the source and target zones to determine the population allocation to the target zone (Goodchild et al., 1993). Pycnophylactic interpolation, assuming population similarity across nearby regions, supplements area weighting with a smoothing step. This approach effectively mitigates abrupt population value changes at source zone edges by applying a weighted average to each grid's nearest neighbors (Tobler, 1979).

Implementing the simple areal interpolation method, which requires minimal ancillary data, presents a challenge in accurately capturing local population differences. To overcome this, dasymetric mapping was innovated as an extension to areal interpolation (Fisher and Langford, 1995; Mennis, 2009). This method employs ancillary data to subdivide the source zone into petite subzones, capable of reflecting spatial population variations. Subsequently, areal interpolation is applied within each of these subzones (Petrov, 2012). The source zone's segmentation approach primarily classifies dasymetric mapping into two types: binary and multi-class. Binary dasymetric mapping typically assigns two subzones to each source zone: populated and unpopulated, with only the populated zones being eligible to receive population assignment from the source zones (Langford and Unwin, 1994). Contrastingly, multi-class dasymetric mapping incorporates ancillary variables like land cover to partition the source zone into multiple subzones (Su et al., 2010). The population proportion in every subzone can be appraised via subjective determination, selective sampling, or statistical modeling (Mennis, 2003; Langford, 2006). With advancements in modeling technology, traditional dasymetric mapping has been upgraded with machine learning and deep learning techniques to form intelligent dasymetric mapping (Mennis and Hultgren, 2006). Backed by an intricate model, intelligent dasymetric mapping has demonstrated a commendable capacity to associate ancillary data with

census data, proving significantly beneficial for population disaggregation (Gervasoni et al., 2018; Šimbera, 2020; Chen et al., 2024).

Surface-based areal interpolation, a salient branch of areal interpolation, aims at constructing a statistical surface with diverse source-zone data (Langford, 2013). By utilizing an adaptive kernel estimator, Bracken and Martin (1989) devised a discrete surface model, which attributed the population of the source zones from population-weighted centroids to the proximate grids. For a precise depiction of the population's spatial distribution, a plethora of address code records disseminated across the source zones were considered for the inaugural population points (Harris and Longley, 2000). Zhang and Qiu (2011) introduced two patterns of population decay with respect to distance from point data encompassing linear and nonlinear; this expanded the prospects for population surface modeling. The growth of Volunteer Geographic Information (VGI) has led to Points-of-Interest (POIs) gaining prominence as a substantial indicator of high population density (Psyllidis et al., 2022). Bakillah et al. (2014) confirmed the possibility of crafting the population surface with superior accuracy by perceiving POIs filtered via the quadtree division as high-density points.

Despite their usefulness, the population disaggregation methods discussed above suffer from certain limitations. First, they struggle to adequately represent the spatial heterogeneity of populations due to traditional dasymetric methods assigning population based on an area-weighting strategy within each subzone. Second, existing point-based surface modeling methods overlook some population-related information, such as point aggregation patterns, and do not account for population variations around different high-density points. Third, intelligent dasymetric methods often require substantial data input which can constrain the achievement of precise results at a small-scale. This study endeavors to devise a small-scale population gridding method that overcomes these limitations and enhances the accuracy of population disaggregation. We introduce the three-weight surface model that fuses basic dasymetric mapping and point-based surface modeling. The model uses three-dimensional building information and Points of Interest (POIs) to bolster the spatial representation of the population. (Ural et al., 2011; Palacios-Lopez et al., 2022). The building-volume weight, the first of the three weights, is derived from population assignment based on building volume information (Alahmadi et al., 2013). The other two variable weights (POI-center weight and POI-distance weight) are constructed based on POI categories and aggregation patterns to dynamically adjust the population fitting process. In conjunction with this, a parameter search strategy aimed at minimizing the Mean Relative Error (MRE) is proposed for each source zone to build the optimal model. Lastly, a quantitative and

comparative assessment of the population grid generated by this model is conducted based on its accuracy.

2 Material and methods

2.1 Study area

Our research focuses on three adjacent counties in the southern part of Guizhou Province, China: Huishui, Luodian, and Pingtang (Fig. 1). This region, characterized by a predominantly mountainous and hilly topography with elevations spanning 242 m to 1691 m, experiences a humid subtropical monsoon climate, characterized by warm and humid conditions. The cumulative area of Huishui, Luodian, and Pingtang is 8290.92 km², with respective values being 2471.84 km², 3013.08 km², and 2806.00 km². The 2020 population data from the Seventh National Population Census Report of China indicates a total of 887846 inhabitants in these counties: Huishui with 395878, Luodian with 257551, and Pingtang with 234417. Figure 1 shows that there are 32 towns in the study area: 11 in Huishui, 10 in Luodian, and 11 in Pingtang, further subdivided into 544 lower-level administrative units, or villages.

2.2 Method outline

The present study progresses sequentially through three

stages, as shown in Fig. 2. These stages comprise data sources and processing, volume-based dasymetric mapping, and three-weight surface modeling. The primary stage involves collecting and processing data from five unique sources. Volume-based dasymetric mapping, the second stage, utilizes building height data to compute the building-volume weight via multi-class dasymetric mapping. The final stage, three-weight surface modeling, begins by generating the last two weights (POI-center weight and POI-distance weight) according to the POI data. Subsequently, the three-weight surface model is constructed using a combined weight derived from the aforementioned three weights. This model assigns the census population to the populated grids located within the source zones.

2.3 Data sources and processing

Five distinct data sources underpinned this study: building outline data, building height data, POI data, census data and administrative boundary data (Table 1). The building outline data was derived from Tianditu Map, a national geo-information service platform. It was initially extracted in meter-scale vector polygon format via threshold segmentation following local storage of vector map tiles. To enhance the alignment with the building outline data's resolution, we favored the CNBH-10 m product (China building height at 10 m resolution) for improved accuracy in building volume calculations (Wu et al., 2023). The Amap Open Platform, an open-source platform offering a broad collection of spatio-temporal data and services to the public, supplied the POI data through its free APIs (Application Programming Interfaces). The Amap POI data utilizes a three-tier classification, containing 23 top categories, 146 intermediate categories, and 902 subcategories. As of 2020, the study area covered 32121 records across 15 top categories such as transportation, healthcare, catering, and more from the POI data. The census data utilized in this study, collected at both village and town levels, were extracted from the Seventh National Population Census report of China, which was published by the National Bureau of Statistics on May 11, 2021. Lastly, the administrative boundary data in vector format, ranging from village to county level, were obtained from the LandCloud Platform of China.

Prior to the population disaggregation within the study area, we implemented a sequence of preprocessing tasks on the acquired data. Initially, we transformed the vector building outline data into a raster building layer, consistent in resolution with the building height data (10 m). Next, to prevent population misallocations, unpopulated zones, including roads and factories, were eliminated from the raster building layer. Owing to the partial lack of building height data relative to the raster building layer, we supplemented the missing locations

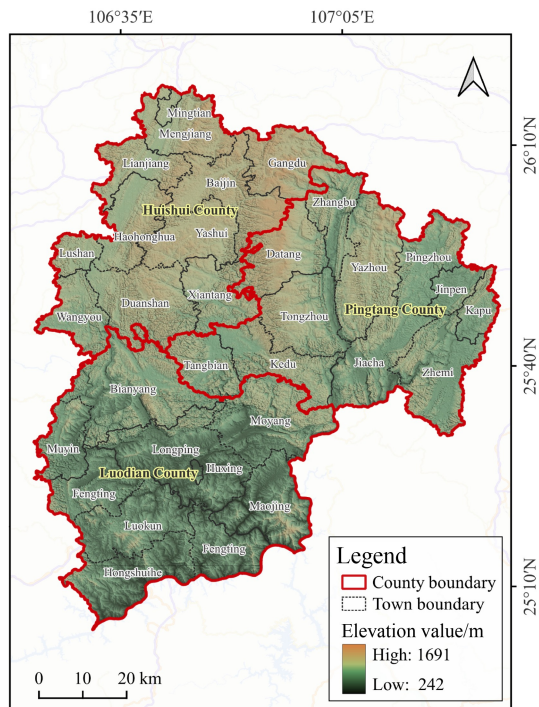


Fig. 1 Overview of the study area. Huishui, Luodian and Pingtang Counties are in Guizhou Province. Spatial distribution of the 32 towns within the study area.

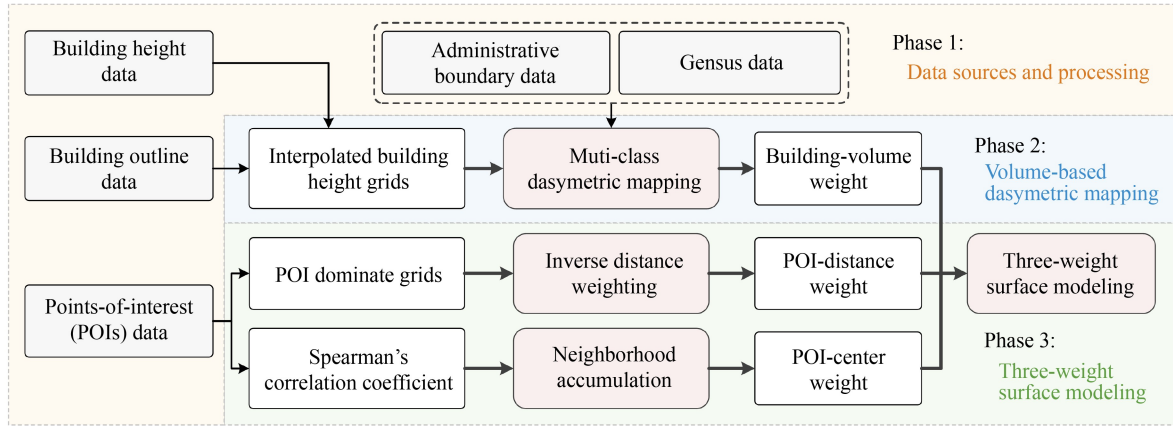


Fig. 2 The framework for the three-weight surface modeling approach proposed in this study.

Table 1 Data used in this case study

Data	Format	Resolution	Year	Source
Building outline data	Polygon features	2 m	2019–2021	Tianditu Map
Building height data	Raster	10 m	2020	CNBH-10 m
POI data	Point features	n.a.	2020	Amap Open Platform
Census data	Table	n.a.	2020	China National Bureau of Statistics
Administrative boundary data	Line features	n.a.	2020	LandCloud Platform

with height values computed by averaging the values of surrounding grids. Finally, for analytical convenience, we reclassified the building height values into three categories (low, medium, and high) by employing the Jenks algorithm (Karunarathne and Lee, 2019).

2.4 Volume-based dasymetric mapping

The application of volume-based dasymetric mapping is an innovative attempt at utilizing ancillary data for multi-class dasymetric mapping. Building height data, serving as the primary ancillary data, is integrated with multi-class dasymetric mapping to facilitate population estimation based on building volume. The refinement of multi-class dasymetric mapping mitigates population misallocations by segmenting the source zone into several subzones, each with a similar population distribution. Multi-class dasymetric mapping entails two population transfers: Firstly, from the source zone to the subzone, and secondly, from the subzone to the target zone (Su et al., 2010). This concept can be algebraically expressed via the multi-class dasymetric mapping formula as shown below:

$$\widehat{P}_t = \sum_{s=1}^S \sum_{h=1}^H \frac{A_{t,s,h} \cdot P_{s,h}}{A_{s,h}} = \sum_{s=1}^S \sum_{h=1}^H A_{t,s,h} \cdot D_{s,h}, \quad (1)$$

where \widehat{P}_t is the estimated population of target zone t ; $A_{t,s,h}$ represents the area of the overlapping region (belonging to building height class h) between the target zone t and source zone s ; $P_{s,h}$ refers to the total population

corresponding to building height class h , assigned from the source zone s ; $A_{s,h}$ is the cumulative area of regions assigned to building height class h within source zone s ; $D_{s,h}$ indicates the relative population density of the regions within building height class h in source zone s ; S stands for the number of source zones; H denotes the number of building height classes.

To ascertain the value of $P_{s,h}$ in Eq. (1), the linear regression model without an intercept (Yuan et al., 1997) is adapted to fit as Eq. (2). To circumvent the issue of negative regression coefficients outlined by Moxey and Allanson (1994), the model's convergence criteria is transitioned from Least Squares Estimation (LSE) to Non-Negative Least Squares Estimation (nnLSE):

$$P_s = \left(\sum_{h=1}^H A_{s,h} \cdot \alpha_h \right) + \varepsilon_0, \quad (2)$$

where P_s represents the actual population of source zone s ; $A_{s,h}$ indicates the area of the region belonging to the building height class h in source zone s ; α_h denotes the coefficient corresponding to the building height class h ; H refers to the number of building height classes; ε_0 is the random error term.

During the population gridding procedure, the population of the source zone is apportioned to an array of small grids, each corresponding to distinctive building height classes. Thus, the area of regions equates to the grid count, expressed as $A_{t,s,h} = N_{t,s,h}$ and $A_{s,h} = N_{s,h}$. Given that the population of a specific grid is presumed to be allocated only from a single source zone s and a particular building height class h (i.e., $N_{t,s,h} = 1$), the ensuing

population of grids can be estimated as follows:

$$\widehat{P}_{g,s,h} = \frac{P_{s,h}}{N_{s,h}} = \frac{P_s \cdot \alpha_h}{N_{s,h} \cdot \sum_{h=1}^H \alpha_h}, \quad (3)$$

where $\widehat{P}_{g,s,h}$ represents the estimated population of the grids with building height class h in the source zone s ; $N_{s,h}$ refers to the quantity of grids with building height class h in the source zone s ; α_h is the linear regression coefficient for building height class h .

2.5 Three-weight surface modeling

The three-weight surface model is constructed utilizing three main components—building-volume weight (w_{volum}), POI-center weight (w_{center}) and POI-distance weight (w_{distance}). The census population within each source zone is apportioned to populated grids based on an adjusted weight (w_{adjusted}). This adjusted weight is a relative measure, calculated from the preceding three weights as illustrated in Eq. (4):

$$w_{\text{adjusted}} = w_{\text{volum}} \cdot w_{\text{center}} \cdot w_{\text{distance}}. \quad (4)$$

The fundamental weight, building-volume weight (w_{volum}), is equivalent to the grid population resulting from volume-based dasymetric mapping. Additionally, a POI conversion rule is defined, transitioning from vector POIs to a POI-center weight grid as designated in Eq. (5). The POI-center weight (w_{center}) is calculated using this rule, aggregating the Spearman's correlation coefficient (as outlined in Eq. (6)) across various POI categories within neighboring grids:

$$w_{\text{center}}^{(row,col)} = \sum_{c=POI_Category}^{S_POIs(k,row,col)} \rho_c, \quad (5)$$

$$\rho_c = \frac{\sum_{s=1}^S (P_s - \bar{P}) \cdot (N_{cs} - \bar{N}_c)}{\sqrt{\sum_{s=1}^S (P_s - \bar{P})^2 \cdot \sum_{s=1}^S (N_{cs} - \bar{N}_c)^2}}, \quad (6)$$

where $S_POIs(k,row,col)$ represents all POIs that are searched within k rings (8-connected fields) around the grid at position (row,col) ; ρ_c is the Spearman's correlation coefficient between POI category c and census population; P_s stands for the actual population in source zone s ; \bar{P} denotes the average population of all source zones; N_{cs} refers to the number of POIs belonging to the category c in source zone s ; \bar{N}_c is the average number of POIs belonging to the category c ; S is the number of source zones.

POIs serve as crucial indicators of population density centers in this study, with the distance from these POIs delineating the extent of their influence on the population of nearby locations. POI-distance weight (w_{distance}) signifies the adjustments of population fitting at a specific location due to its proximity to the nearest POI (Zhang

and Qiu, 2011). This relationship is defined by an inverse distance weighting function, as illustrated below:

$$w_{\text{distance}}^{(row,col)} = \left(1 - \frac{\lambda_{(row,col)}}{\lambda_{\text{max}}}\right)^q \quad (q > 0), \quad (7)$$

where $\lambda_{(row,col)}$ is the distance of grid at position (row,col) from its nearest POI; The maximum global value of $\lambda_{(row,col)}$ in the study area is represented by λ_{max} ; q denotes the distance decay term.

2.6 Accuracy assessment

In this study, the accuracy of the population grid derived from the three-weight surface model is evaluated at the village scale using root mean square error (RMSE), mean absolute error (MAE), and mean relative error (MRE) (Cartagena-Colón et al., 2022). RMSE is computed by taking the square root of the average of the squared differences between the actual and estimated population across all villages, as demonstrated in Eq. (8). MAE, which represents the mean of the absolute errors, ensures the interpretability of the accuracy assessment, thanks to its consistency with the census population units, as illustrated in Eq. (9). MRE, which highlights the average percentage of absolute errors in relation to the actual values, offers a smaller interval and improved resilience to outliers (Pavía and Cantarino, 2017; Li and Zhou, 2018). The calculation for MRE is provided in Eq. (10):

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \quad (8)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|, \quad (9)$$

$$\text{MRE} = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{y_i}, \quad (10)$$

where y_i represents the census population in village i ; \hat{y}_i denotes the estimated population in village i ; n stands for the total number of villages.

3 Results

The three-weight surface model provides a flexible control mechanism over the population gridding process by adjusting two parameters (q in w_{distance} and k in w_{center}). This study experimented with five discrete values of the parameter k to generate POI-center weight grids, hereinafter referred to as w_{center} grid. Figure 3 displays the spatial distribution of POIs within a specified region, accompanied by five w_{center} grids derived using different values of parameter k . Upon setting k at 0, the Spearman's correlation coefficients, corresponding to individual POI categories, were solely integrated into the grids occupied

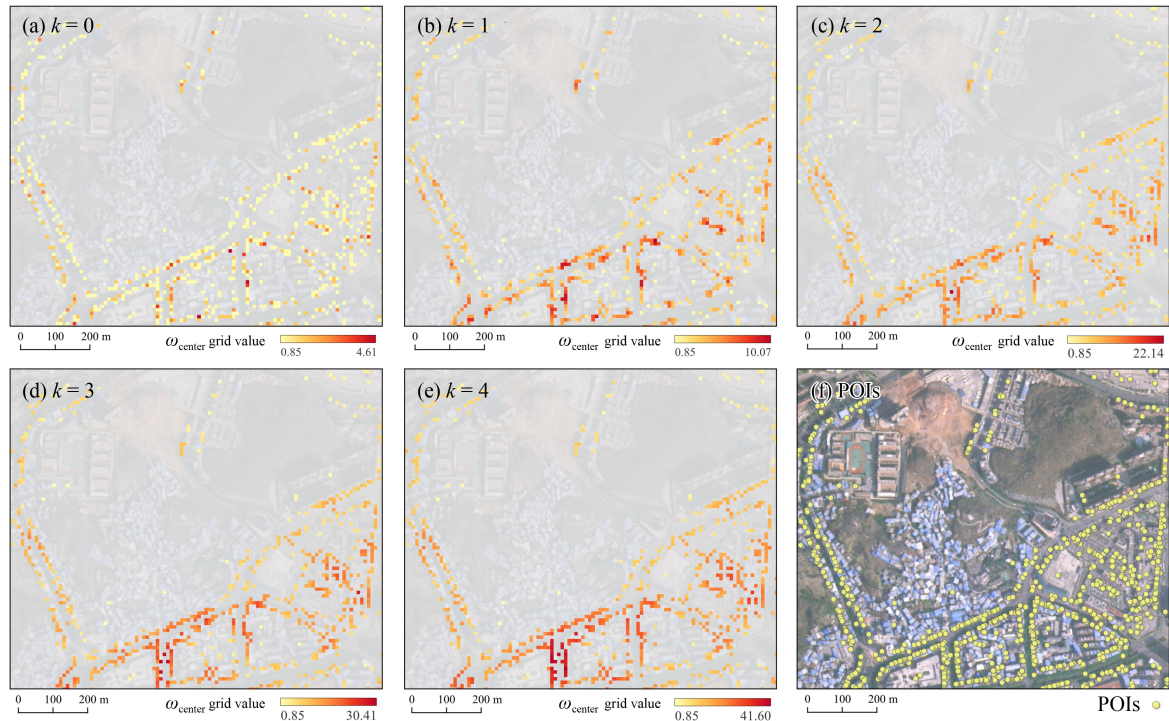


Fig. 3 Five w_{center} grids generated using parameter k values ranging from 0 to 4 and POIs' spatial distribution in a certain region. (a) The w_{center} grid generated when $k = 0$; (b) The w_{center} grid generated when $k = 1$; (c) The w_{center} grid generated when $k = 2$; (d) The w_{center} grid generated when $k = 3$; (e) The w_{center} grid generated when $k = 4$; (f) The spatial distribution of POIs.

by POIs, thereby rendering similar w_{center} grid value (Fig. 3(a)). When k was set at 1, POI-rich regions manifested remarkably higher values within the w_{center} grid than those exhibited by POI-deficient regions (Fig. 3(b)). A setting of k at 2 led to smoother w_{center} grid values, as the expansion of the accumulative range augmented the size of the high-value regions (Fig. 3(c)). Similar distributions of w_{center} grid value were observed at k values of 3 and 4, although exhibiting a smoother appearance but with a broader disparity between the maximum and minimum values as k increased (Figs. 3(d) and 3(e)). It is discernible that the adoption of the neighborhood accumulation rule engenders more prominent w_{center} grid values within regions characterized by a higher concentration of POI. These variations underscore the capability of the w_{center} (POI-center weight) to depict POI aggregation patterns across multiple scales aided by the variable parameter k .

Another adjustable component in the three-weight surface model is the parameter q in w_{distance} . In this research, we selected a series of q -values ranging between 0 and 10 in increments of 0.2. These q -values were then paired in groups with the aforementioned five discrete k -values. To evaluate the model's performance within the study area, we generated a series of population surfaces by utilizing all paired parameter groups $\{(k, q)\}$. Three accuracy metrics, RMSE, MAE, and MRE, were computed to examine the variance in accuracy among models constructed with different enumerated parameter

groups. Figure 4 presents five accuracy curves corresponding to each metric. Each curve illustrates the fluctuation of the model's overall accuracy in relation to q , while maintaining a constant value for k . Two salient trends can be discerned in the RMSE, MAE, and MRE as q increases: they either undergo an initial decline succeeded by a surge or simply experience an unabated increase. This tendency might stem from the deficient fitting of the three-weight surface model when the parameter group $\{(k, q)\}$ comprising k and q adopt extreme values. In contrast, employing moderate values for parameters k and q could potentially enhance the population modeling, leading to dips in the error curve. The subsequent analysis scrutinized the impact of varying the q value on the overall accuracy, with the k value held constant. Upon equating k to 0, the RMSE demonstrated an initial decline before ascending slowly at q equals to 6.4 with augmenting q values. Both the MAE and MRE exhibited a comparable trajectory, shifting from a decreasing to an increasing trend, with their respective turning points transpiring at q equals 0.8. Upon setting k equal to 1, the RMSE noticeably lowered relative to its value at k equal to 0. Its changing pattern remained, despite the turning point shifting to q being 2.4. Notably, the MAE and MRE did not consistently stay beneath the levels recorded when k equals 0. Compared to the condition where k equals 0, the MAE decreased just prior to the point where q equals 4.2, while the MRE performed marginally higher subsequent to q equals 0.2.

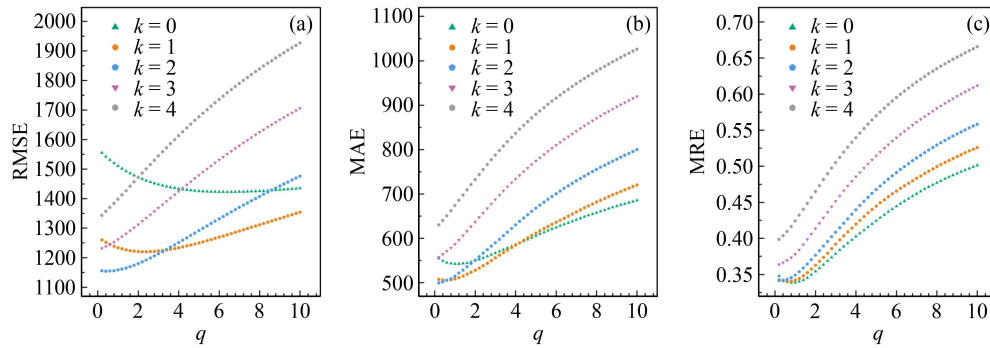


Fig. 4 The overall accuracy metrics for models constructed using different parameter groups $\{(k,q)\}$ within the study area. (a) The overall RMSEs; (b) The overall MAEs; (c) The overall MREs.

This outcome could probably be attributed to the susceptibility of RMSE to outliers, indicating that the RMSE is more vulnerable to extreme values, thus leading to rapid fluctuations. Upon reaching k values of 2, 3, and 4, the RMSE, MAE, and MRE all exhibited an escalating trend in relation to increasing q values. Moreover, there was no convergence among the three k -curves for k equals 2, 3, and 4 due to the consistent ascension of the metrics with the rise in k values. As per the displayed results, the RMSE established its lowest global point at k being 2 and q at 0.4, while both the MAE and MRE demonstrated their lowest global points at k equating to 2 and q equivalent to 0.2. Given the superior robustness of MAE and MRE to outliers, the optimal parameter group was determined to be $(k = 2, q = 0.2)$. This suggests that the three-weight surface model, built with this optimal parameter group, possesses the highest overall accuracy within the study area.

To explore the potential for improved model accuracy afforded by the use of a unique parameter group within each source zone, we implemented a strategy that minimized the MRE at the town level to identify the ideal parameter group for each town. This approach provides the possibility to iteratively optimize local parameter groups for enhanced performance of town-level models. Figure 5 illustrates the relative differences in RMSE, MAE, and MRE at the town level, computed separately for two models, one utilizing the optimal local parameter groups and the other using the global parameter group. Negative values in Figs. 5(a), 5(b), 5(c) indicate that the local parameter model yields better accuracy performance, or lower accuracy metrics, compared to the global parameter model. As depicted in Fig. 5(a), upon transitioning from the global parameter model to the local parameter model, there's a noticeable variation in the town-level RMSE. Out of the 32 towns studied, the RMSE observed a decline in 21 towns, an increase in 9, and remained consistent in one. The mean reduction in RMSE amounts to 28.19%, while the average upsurge is recorded at 13.33%. This result underscores the beneficial implication of employing local parameter groups toward minimizing the RMSE in a greater part of the towns. As

demonstrated in Fig. 5(b), an observable decrease in the MAE transpired in 27 of the 32 towns, reflecting an average decrease of 26.43%. A minor group of four towns reported an increase in the MAE, generating an average escalation of 12.48%. Interestingly, the reduction in MAE pervades across more towns in comparison to the RMSE, even though the MAE's decrease proves marginally smaller when utilizing local parameter groups in contrast with the RMSE. Analyzing Fig. 5(c) reveals that, except for one that remained unchanged, the other 31 towns underwent various reductions in the MRE. The overall impact on town-level MRE reduction is notable, showing an average decrement of 25.67%. Evidently, a strategy that focuses on minimizing the town-level MRE yields the most substantial improvement in the MRE. The minimum MREs for each town, obtained by the iterative adjustment of parameters k and q , are exhibited in Fig. 5(d). This indicates that the model can efficiently control the values of MRE, as illustrated by a mere six towns registering MRE values exceeding 0.3. Evidently, the precision of the three-weight surface model at the town-level tends to improve, generally, by applying local parameter groups rather than a single global parameter group. However, fluctuations in the accuracy differences concerning the RMSE and MAE are observed. The RMSEs of nine towns and MAEs of four towns unexpectedly escalated after implementing local parameter groups. Even with a universal decline in MRE for all towns, the rise in RMSE and MAE for a handful of towns indicates that the local parameter model could potentially introduce new outliers affecting the assessment of RMSE and MAE at town-level accuracy. This finding further proves the reliability of using MRE as the benchmark for parameter selection.

This study undertook a comparative experiment to validate the supremacy of the three-weight surface modeling over its component methodologies, volume-based dasymetric mapping and point-based surface modeling, in the domain of population disaggregation within the designated study area. The five population disaggregation techniques encompassed within this experiment are delineated in Table 2. Method 1 applies

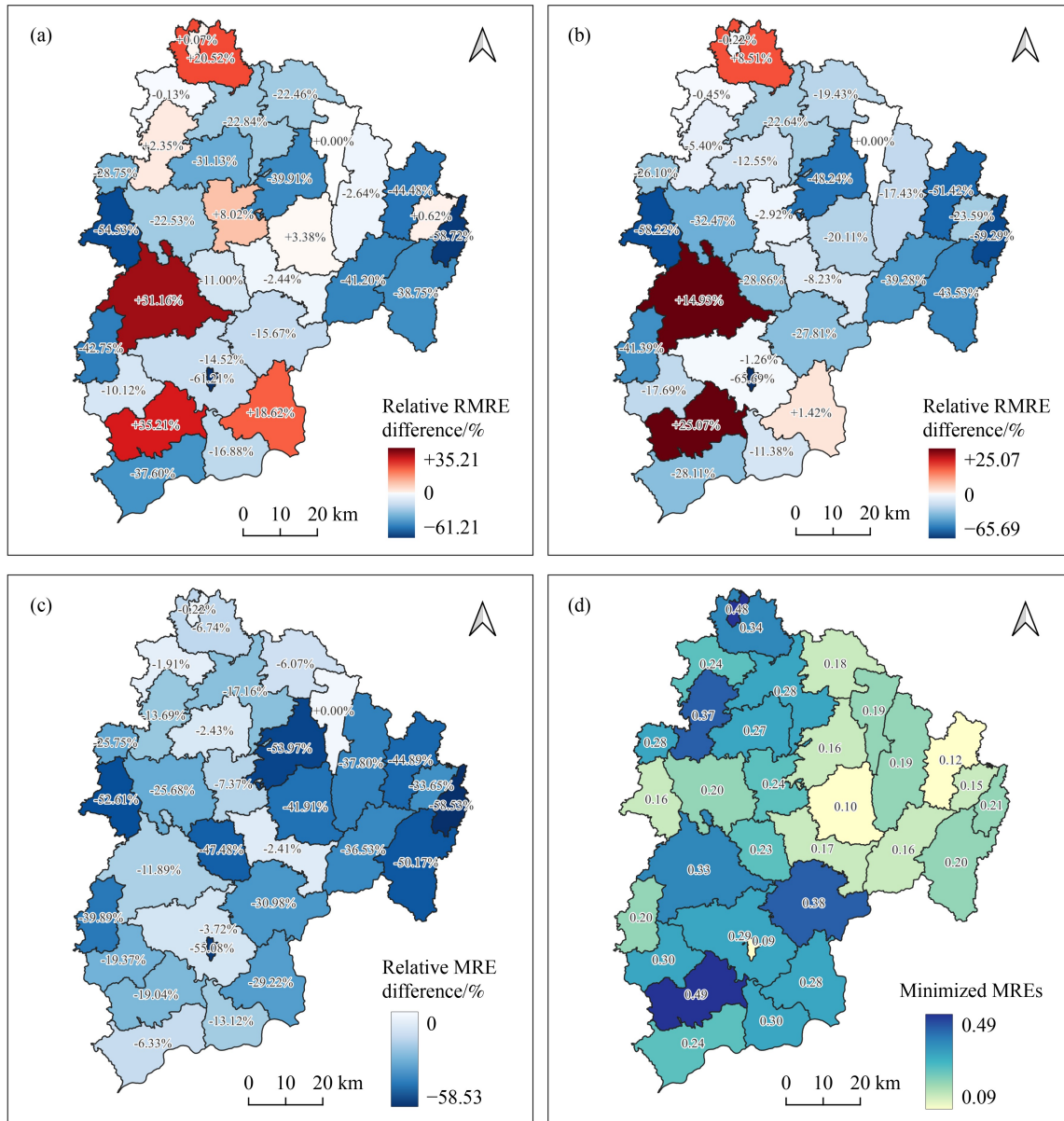


Fig. 5 Accuracy differences at town-level between the three-weight surface models constructed using the optimal local and global parameter groups. (a) Relative RMSE differences. (b) Relative MAE differences. (c) Relative MRE differences. (d) Minimized MREs.

volume-based dasymetric mapping, using linear regression to evaluate the population within grid cells associated with each building volume class. Method 2, also known as point-based surface modeling with a global parameter q , creates a population surface entirely dependent on the distance from POIs (similar to w_{distance}). This method employs the same optimal decay term q universally across all towns. Method 3 is similar to Method 2, with the only distinction being it adopts a different optimal parameter q specific to each town. Method 4 and Method 5, both three-weight surface modeling approaches, are devised by augmenting and applying the volume-based dasymetric mapping with point-based surface modeling. In Method 4, a uniform

Table 2 Population disaggregation methods compared in this study

Method	Description
Method 1	Volume-based dasymetric mapping
Method 2	Point-based surface modeling (global q)
Method 3	Point-based surface modeling (local q)
Method 4	Three-weight surface modeling (global k and q)
Method 5	Three-weight surface modeling (local k and q)

parameter group utilized for all towns is determined by minimizing the overall MRE. While in Method 5, distinctive parameter groups specific to each town are determined by minimizing the town-specific MREs, allowing the creation of a unique population surface for

each town.

Table 3 showcases three overall accuracy metrics (RMSE, MAE, and MRE), along with the percentage of villages (target zones) where the MRE remained beneath 0.1, 0.3, and 0.5 thresholds. Evidently, Method 2 consistently outperforms Method 1 by producing lower values for RMSE, MAE, and MRE, which respectively imply differences of 44, 33, and 0.0259. In addition, Method 2 displays a superior concentration in village-level MREs toward lower values, indicated by an increased percentage of villages with MRE below 0.1, 0.3, and 0.5 by 0.98%, 2.54%, and 3.72% respectively, compared to Method 1. Between the two point-based surface modeling methods, Method 3 outperforms Method 2 by yielding lower RMSE, MAE, and MRE. Interestingly, Method 2 outdoes Method 3 by possessing a higher proportion of villages with an MRE below 0.3 and 0.5. Amid the unique strengths and weaknesses both Method 2 and Method 3 offer in relation to accuracy metrics, this compelling observation accentuates the superior performance of point-based surface modeling over volume-based dasymetric mapping in the study area. Concerning Method 4, the three-weight surface model, constructed using the optimal global parameter group, has achieved a RMSE of 1155 and a MAE of 500, which are lower than those of the three aforementioned methods. Nevertheless, the MRE of Method 4 is roughly 0.04 higher than that of Method 3 and 0.02 higher than Method 2. This discrepancy could potentially be due to the misfit of the global parameter group for all towns. While particular regions fitting poorly may detriment the overall accuracy, their impacts may not necessarily be reflected in the RMSE and MAE, possibly veiled by outliers. Method 5 uses a set of optimal local parameter groups to construct the three-weight surface model. It can be clearly observed that its RMSE, MAE and MRE are

lower compared to Method 4 (46, 78, and 0.0784 lower, respectively). Furthermore, among all experimental methods, Method 5 records the highest proportion of villages (29.94%, 66.73%, and 87.28%) with MREs below 0.1, 0.3, and 0.5. The empirical evidence demonstrates a significant enhancement in the precision of the three-weight surface model when local parameter groups are utilized. According to these results, the application of local parameter groups to the three-weight surface model confers greater accuracy relative to alternative methods of population disaggregation.

Table 4 reveals the overall accuracy metrics of the population grid that was developed using the local parameter three-weight surface model (LPTW-POP), in contrast with other gridded population datasets, such as LandScan, WorldPop, and GHS-POP. The accuracy of LandScan is clearly the lowest, as evidenced by a MRE of 0.6494, with over 42% of villages confronting an MRE of 50% or greater. Although WorldPop's accuracy is inferior to that of GHS-POP, as indicated by measurements of RMSE, MAE, and MRE, its village-level MREs tend to cluster toward lower values. This clustering tendency could be a reflection of WorldPop's error instability, which results in a small number of villages with high errors that inflate the overall values of RMSE, MAE, and MRE. Compared to the aforementioned three open-source datasets, LPTW-POP demonstrates superior accuracy. The RMSE, MAE, and MRE of LPTW-POP are approximately 52.34%, 56.19%, and 48.91% of those recorded for WorldPop. Furthermore, the metric values for LPTW-POP are considerably lower compared to GHS-POP, demonstrating a decrease by 27.52%, 37.94%, and 43.39% for RMSE, MAE, and MRE, respectively.

Figure 6(a) illustrates the 10 m LPTW-POP within the designated study area. Uninhabited zones were notably excluded and void of any population assignment. Within

Table 3 Overall accuracy performance of five population disaggregation methods in the study area

Method	RMSE	MAE	MRE	MRE<0.5	MRE<0.3	MRE<0.1
Method 1	1655	573	0.3477	81.21%	63.41%	25.05%
Method 2	1611	540	0.3218	84.93%	65.95%	26.03%
Method 3	1511	513	0.3041	84.54%	65.75%	28.18%
Method 4	1155	500	0.3414	83.37%	56.16%	18.40%
Method 5	1109	422	0.2630	87.28%	66.73%	29.94%

Table 4 Overall accuracy performance of the population grid generated by the local parameter three-weight surface model and other gridded population datasets

Method	RMSE	MAE	MRE	MRE<0.5	MRE<0.3	MRE<0.1
LandScan	2398	984	0.6494	57.34%	36.99%	13.50%
WorldPop	2119	751	0.5377	71.62%	53.62%	17.22%
GHS-POP	1530	680	0.4646	61.84%	40.90%	14.29%
LPTW-POP¹⁾	1109	422	0.2630	87.28%	66.73%	29.94%

Note: 1) The population grid generated by the local parameter three-weight surface model proposed in this study.

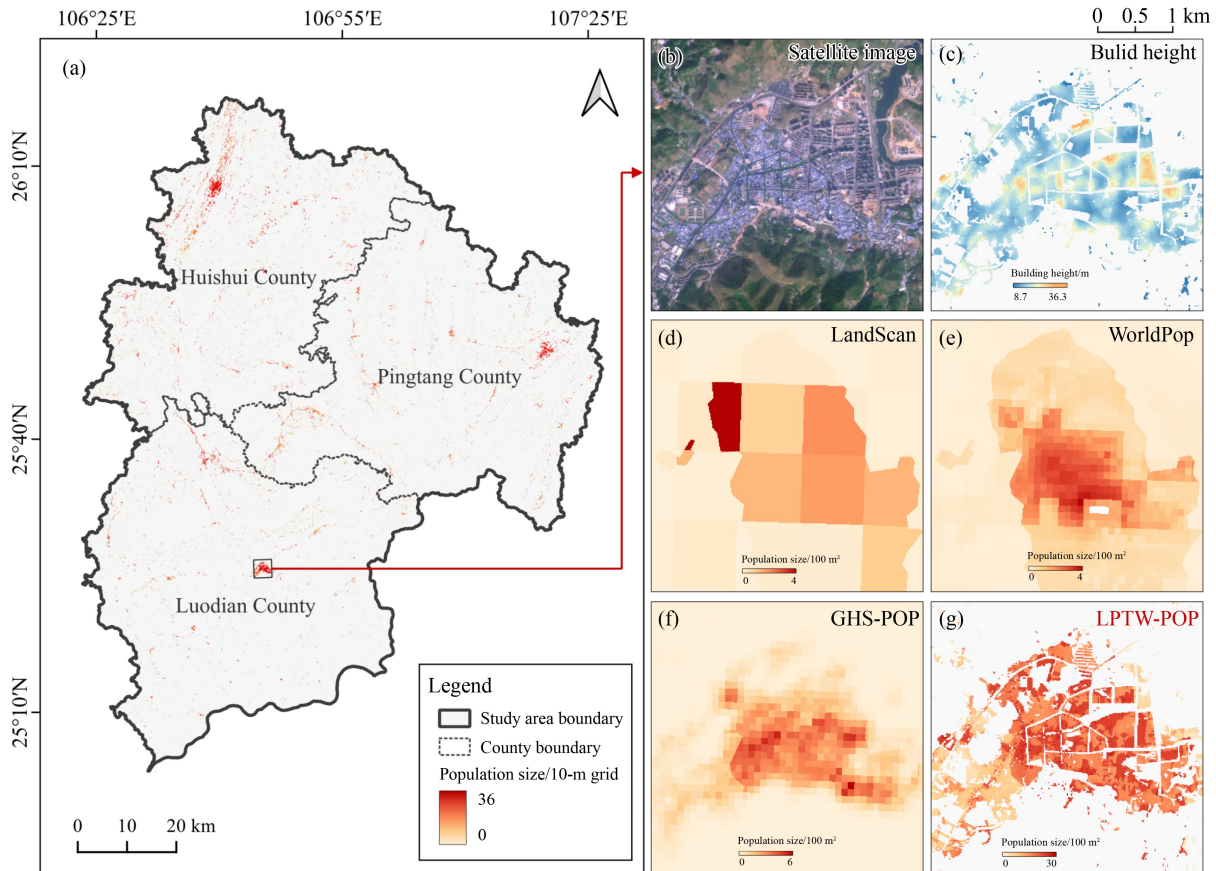


Fig. 6 LPTW-POP in the study area and the comparison of population distribution in the partial area from different datasets. (a) LPTW-POP in the study area; (b) Satellite image in the partial area; (c) Build height grid in the partial area; (d) LandScan in the partial area; (e) WorldPop in the partial area; (f) GHS-POP in the partial area; (g) LPTW-POP in the partial area.

the LPTW-POP grid of the analyzed area, population count significantly fluctuates, ranging from 0 to 36 across various locations. Figure 6(b) presents a satellite image of a specific region, whereas Fig. 6(c) portrays the associated building heights displayed in a grid format. To facilitate deeper analysis of differences among the gridded population datasets, the grid values across the three datasets, LandScan, WorldPop, and GHS-POP, were standardized. This ensured a uniform representation of population size in units per 100 m². Upon visually comparing Figs. 6(d) and 6(e), it becomes clear that WorldPop offers a considerably greater resolution and notably more accurate depiction of population distribution than LandScan. Nonetheless, both WorldPop and LandScan display population densities between 0 and 4 persons per 100 m², indicating considerable underestimation within this region. Regarding the GHS-POP, it is possible to observe the spatial outlines of urban areas from the population grid values (Fig. 6(f)). The maximum density value signified by 6 underpins the notion that GHS-POP tends to align more closely to the actual population distribution than the previously mentioned datasets. Figure 6(g) demonstrates that LPTW-POP encapsulates a broader range of spatial population

information, closely aligning with the outlines of the built-up areas. Additionally, within this region, LPTW-POP attains its maximum density value of 30, which corresponds with the building height measurements. Hence, this characteristic seems to render LPTW-POP more credible. The findings indicate that the incorporation of building height constraints mitigates issues commonly observed in other datasets, such as the conservative estimation of population figures in urban areas. These results also validate the ability of the three-weight surface modeling method to produce accurate gridded population data that are closer to the actual population situation in a small area such as the study area.

4 Discussion

This section analyzes the advantages of the three-weight surface modeling in contrast to other methods for small-scale population disaggregation. The mechanism by which the three weights (building-volume weight, POI-center weight, and POI-distance weight) collectively contribute to the development of an enhanced accuracy population surface, is also uncovered. We have

subsequently underscored some aspects in which the three-weight surface model can be further optimized, including model stability, parameter tuning, accuracy evaluation and data utilization.

When dealing with population disaggregation in small areas, the scarcity of data poses a challenge to the effective application of intelligent methodologies such as spatial regression for achieving an optimal fitting (Monteiro et al., 2019). However, traditional population gridding methods fall short in characterizing the spatial heterogeneity of regional population. In simple dasymetric mapping, the subzones' population is presumed to follow a uniform distribution, which doesn't always conform to the actual population situation (Baynes et al., 2022). In point-based surface modeling, the unique roles of different points in population estimations have not been adequately acknowledged. This modeling strategy employs a static attenuation pattern for the population surrounding various high-density points (Zhang and Qiu, 2011).

Based on the outcomes of the experiments conducted in this research, the flexibility of population estimation is heightened within the three-weight model due to the inherent variability of its multi-weight structure. The product of volume-based dasymetric mapping, denoted as the first weight w_{volume} , furnishes the model with a foundational population assignment predicated on building volume. This model's dynamic explanatory capacity is further enhanced by the additional two weights, w_{center} and w_{distance} , both of which are computed using POIs. The w_{center} reflects the simulation based on various POI categories and aggregation patterns related to population distribution. In computing the w_{center} , Spearman's correlation coefficients delineate the association of various categories of POIs with the population (Bakillah et al., 2014). The aggregation patterns of POIs are embodied in the correlation coefficients' neighborhood accumulation rule. In a region dense with POIs, w_{center} grid values associated with the center tend to aggregate based on the correlation coefficients stemming from a more substantial number of surrounding POIs. These aggregated values end up being higher than those in regions sparse with POIs. Viewed from another perspective, the disparity in w_{center} grid values serves as an indirect indicator of the population imbalance between urban and rural areas. This correlation can be attributed to the fact that POIs tend to cluster densely in urban regions while being more thinly spread in rural ones. The parameter k is designed to dynamically capture the disparity by altering the gap value of w_{center} between urban and rural areas. The POI-distance weight, w_{distance} , is developed to portray the impact of w_{center} grids on their corresponding surrounding grids. Leveraging the inverse distance weighting function, the parameter q in w_{distance} can effectively represent different decay rates, which dictate how the influence of w_{center} grids

diminishes with increasing distance (Langford, 2013). The results of this study have verified the enhanced fitting proficiency of the surface model with this three-weight integration, resulting in better population disaggregation accuracy. Two adjustable parameters k and q , enable the model to search iteratively for the most optimal population surface model within each source zone. The inadequacy of the global parameter model reveals that different source zones might exhibit unique characteristics in population distribution. Conversely, the local parameter model has the capacity to develop a distinct fitting for each source zone. This feature is deemed pivotal for enhancing performance in the three-weight surface model.

Nonetheless, there are several aspects of this study that need further optimization. First, a discrete neighborhood range is selected to compute w_{center} , due to the discrete nature of parameter k . In this instance, it is not possible to seamlessly convert the aggregation patterns of the POIs to w_{center} values, resulting in the incapability of identifying a truly optimal parameter k that perfectly aligns with the population distribution. Hence, improvements in the calculation of w_{center} can be realized through the application of the weighted accumulation method. Prior to their accumulation to grids, the correlation coefficients are multiplied by positional weights calculated using a smoothing inverse distance function like the Gaussian function. This process consequently modifies the role of parameter k to function arguments, thereby achieving a better fitting through iterative adjustments of these arguments. Second, integrating the weights in a multiplicative manner, as presented in Eq. (4), could lead to stability issues in the three-weight surface model. Extreme values of w_{volume} might not be rectifiable by w_{center} and w_{distance} , thereby leading to substantial errors in some regions. Consequently, there is a need to devise more effective weight combination methods to enhance the model's robustness. Third, the accuracy assessment system employed in the present study is not comprehensive, thereby resulting in a deficient detection of local outliers. Enhancing the outlier detection capability can potentially be achieved by computing the error variance for each source zone.

Furthermore, there exist opportunities to expand upon the present work. The three-weight surface model, along with its iterative optimization procedure, harbors the potential for extrapolation to broader regions. We conducted population disaggregation within a limited scope (three counties) through the application of census data at both town and village levels. Census data at the village scale is utilized in both the optimization of parameters and the verification of accuracy. Consequently, by upgrading these two tiers of census data, the three-weight surface model can be potentially applied on a more extensive scale. Moreover, it is crucial to explore the relationship between the parameter groups

$\{(k,q)\}$ and the population in source zones to construct a more universal model that eliminates the need for parameter iteration based on second-level census data. The potential for future enhancements in the three-weight surface model, particularly with regards to data utilization, is significant. Building height data accrued from LiDAR, for instance, can provide increased accuracy even in smaller areas. By utilizing advanced deep learning methods for segmenting remote sensing images, an elevated level of timeliness related to building outline data can be achieved. Lastly, comprehensive population fitting could be augmented by incorporating multi-source spatial data products, such as road network data, digital elevation data, and nighttime light data, into the weight calculation of the three-weight surface model.

5 Conclusions

This study introduces a lightweight method, named three-weight surface modeling, appropriate for small-scale population gridding. This method illustrates the spatial heterogeneity of the population through three key weights: building-volume weight, POI-center weight and POI-distance weight. The incorporation of adjustable parameters from the two POI-associated weights bestows upon the three-weight surface model a more robust capability for population modeling than existing methods, such as dasymetric mapping and point-based surface modeling. The strategy of minimizing MRE for parameter search has proven effective in constructing the optimal fitting surface. Upon comparing evaluation metrics of various population disaggregation methods within the study area (three counties in southern Guizhou Province, China, named Huishui, Luodian, and Pingtang), it is apparent that the three-weight surface model using local parameter groups holds a significant advantage. This model achieves an overall RMSE, MAE, and MRE of 1109, 422, and 0.2630, respectively, outperforming common methods by a substantial margin. Furthermore, by utilizing the higher resolution of building outline data and building height data, the model derived 10-m population grid (LPTW-POP) is capable of representing a more detailed population distribution than prevalent gridded population datasets such as LandScan, WorldPop, and GHS-POP. Moreover, the three-weight surface model stands to have its applicability extended to more regions, given the availability of accessible and wide-coverage ancillary data, an area ripe for ensuing research.

CRedit authorship contribution statement

Yucheng Zhou: Conceptualization, Data curation, Formal analysis, Methodology, Resources, Validation, Visualization, Writing – original draft, Writing – review and editing. Ling Qin: Conceptualization, Data curation, Formal analysis, Methodology, Resources, Validation, Writing – review and editing. Yirun Chen: Conceptualization, Data curation, Formal analysis, Methodology, Resources, Visualization. Le Wang:

Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources. Xinyan Huang: Data curation, Investigation, Methodology, Resources, Visualization. Liangfeng Zhu: Conceptualization, Formal analysis, Funding acquisition, Methodology, Resources, Supervision.

Acknowledgments We acknowledge the support of the Natural Science Foundation of Shanghai Municipality (No. 24ZR1420500), the Project of Yulin Science, and Technology Light (No. 2024-KJZG-KXJ-005), and the Project of International Research Center of Big Data for SDGs (No. CBAS2024SDG005).

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

References

- Alahmadi M, Atkinson P, Martin D (2013). Estimating the spatial distribution of the population of Riyadh, Saudi Arabia using remotely sensed built land cover and height data. *Comput Environ Urban Syst*, 41: 167–176
- Bakillah M, Liang S, Mobasheri A, Jokar Arsanjani J, Zipf A (2014). Fine-resolution population mapping using OpenStreetMap points-of-interest. *Int J Geogr Inf Sci*, 28(9): 1940–1963
- Baynes J, Neale A, Hultgren T (2022). Improving intelligent dasymetric mapping population density estimates at 30 m resolution for the conterminous United States by excluding uninhabited areas. *Earth Syst Sci Data*, 14(6): 2833–2849
- Bracken I, Martin D (1989). The generation of spatial population distributions from census centroid data. *Environment and Planning A: Economy and Space*, 21(4): 537–543
- Cartagena-Colón M, Mattei H, Wang C (2022). Dasymetric mapping of population using land cover data in JBNERR, Puerto Rico during 1990–2010. *Land (Basel)*, 11(12): 2301
- Chen Y, Xu C, Ge Y, Zhang X, Zhou Y (2024). A 100 m gridded population dataset of China's seventh census using ensemble learning and big geospatial data. *Earth Syst Sci Data*, 16(8): 3705–3718
- CIESIN (2018). Gridded Population of the World, Version 4 (GPWv4): Population Count, Revision 11. Palisades, New York: NASA Socioeconomic Data and Applications Center (SEDAC)
- Eicher C L, Brewer C A (2001). Dasymetric mapping and areal interpolation: implementation and evaluation. *Cartogr Geogr Inf Sci*, 28(2): 125–138
- European Commission (2023). GHSL data package 2023. Publications Office of the European Union, Luxembourg
- Fisher P F, Langford M (1995). Modelling the errors in areal interpolation between zonal systems by Monte Carlo simulation. *Environment and Planning A: Economy and Space*, 27(2): 211–224
- Gervasoni L, Fenet S, Perrier R, Sturm P (2018). Convolutional neural networks for disaggregated population mapping using open data. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)
- Goodchild M F, Anselin L, Deichmann U (1993). A framework for the areal interpolation of socioeconomic data. *Environment and Planning A: Economy and Space*, 25(3): 383–397

- Goodchild M F, Lam N (1980). Areal interpolation: a variant of the traditional spatial problem. *Geo-Processing*, 1(3): 297–312
- Harris R J, Longley P A (2000). New data and approaches for urban analysis: modelling residential densities. *Trans GIS*, 4(3): 217–234
- Karunaratne A, Lee G (2019). Estimating hilly areas population using a dasymetric mapping approach: a case of Sri Lanka's highest mountain range. *ISPRS Int J Geoinf*, 8(4): 166
- Lam N S N (1983). Spatial interpolation methods: a review. *Am Cartogr*, 10(2): 129–150
- Langford M (2006). Obtaining population estimates in non-census reporting zones: an evaluation of the 3-class dasymetric method. *Comput Environ Urban Syst*, 30(2): 161–180
- Langford M (2013). An evaluation of small area population estimation techniques using open access ancillary data. *Geogr Anal*, 45(3): 324–344
- Langford M, Unwin D J (1994). Generating and mapping population density surfaces within a geographical information system. *Cartogr J*, 31(1): 21–26
- Li X, Zhou W (2018). Dasymetric mapping of urban population in China based on radiance corrected DMSP-OLS nighttime light and land cover data. *Sci Total Environ*, 643: 1248–1256
- Mei Y, Gui Z, Wu J, Peng D, Li R, Wu H, Wei Z (2022). Population spatialization with pixel-level attribute grading by considering scale mismatch issue in regression modeling. *Geo Spat Inf Sci*, 25(3): 365–382
- Mennis J (2003). Generating surface models of population using dasymetric mapping. *Prof Geogr*, 55(1): 31–42
- Mennis J (2009). Dasymetric mapping for estimating population in small areas. *Geogr Compass*, 3(2): 727–745
- Mennis J, Hultgren T (2006). Intelligent dasymetric mapping and its application to areal interpolation. *Cartogr Geogr Inf Sci*, 33(3): 179–194
- Monteiro M, Martins B, Murrieta-Flores P, Pires J M (2019). Spatial disaggregation of historical census data leveraging multiple sources of ancillary information. *ISPRS Int J Geoinf*, 8(8): 327
- Moxey A, Allanson P (1994). Areal interpolation of spatially extensive variables: a comparison of alternative techniques. *Int J Geogr Inf Syst*, 8(5): 479–487
- Palacios-Lopez D, Esch T, MacManus K, Marconcini M, Sorichetta A, Yetman G, Zeidler J, Dech S, Tatem A J, Reinartz P (2022). Towards an improved large-scale gridded population dataset: a pan-european study on the integration of 3D settlement data into population modelling. *Remote Sens (Basel)*, 14(2): 325
- Pavía J M, Cantarino I (2017). Can dasymetric mapping significantly improve population data reallocation in a dense urban area. *Geogr Anal*, 49(2): 155–174
- Petrov A (2012). One hundred years of dasymetric mapping: back to the origin. *Cartogr J*, 49(3): 256–264
- Psyllidis A, Gao S, Hu Y, Kim E-K, McKenzie G, Purves R, Yuan M, Andris C (2022). Points of Interest (POI): a commentary on the state of the art, challenges, and prospects for the future. *Comput Urban Sci*, 2(1): 20
- Qiu Y, Zhao X, Fan D, Li S, Zhao Y (2022). Disaggregating population data for assessing progress of SDGs: methods and applications. *Int J Digit Earth*, 15(1): 2–29
- Rose A, McKee J, Sims K, Bright E, Reith A, Urban M (2021). *LandScan Global 2020* [Global]. Oak Ridge National Laboratory
- Šimbera J (2020). Neighborhood features in geospatial machine learning: the case of population disaggregation. *Cartogr Geogr Inf Sci*, 47(1): 79–94
- Sinha P, Gaughan A E, Stevens F R, Nieves J J, Sorichetta A, Tatem A J (2019). Assessing the spatial sensitivity of a random forest model: application in gridded population modeling. *Comput Environ Urban Syst*, 75: 132–145
- Su M D, Lin M C, Hsieh H I, Tsai B W, Lin C H (2010). Multi-layer multi-class dasymetric mapping to estimate population distribution. *Sci Total Environ*, 408(20): 4807–4816
- Tatem A J (2017). WorldPop, open data for spatial demography. *Sci Data*, 4(1): 170004
- Thomson D R, Stevens F R, Chen R, Yetman G, Sorichetta A, Gaughan A E (2022). Improving the accuracy of gridded population estimates in cities and slums to monitor SDG 11: evidence from a simulation study in Namibia. *Land Use Policy*, 123: 106392
- Tobler W R (1979). Smooth pycnophylactic interpolation for geographical regions. *J Am Stat Assoc*, 74(367): 519–530
- United Nations (2021). *Global Population Growth and Sustainable Development*. United Nations Department of Economic and Social Affairs, Population Division
- Ural S, Hussain E, Shan J (2011). Building population mapping with aerial imagery and GIS data. *Int J Appl Earth Obs Geoinf*, 13(6): 841–852
- Weber E M, Seaman V Y, Stewart R N, Bird T J, Tatem A J, McKee J J, Bhaduri B L, Moehl J J, Reith A E (2018). Census-independent population mapping in northern Nigeria. *Remote Sens Environ*, 204: 786–798
- Wu W B, Ma J, Banzhaf E, Meadows M E, Yu Z W, Guo F X, Sengupta D, Cai X X, Zhao B (2023). A first Chinese building height estimate at 10 m resolution (CNBH-10 m) using multi-source earth observations and machine learning. *Remote Sens Environ*, 291: 113578
- Yuan Y, Smith R M, Limp W F (1997). Remodeling census population with spatial information from LandSat TM imagery. *Comput Environ Urban Syst*, 21(3–4): 245–258
- Zeifman L, Hertog S, Kantorova V, Wilmoth J (2022). *A World of 8 Billion*. United Nations Department of Economic and Social Affairs, Population Division
- Zhang C, Qiu F (2011). A point-based intelligent approach to areal interpolation. *Prof Geogr*, 63(2): 262–276