Published in final edited form as:

Int J Appl Earth Obs Geoinf. 2024 June ; 130: . doi:10.1016/j.jag.2024.103882.

Unraveling near real-time spatial dynamics of population using geographical ensemble learning

Yimeng Song[a,*], **Shengbiao Wu**[b], **Bin Chen**[b], **Michelle L. Bell**[a,c]

[a]School of the Environment, Yale University, New Haven, CT 06511, USA

[b]Future Urbanity & Sustainable Environment (FUSE) Lab, Division of Landscape Architecture, Faculty of Architecture, The University of Hong Kong, Hong Kong Special Administrative Region

[c]School of Health Policy and Management, College of Health Sciences, Korea University, Seoul, South Korea

Abstract

Dynamic gridded population data are crucial in fields such as disaster reduction, public health, urban planning, and global change studies. Despite the use of multi-source geospatial data and advanced machine learning models, current frameworks for population spatialization often struggle with spatial non-stationarity, temporal generalizability, and fine temporal resolution. To address these issues, we introduce a framework for dynamic gridded population mapping using open-source geospatial data and machine learning. The framework consists of (i) delineation of human footprint zones, (ii) construction of muliti-scale population prediction models using automated machine learning (AutoML) framework and geographical ensemble learning strategy, and (iii) hierarchical population spatial disaggregation with pycnophylactic constraint-based corrections. Employing this framework, we generated hourly time-series gridded population maps for China in 2016 with a 1-km spatial resolution. The average accuracy evaluated by root mean square deviation (RMSD) is 325, surpassing datasets like LandScan, WorldPop, GPW, and GHSL. The generated seamless maps reveal the temporal dynamic of population distribution at fine spatial scales from hourly to monthly. This framework demonstrates the potential of integrating spatial statistics, machine learning, and geospatial big data in enhancing our understanding of spatio-temporal heterogeneity in population distribution, which is essential for urban planning, environmental management, and public health.

[*]Corresponding author. yimeng.song@yale.edu (Y. Song).

CRediT authorship contribution statement

Yimeng Song: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Shengbiao Wu:** Writing – review & editing. **Bin Chen:** Writing – review & editing. **Michelle L. Bell:** Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

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.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jag.2024.103882.

Keywords

Population spatialization; Data fusion; Social sensing; AutoGluon; Geospatial big data; GeoAI; Human mobility

1. Introduction

Data on the fine-scale distribution of population play an important role across various domains, including but not limited to disaster reduction (Sudmeier-Rieux et al., 2021), public health (Tran et al., 2022), urban planning (Chen et al., 2022), and global change studies (Goodwin et al., 2023). Census and survey data, while foundational, face limitations such as infrequent updates, limited spatial granularity, and confinement to administrative boundaries, affecting their usefulness in many contexts (Batista e Silva et al., 2020). Moreover, information that captures the dynamic nature of population distribution, moving beyond static residence-based statistics to include place-of-activity, is critical for diverse real-world applications, such as transport planning (Nieuwenhuijsen, 2020), epidemic surveillance (Li et al., 2020), and hazard exposure assessment (Caplin et al., 2019). In response to these challenges and demands, dynamic gridded population data, providing a comprehensive spatio-temporal representation of population distribution at a finer scale, are recognized as a promising solution (Leyk et al., 2019).

Gridded population mapping, or population spatialization, involves disaggregating population from larger geographic units to finer target units using spatial disaggregation methods such as areal interpolation, dasymetric mapping, statistical modeling, and machine learning modeling (Batista e Silva et al., 2020). Areal interpolation (e.g., areal weighting) uniformly redistributes census data across target grid cells based on spatial overlap (Doxsey-Whitfield et al., 2015), but may not capture the actual spatially heterogeneous features of population distributions. Dasymetric mapping, on the other hand, uses ancillary variables such as land cover, topography, and street networks to create more reasonable population redistribution schemes by assuming relationships between the population and ancillary variables (Freire et al., 2016; Wei et al., 2021). Statistical modeling and machine learning refine the dasymetric mapping method using regression relationships between the target variable (i.e., population size/density) and auxiliary variables (Stevens et al., 2015; Tatem, 2017). In practice, researchers typically do not use only one method alone, but combinations of methods to capitalize on the strengths of each and improve the overall accuracy of population mapping (Leyk et al., 2019).

While advancements in population spatialization methodologies are growing, they still face many challenges. A primary concern is spatial non-stationarity, a phenomenon where relationships vary across different locations, which is particularly challenging for areas with significant spatial heterogeneity in population distribution (Cockx & Canters, 2015). Efforts have been made to address this issue through the application of localized models, such as geographical weighted regression (GWR) (Wang et al., 2018), and geographical random forest (Georganos et al., 2021). Such models provide localized relationships between population and ancillary variables, rather than assuming spatially uniform relationships among variables as in global models (Xu et al., 2021). An important advance in addressing this concern has been the emergence of hybrid approaches, such as the integration of Random Forest, XGBoost, and GWR (Tu et al., 2022). Nevertheless, the potential of hybrid approaches for solving spatial non-stationarity remains largely unexplored, especially in maximizing the benefits of various machine learning models across distinct geographic contexts.

Addressing the challenges of temporal generalization within population spatialization frameworks is crucial for achieving dynamic population mapping with a fine temporal resolution. Since around 2010, the use of human digital footprint data, such as mobile phone records, geotagged social media, and location-based service (LBS) data from mobile applications, has become increasingly prevalent for capturing the dynamic spatial patterns of populations (Batista e Silva et al., 2020; Cheng et al., 2022; Song et al., 2018; Tsou et al., 2018). A distinguishing characteristic of these data is its temporal variability in magnitude, such as the daily fluctuations observed in Twitter/X usage (Tsou et al., 2018). This variability challenges the direct application of a time-specific framework across different times, thereby constraining their temporal generalization capabilities. Current frameworks often fail to adequately address this issue, leading to the generated dynamic population maps that represent broader temporal aggregates, such as monthly average (Cheng et al., 2022), diurnal average (Batista e Silva et al., 2020), or yearly averages for each individual hour (Tu et al., 2022), despite the availability of digital footprint data capable of supporting finer temporal analysis. Consequently, the task of mitigating the effects of digital footprint data’s magnitude fluctuations to improve a framework’s ability for temporal generalization and to facilitate the generation of detailed time-series dynamic population maps is a pressing concern.

In this study, we introduced an innovative framework for population spatialization to address the identified challenges, designed to produce comprehensive, large-scale, time-series gridded population maps. We implemented this framework in a case study in China, demonstrating its capacity to accurately depict the spatial dynamics of population at both monthly and hourly scales. This research aims to explore three critical questions: (1) How can we merge the strengths of various machine learning models with a geographical-sensitive ensemble approach to maximize benefits across different geographical contexts? (2) How can we mitigate the effects of magnitude fluctuations in digital footprint data to enhance the framework’s temporal generalization, thereby facilitating the generation of dynamic population maps with finer temporal resolution? (3) What insights can the generated maps provide into population distribution dynamics across varied spatial–temporal scales?

2. Materials

We constructed a library of spatial data layers for population spatialization. This library comprises twelve categories of datasets. With 2016 as the reference year, we collected the available data closest to 2016. Table 1 presents an overview of the data sources for these datasets and their utilization in Section 3.

Specifically, county-level census data from the 1 % population sample survey in 2015 published in the China Statistical Yearbook 2016 (National Bureau of Statistics of China, 2016) were used as ground truth data. A total of 2851 counties were included in this study, with the exception of Hong Kong, Macau, Taiwan, and islands in the South China Sea (Fig. S1 in Supplementary Material).

Tencent LBS data is a pivotal data source that offers insights into human spatial behavior and dynamics. The data record the real-time locations of active users utilizing Tencent's location-based services (Gong et al., 2020; Song et al., 2018). As one of China's largest internet service providers, Tencent recorded an average of 38 billion LBS requests per day from its 450 million active users worldwide in 2016, with 90 % of those requests generated in mainland China (Song et al., 2018). Tencent LBS has been demonstrated to effectively characterize population distribution and has been widely utilized in various fields such as population prediction (Xu et al., 2021), land cover/use mapping (Gong et al., 2020), environmental and ecology research (Song et al., 2021) and population migration modeling (Zhu et al., 2018). In this study, we collected data generated in 2016 via the method outlined in Song et al. (2018). The raw data is tabulated as the count of LBS requests within each 30 arc-second spatial interval released every 5 min. We aggregated and converted the tabular data into hourly raster data for 2016.

Due to the limited length of the article, we provide detailed information about the other datasets in the Supplementary Material. To ensure spatial correspondence between different raster layers, all raster data were resampled to 1-km resolution using either the nearest neighbor approach (for categorical data) or bilinear interpolation (for continuous data), and projected to the Albers Conical Equal Area projection.

3. Methodology

Several fundamental principles outlined in the literature (Gaughan et al., 2016; Wang et al., 2018) inform the population spatialization framework proposed in this study. These principles are: (1) population distribution should be spatially linked to areas with human activity; (2) every designated area should have a non-negative population count; (3) prediction models need to address spatial non-stationarity; and (4) inherent estimation errors should be managed using a pycnophylactic constraint. Following these principles, our proposed framework consists of the three steps shown in Fig. 1: (i) delineation of human footprint zones, (ii) construction of muliti-scale population prediction models, and (iii) hierarchical population spatial disaggregation.

3.1. Human footprint zone delineation

As population is not distributed over every inch of the earth's surface, we delineate human footprint zones to identify the potential spatial extent of population distribution, utilizing various layers of human activity-related factors. These layers include human digital footprint, human settlement footprint, artificial impervious area, road density, and POI density, as listed in the first five rows of Table 1. By overlaying these layers (Step-i in Fig. 1), we defined the human footprint zones as grids covered by at least one non-zero human activity-related factor. As a result, a total of 4.6 million grids (1*1 km) were identified as human footprint zones, covering roughly 49 % of the study area.

3.2. Multi-scale population prediction models constrcution

The core component of a population spatialization framework is applying a spatial disaggregation method to transform large-scale census data into a detailed grid format. In this study, spatial disaggregation started with predicting multi-scale population density using regression models with a bunch of ancillary variables. We constructed two regression models using an automated machine learning (AutoML) framework and different ensemble learning strategies. The first model is an ensemble model used for grid-scale predicting (hereafter Grid-Model), and the second model is a geographical ensemble model used for county-scale predicting (hereafter County-Model). These two models were later used for the hierarchical population spatial disaggregation in Section 3.3.

3.2.1. Grid-Model construction—We first constructed an ensemble model, the Grid-Model, using AutoGluon, a new AutoML framework that automates data preprocessing, feature engineering, model selection, and tuning (Erickson et al., 2020). Its novel multi-layer stacking strategy (Step-ii in Fig. 1) consists of a base layer with diverse models and several subsequent layers. To save on computing, stackers in each layer reuse identical hyper-parameters. This approach can be considered deep learning with layer-wise training using arbitrary machine learning models. Stacker models in higher layers use previous predictions and original input features during training. The final stacking layer combines predictions using ensemble selection and aggregation methods to improve accuracy and reduce errors (Caruana et al., 2004).

We selected and extracted fourteen distinct ancillary variables with potential explanatory relationships with the target variable of population density, and extracted area-averaged features at the county scale (see Table 2 and Table S1). The reason why we use area-averaged features as both target variable and ancillary variables is to satisfy modal consistency to bolster the scale transferability (from county to grid scales) of the constructed nonlinear regression. Given that the official introduction of the census data emphasizes that the standard time for the demographic status it represents is 1st November (National Bureau of Statistics of China, 2016), for the Tencent LBS data, we trained the model using only the average hourly data in November to ensure temporal matching, which is different from earlier works that utilized annual average LBS data (Cheng et al., 2022; Tu et al., 2022). LBS data from other specific hours were utilized for population density prediction at the corresponding times.

We used automatic multi-layer stacking and 5-fold cross-validation as the parameters for model training. The 2851 county-scale samples were split into 2281 (80 %) for training and 570 (20 %) for validation. The model was trained using all the variable listed in Table 2. The variable Pop_den was transformed to ln(Pop_den) and served as the target variable to ensure the predicted population density greater than zero. The customized base models used by AutoGluon for regression model construction include Random Forest algorithms, Extreme

Random Tree algorithms, K-Nearest Neighbor algorithms, Boosting Tree algorithms, and Neural Network algorithms.

3.2.2. County-Model construction—

We constructed the County-Model based on the trained ensemble model (i.e., Grid-Model) by intorducing a geographical ensemble learning strategy. Specifically, due to the existence of positive spatial autocorrelation of county-scale population density in the study area (Moran' I = 0.576, p < 0.01), aggregation methods used in AutoGluon (e.g., Bayesian Model Averaging, Coopetitive Soft Gating Ensemble) do not fully utilize the strengths of base models across different geographic regions to address the potential spatial non-stationarity. To tackle this problem, we refined the aggregation method by incorporating the principles of GWR to assign different geographical weights to the predictions of various base models for more effective aggregation (Step-ii in Fig. 1). The final predicted population density of a county was received via Eq. (1):

$$P_i = \beta_0(u_i, v_i) + \sum_j M_{ij}\beta_j(u_i, v_i) + \varepsilon_i \tag{1}$$

where (u_i, v_i) denotes the coordinates of the geometric center of the county i, $\beta_0(u_i, v_i)$ is the intercept value, ε_i is a random error term, M_{ij} is the prediction of base models j, and $\beta_j(u_i, v_i)$ represents a set of weights to be assigned to M_{ij}. The estimation of weights $\beta_j(u_i, v_i)$ is given by Eq.(2):

$$\hat{\beta}(u_i, v_i) = \left[M^T W(u_i, v_i) M\right]^{-1} M^T W(u_i, v_i) P \tag{2}$$

where M is a matrix of the base models' predictions, P is a vector of the ground truth value of counties, and $W(u_i, v_i)$ is the spatial weight matrix generated from the adaptive Gaussian kernel function. The optimal number of county neighbors was chosen by using a cross-validation method (CV) (Bowman, 1984).

Notably, we chose not to use this geographical ensemble learning strategy for Grid-Model construction, recognizing that the spatial non-stationarity relationships built at the county level might not effectively translate to grid-scale predictions due to scale variations, which is also supported by the observations detailed in Section 4.2.

3.3. Hierarchical population spatial disaggregation

Based on the constructed County-Model and Grid-Model, we proposed a hierarchical population spatial disaggregation approach for the hourly population mapping (Step-iii in Fig. 1). For a given hour, the approach begins with county-scale population density prediction using the County-Model. For the prediction, all ancillary variables remain constant as model training, except Tencent LBS data, which were updated to align with the corresponding hour. Prior to prediction, we employed a correction method based on pycnophylactic constraints (hereafter pycnophylactic correction) to ensure that the magnitude of LBS count in any given hour is consistent with the hourly average count in November. The correction procedure is outlined in Eq. (3):

$$lbs_{it}^{'} = lbs_{it} \times \frac{LBS_{11}}{LBS_t} \tag{3}$$

where lbs_{it} and lbs_{it}*'* represent the raw and corrected LBS count of county/grid i at hour t, respectively. LBS_{11} and LBS_t denote the LBS count across the entire study area in November (hourly average) and at hour t, respectively.

Following the county-scale prediction, a pycnophylactic correction was implemented on the predicted county-scale population density, as specified by Eq. (4):

$$Pc_i = \widehat{Pc_i} \times \frac{\sum_i^n TPc_i}{\sum_i^n \widehat{Pc_i} \times Area_i} \tag{4}$$

where Pc_i and $\widehat{PC_i}$ represents the corrected and the originally predicted population density of county i, respectively; TPc_i and $Area_i$ refer to the population in census data and the area of county i, respectively.

We then performed grid-scale prediction using the Grid-Model with all the ancillary variables at grid-scale (1*1 km) as well as the corrected LBS data of the corresponding hour. Notably, predictions were confined to grids within the human footprint zone as delineated in Section 3.1, and grids outside this zone were assigned a population density of zero.

Last, we applied pycnophylactic correction again to the predicted grid-scale population density based on the corrected county-scale prediction, as specified in Eq. (5):

$$Pg_{ij} = \widehat{Pg_{ij}} \times \frac{Pc_i \times Area_i}{\sum_j^n \widehat{Pg_{ij}}} \tag{5}$$

where Pg_{ij} and $\widehat{Pg_{ij}}$ represent the corrected and original predicted population densities of grid j in county i, respectively; Pc_i and $Area_i$ maintain the same definitions as in Eq. (4).

3.4. Accuracy validation

Four distinct metrics were used for accuracy validation and comparative analysis, including root mean square deviation (RMSD, Eq.6), relative root mean square deviation (%RMSD, Eq.7), mean absolute error (MAE, Eq.8), and the coefficient of determination (R^2).

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n} (p_i - \widehat{p_i})^2}{n}} \tag{6}$$

$$\%RMSD = \frac{RMSD}{\frac{1}{n}\sum_{i=1}^{n}\hat{p}_i} \tag{7}$$

$$MAE = \frac{\sum_{i=1}^{n}|p_i - \hat{p}_i|}{n} \tag{8}$$

where p_i and $\hat{p}_i$ is the prediction and ground truth of county i, respectively.

The final generated hourly gridded population maps in 2016 were aggregated to county-level annual averages and validated for accuracy based on census data. Additionally, we compared the recieved annual average map with four established population datasets, including WorldPop, LandScan, GPW, and GHSL.

4. Results

4.1. Model performance and feature importance

An ensemble model with a two-layer stacking structure and eleven base models was generated using AutoGluon framework (Table S2). This ensemble model was then used as Grid-Model as well as the basis for constructing the County-Model. For the selected base models, the topperforming ones measured by RMSD and R^2 are CatBoost, LightGBM, and XGBoost in both training and testing procedures (Fig. 2a–b, Table S2), yet the KNeighborsDist and KNeighborsUnif models perform relatively poorly but have the shortest training time. The ensemble model (WE) has the best performance, achieving an RMSD of 0.27 and R^2 of 0.981 in testing. The performance measured by %RMSD and MAE (Fig. 2c, Table S2) is consistent with those by RMSD and R^2. As an automatic multi-layer stacking strategy was used for the model training, the generated two-layer stacking structure shows that increasing the number of layers does not substantially improve the prediction performance of the model.

We used permutation importance to evaluate the contributions of different features (i.e., ancillary variables) to the model's predictive accuracy, as measured by RMSD. A higher score indicates a more siginificant impact on accuracy, while a negative score suggests the feature may detract from model performance, implying that removing such features could enhance predictions (Erickson et al., 2020). Fig. 2d displays the importance scores of the used fourteen features. The result shows that the LBS data has the highest importance score of 1.592, highlighting its critical role in population density prediction, followed by urban coverage (0.258) and cropland coverage (0.172). Additionally, all the fourteen selected features have positive importance scores ($p < 0.01$, Table S3), suggesting they are unlikely to affect the prediction results adversely.

4.2. Performance in multi-scale population density prediction

Based on their pre-corrected prediction in November, we compared the constructed County-Model and Grid-Model performance at different scales. For grid-scale assessment, we averaged grid values to match corresponding county scales.

For county-scale prediction, the Grid-Model obtained satisfactory accuracy with an R^2 of 0.977 and RMSD of 480 (Fig. 3a), while the accuracy of the County-Model further improved with an R^2 of 0.999 and RMSD of 279 (Fig. 3b). The better accuracy of the County-Model indicates that this geographical ensemble learning-based model had solved the spatial non-stationary issue in county-scale prediction to certain extent.

For grid-scale prediction, the Grid-Model exhibited notable performance (Fig. 3c), achieving an R^2 of 0.926 and an RMSD of 857. Given that the Grid-Model was trained using census data at the county scale, such high accuracy highlights the fine-scale transferability of the Grid-Model. In using the County-Model for grid-scale prediction, we used Kriging interpolation to transform geographical weights at the county scale to the grid scale, thus enabling the geographical ensemble learning strategy. However, the County-Model does not exhibit superior performance in grid-scale prediction (Fig. 3d), with an R^2 of 0.748 and an RMSD of 1558. This outcome suggests that the spatial non-stationary relationship initially formulated for county-scale prediction could not directly translate to grid-scale prediction. Therefore, the strong predictive performance of the County-Model used for county-scale population density prediction, together with the effectiveness of the Grid-Model used for grid-scale population density prediction, underscores the rationale behind the hierarchical population spatial disaggregation framework we proposed in Section 3.3.

Fig. 3e–h and 3i–l show the relationships between residuals, relative residuals (calculated as residuals divided by population density), and population density across models. In general, we observe that residuals show limited variation as population density increases. Specifically, in county-scale population density predictions, when using the Grid-Model (Fig. 3e and i), a slight negative association is observed between residuals (or relative residuals) and population density. This suggests a higher likelihood of underestimation in counties with high population density. However, when using the County-Model (Fig. 3f and j), the residuals are predominantly zero, with larger relative residuals only noticeable in counties with low population density. This implies that the County-Model performs sub-optimally in estimating populations in areas with low population density, but this limitation has minimal impact on the overall population estimation.

Turning to grid-scale population density prediction, the Grid-Model outperforms the County-Model in both residual measures (Fig. 3g–h) and relative residual measures (Fig. 3k–l). Furthermore, when using the Grid-Model, the negative relationship previously identified between residuals and population density in county-scale prediction becomes less pronounced in grid-scale prediction. Moreover, the near-zero correlation between relative residuals and population density creates favorable conditions for applying a pycnophylactic correction to the predictions.

4.3. Dynamic population distribution in China

We generated hourly population distribution maps for China in 2016 at a spatial resolution of 1 km utilizing the proposed population spatialization framework. Fig. 4a illustrates the annual average population distribution across China, revealing a notable concentration of population in the eastern and coastal regions, moderate density in central areas, and sparse population in the western and remote parts of mainland China. This observed pattern is primarily influenced by a combination of factors, including economic development, geography, climate, and natural resources. Fig. 4b–d present the population distribution within three major urban agglomerations in China: Beijing-Tianjin-Hebei (Fig. 4b), Yangtze River Delta (Fig. 4c), and Pearl River Delta (Fig. 4d), highlighting the general differences in population distribution between urban and rural areas, as well as the gradient of population density from urban to rural regions. They distinctly depict high population density in core urban areas, a slightly reduced density in surrounding satellite cities, and a lower density in rural areas.

The generated hourly population maps reveal the spatial distribution changes resulting from human activities throughout the day. Fig. 5a displays the hourly correlation matrix of population density for Monday, November 7, across China. Several relatively stable periods emerge from these patterns, each displaying high correlations in hourly population distribution within that phase. The first phase, spanning from 9:00 to 17:00, corresponds to people’s commute to work and school. A slight fluctuation around noon suggests midday movements for dining and other non-work activities. The second phase, from 18:00 to 22:00, represents a shift in population due to post-work and post-school activities. The third period, from 23:00 to 4:00, captures when most residents are at home. A notable change in distribution occurs from 5:00 to 8:00 as people move from their homes to various activity areas. Distinct daily population density shifts can be observed in regions based on their primary functions. For instance, commercial areas like Zhucheng, a County in Weifang City (C1 in Fig. S2a), see higher population densities during the day, whereas residential zones like Liangqing, a district in Nanning City (C2 in Fig. S2a), experience increased densities in the evening. Some areas, like Shuangfeng, a County in Loudi City of Hunan Province (C3 in Fig. S2a), undergo relatively minor population shifts throughout the day.

Beyond the hourly dynamics, there are also variations in population distribution across different months. Fig. 5d displays the correlation matrix of monthly population density throughout 2016. The population distribution typically shows higher similarities between adjacent months, with differences increasing as the time gap widens. For instance, the most significant difference is seen between February and November. However, for the country as a whole, there have not been drastic changes in population distribution throughout the year. The variations in monthly population distribution reveal movements due to factors such as holidays, economic activities, the education system, climate, and seasonal work, manifesting in different temporal patterns across regions. For instance, economically active urban areas such as Qingzhen County in Guiyang City (C1 in Fig. S2b) usually attract labor, leading to higher population densities throughout most of the year, but experience a significant decrease in January-February as the labor force returns home for the Chinese New Year. In contrast, areas with lower economic activity or rural regions see an influx of returning labor in January-February, causing a temporary population surge, such as in Yanting County of Mianyang City (C2 in Fig. S2b). On the other hand, some areas like Chibi County in Xianning City (C3 in Fig. S2b) maintain a consistent density, attributable to limited population movements.

Moreover, population distribution changes display distinct spatial characteristics across different temporal scales. For example, the variation in population distribution within a day primarily reflects transient shifts due to daily work and life-related activities. Fig. 5b–c compares the population distribution between 8:00 and 14:00 on Monday, November 7, 2016, in and around Shanghai and Shenzhen. By 14:00, only regions in the cities' cores with a high concentration of commercial activities show a heightened population density (indicated by the blue areas) compared to 8:00 in the morning. In contrast, monthly variations in population distribution are more indicative of urban inflows or outflows driven by seasonal factors. As depicted in Fig. 5e–f, there is a noticeable decrease in the whole urban areas of Shanghai and Shenzhen (indicated by the dark blue areas) during February, likely attributed to labor and students returning home for the Chinese New Year. Concurrently, there is an increase in the less economically developed *peri*-urban and rural areas with lower population density (indicated by the orange areas) compared to November.

We averaged the hourly gridded population map for November for temporal consistency with census data, and conducted a comparative analysis with four established population datasets. Regardless of the validation parameter used, such as RMSD, %RMSD, MAE, or R^2, our results consistently demonstrated superior accuracy when compared to the other datasets (Table S4). Our map achieved an RMSD of 324.9, outperforming LandScan, WorldPop, GPW, and GHSL, which reported larger RMSD values of 705.4, 817.1, 1079.8, and 775.3, respectively.

5. Discussion

The study presents a novel framework for dynamic population spatialization. The framework was applied to generate hourly gridded population distribution maps for China in 2016. The results show that it can accurately map population distribution with high temporal resolution.

5.1. Effects of incorporating human footprint zones

The usage of human footprint zones significantly improves the accuracy of population spatialization, especially for sparsely populated areas, such as desert areas and mountainous regions. Fig. 6 compares the effects of using and neglecting human footprint zones in Alxa Left Banner, Inner Mongolia, a low-density area (about 171,200 residents spanning 79,813 km^2 as of 2015). Populations are largely confined to small regions within the county boundary (red boundary in Fig. 6a), while expansive desert areas likely have minimal human activity. Without incorporating human footprint zones and scattering populations across all county grids, a significant population count could be mistakenly assigned to desert areas, resulting in a 38 % underestimation within the actual habitation zones in this region (Fig. 6b–c).

5.2. Scale transferability and variables extraction

In our proposed population spatialization framework, scale transferability is enhanced by using area-averaged features for both target and ancillary variables (e.g., population density, LBS density, mean of NDVI, urban coverage), instead of area-aggregated features (e.g., total population, total urban area) used in some previous studies (Cheng et al., 2022; Tu et al., 2022). We employ area-averaged features to maintain consistency and reduce scale dependence, which is critical when a model trained on large-scale (e.g., county-scale) data is applied to small-scale (e.g., grid-scale) predictions, especially in the presence of nonlinear relationships. Specifically, area-averaged features normalize data across spatial extents, representing average conditions per unit area, which minimizes the impact of scale changes and increases robustness in scale transfer, thus improving the model's scale transferability.

Additionally, the value range of the target variable (i.e., population density) should also be considered carefully during the variable extraction. For example, when target variables are extracted only within the human footprint area rather than within the administrative boundary, the performance of grid-scale prediction decreases with an R^2 value of 0.67 and RMSD of 3978 (Table S5). This degradation can be attributed to the higher population density within human footprint zones (Table S6), which leads to sample selection bias and ultimately results in the model's poor predictive performance in low-density areas at the grid scale. Therefore, incorporating samples with a wide value range of the target variable during the feature extraction is also critical to ensure robust scale transferability.

5.3. Trade-off of prediction accuracy and temporal generalization

In County-Model development, estimating geographic weights necessitates a balance between fitting accuracy and temporal generalization. In previous studies, alternative approaches were also used to estimate geographic weights, for example, using Pop_den transformed from the predicted ln(Pop_den) (Tu et al., 2022), rather than directly using the predicted ln(Pop_den) as in this study. Such strategy yields notable fitting accuracy (R^2 = 0.999, RMSD = 40, see Fig. S3a–c) in the model construction. However, the model was making unreasonable predictions when applied to other periods, evident in numerous negative values and unrealistic population density fluctuations (see Fig. S3d–f). For instance, a 51.5 % average change in county-scale population density occurs between February and November. Therefore, in the geographic weight estimation, we prioritized the temporal generalization of the County-Model while preserving a satisfactory fitting accuracy.

5.4. Enhancing performance with hierarchical spatial disaggregation

The hierarchical spatial disaggregation strategy with pycnophylactic correction enhances the spatial and temporal performance of the population spatialization framework. The spatial performance enhancement is in dealing the boundary effect, an inherent drawback effect of top-down population spatialization that previous studies have often tackled through complex interpolation methods (Cheng et al., 2022; Liu et al., 2008). Our method efficiently addressed this issue through input and output corrections before and during hierarchical spatialization, achieving continuous and reasonable spatial variation in population distribution. The temporal performance enhancement lies in promoting the limited temporal generalization capabilities of the framework, which refers to the prediction bias caused by the variation in the magnitude of LBS data over time (Fig. 7a–b). Fig. 7c–e shows the comparison of the results obtained using different population spatialization strategies in and around Shenzhen city. Specifically, because November is associated with more LBS data records than February (Fig. 7a), population spatialization using the raw LBS data without any pycnophylactic correction results in a larger population in November than in February almost everywhere (Fig. 7c). When we completed the initial correction of the national LBS data through Eq. (3), the bias in the result was reduced, but it remains difficult to accurately identify the difference in population distribution between the two months (Fig. 7d). With the hierarchical spatialization and correction, the bias is further reduced, and the general decrease in population density in Shenzhen in February compared to November due to the Chinese New Year is better characterized (Fig. 7e).

5.5. Framework adaptability and potential applications

The proposed population spatialization framework is based on a design principle of utilizing concise public data and provides a basis for reproducibility in different countries and regions. By replacing Tencent LBS data with comparable data sources that could capture the dynamics of population distribution, the framework can be effectively adapted to various geographical contexts. These alternative data sources may include cellular signaling data (Deville et al., 2014), social media checkin data from platforms such as Twitter/X (Longley & Adnan, 2016), Weibo (Song et al., 2019), and any other data that proves to accurately reflect the nature of population movements and distribution dynamics. The newly introduced regression model construction strategy integrates the strengths of established machine learning models while incorporating a geographical-sensitive ensemble approach. This innovative approach is tailored to recognize spatial non-stationarity and maximize the efficacy of various models over a range of geographical contexts.

6. Limitations

This study also faces certain limitations. The first limitation lies in the validation of the mapping results at grid scale, which is also a limitation faced by all similar works (Leyk et al., 2019). Due to the constraints in the availability of Chinese data, only county-scale demographic data are publicly accessible, hindering the direct validation of grid-scale outcomes. In previous studies, town-scale data have been utilized for validation purposes (Cheng et al., 2022). However, the accuracy of this data is uncertain as they were given that they were assembled from multiple data sources. Moreover, the lack of multi-temporal data also restricts the validation of accuracy at finer temporal scales. Thus, while our validation by aggregating grid-scale to county-scale provides promising evidence for the framework's capability, the incorporation of higher spatial resolution and multi-temporal ground truth data is crucial for future enhancements. The second limitation concerns the representation of LBS data. The use of smartphones is comparatively lower among the elderly and children, leading to potential discrepancies in LBS data sampling across different age groups. Despite our results demonstrating satisfactory accuracy across all scales, the inherent sampling bias in LBS data will unavoidably impact the accuracy of these results. As such, the limitations of data availability, resolution, and potential sampling bias present avenues for further refinement in future research.

7. Conclusions

This study introduced an innovative framework for high temporal-resolution grided population mapping, leveraging open-source geospatial data, automated machine learning, and geographical ensemble learning techniques. To adhere to the fundamental principles of population spatialization and address the shortcomings in previous studies, the framework comprised three main steps: (1) delineation of human footprint zones, (2) population prediction using AutoML framework and geographical ensemble learning, and (3) hierarchical spatial disaggregation, enhanced with pycnophylactic correction. The population maps of China in 2016 produced through the proposed framework showcased remarkable accuracy (RMSD = 325), outperforming existing datasets like LandScan, WorldPop, GPW, and GHSL. Beyond its enhanced accuracy, the generated hourly time-series gridded population maps effectively capture the variations in population distribution due to human mobility across different temporal scales. This study underscores the value of incorporating machine learning, spatial statistics techniques, and geospatial big data for population spatialization, facilitating a nuanced understanding of population distribution and spatial heterogeneity, which is critical for urban planning, environmental management, and public health.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Research reported in this publication was supported by the National Institute on Minority Health And Health Disparities of the National Institutes of Health under Award Number R01MD016054. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data availability

Data will be made available on request.

References

Batista e Silva F, Freire S, Schiavina M, Rosina K, Marín-Herrera MA, Ziemba L, Craglia M, Koomen E, Lavalle C, 2020. Uncovering temporal changes in Europe's population density patterns using a data fusion approach. Nat. Commun 11 (1), 4631. 10.1038/s41467-020-18344-5. [PubMed: 32934205]

Bowman AW, 1984. An alternative method of cross-validation for the smoothing of density estimates. Biometrika 71 (2), 353–360. 10.1093/biomet/71.2.353.

Caplin A, Ghandehari M, Lim C, Glimcher P, Thurston G, 2019. Advancing environmental exposure assessment science to benefit society. Nat. Commun 10 (1), 1236. 10.1038/s41467-019-09155-4. [PubMed: 30874557]

Caruana R, Niculescu-Mizil A, Crew G, & Ksikes A (2004). Ensemble selection from libraries of models Proceedings of the twenty-first international conference on Machine learning, Banff, Alberta, Canada. Doi: 10.1145/1015330.1015432.

Chen B, Wu S, Song Y, Webster C, Xu B, Gong P, 2022. Contrasting inequality in human exposure to greenspace between cities of Global North and Global South. Nat. Commun 13 (1), 4636. 10.1038/s41467-022-32258-4. [PubMed: 35941122]

Cheng Z, Wang J, Ge Y, 2022. Mapping monthly population distribution and variation at 1-km resolution across China. Int. J. Geogr. Inf. Sci 36 (6), 1166–1184. 10.1080/13658816.2020.1854767.

Cockx K, Canters F, 2015. Incorporating spatial non-stationarity to improve dasymetric mapping of population. Appl. Geogr 63, 220–230. 10.1016/j.apgeog.2015.07.002.

Deville P, Linard C, Martin S, Gilbert M, Stevens FR, Gaughan AE, Blondel VD, Tatem AJ, 2014. Dynamic population mapping using mobile phone data. Proc. Natl. Acad. Sci 111 (45), 15888–15893. 10.1073/pnas.1408439111. [PubMed: 25349388]

Doxsey-Whitfield E, MacManus K, Adamo SB, Pistolesi L, Squires J, Borkovska O, Baptista SR, 2015. Taking Advantage of the Improved Availability of Census Data: A First Look at the Gridded Population of the World, Version 4. Papers in Applied Geography 1 (3), 226–234. 10.1080/23754931.2015.1014272.

Erickson N, Mueller J, Shirkov A, Zhang H, Larroy P, Li M, & Smola A (2020). Autogluon-tabular: Robust and accurate automl for structured data. arXiv preprint arXiv:2003.06505.

Freire S, MacManus K, Pesaresi M, Doxsey-Whitfield E, Mills J, 2016. Development of new open and free multi-temporal global population grids at 250 m resolution. Population 250.

Gaughan AE, Stevens FR, Huang Z, Nieves JJ, Sorichetta A, Lai S, Ye X, Linard C, Hornby GM, Hay SI, Yu H, Tatem AJ, 2016. Spatiotemporal patterns of population in mainland China, 1990 to 2010. Sci. Data 3 (1), 160005. 10.1038/sdata.2016.5. [PubMed: 26881418]

Georganos S, Grippa T, Niang Gadiaga A, Linard C, Lennert M, Vanhuysse S, Mboga N, Wolff E, Kalogirou S, 2021. Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. Geocarto International 36 (2), 121–136. 10.1080/10106049.2019.1595177.

Gong P, Chen B, Li X, Liu H, Wang J, Bai Y, Chen J, Chen X, Fang L, Feng S, Feng Y, Gong Y, Gu H, Huang H, Huang X, Jiao H, Kang Y, Lei G, Li A, Xu B, 2020. Mapping essential urban land use categories in China (EULUC-China): preliminary results for 2018. Science Bulletin 65 (3), 182–187. 10.1016/j.scib.2019.12.007. [PubMed: 36659170]

Goodwin S, Olazabal M, Castro AJ, Pascual U, 2023. Global mapping of urban nature-based solutions for climate change adaptation. Nat. Sustainability 6 (4), 458–469. 10.1038/s41893-022-01036-x.

Leyk S, Gaughan AE, Adamo SB, de Sherbinin A, Balk D, Freire S, Rose A, Stevens FR, Blankespoor B, Frye C, Comenetz J, Sorichetta A, MacManus K, Pistolesi L, Levy M, Tatem AJ, Pesaresi M, 2019. The spatial allocation of population: a review of large-scale gridded population data products and their fitness for use. Earth Syst. Sci. Data 11 (3), 1385–1409. 10.5194/essd-11-1385-2019.

Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, Shaman J, 2020a. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). Science 368 (6490), 489–493. 10.1126/science.abb3221. [PubMed: 32179701]

Li X, Zhou Y, Gong P, Seto KC, Clinton N, 2020b. Developing a method to estimate building height from Sentinel-1 data. Remote Sens. Environ 240, 111705 10.1016/j.rse.2020.111705.

Liu XH, Kyriakidis PC, Goodchild MF, 2008. Population-density estimation using regression and area-to-point residual kriging. Int. J. Geogr. Inf. Sci 22 (4), 431–447. 10.1080/13658810701492225.

Longley PA, Adnan M, 2016. Geo-temporal Twitter demographics. Int. J. Geogr. Inf. Sci 30 (2), 369–389. 10.1080/13658816.2015.1089441.

National Bureau of Statistics of China, 2016. China statistical yearbook 2016. China Statistiestics Press.

Nieuwenhuijsen MJ, 2020. Urban and transport planning pathways to carbon neutral, liveable and healthy cities; A review of the current evidence. Environ. Int 140, 105661 10.1016/j.envint.2020.105661. [PubMed: 32307209]

Song Y, Huang B, Cai J, Chen B, 2018. Dynamic assessments of population exposure to urban greenspace using multi-source big data. Sci. Total Environ 634, 1315–1325. 10.1016/j.scitotenv.2018.04.061. [PubMed: 29710631]

Song Y, Huang B, He Q, Chen B, Wei J, Mahmood R, 2019. Dynamic assessment of PM2.5 exposure and health risk using remote sensing and geo-spatial big data. Environ. Pollut 253, 288–296. 10.1016/j.envpol.2019.06.057. [PubMed: 31323611]

Song Y, Chen B, Ho HC, Kwan M-P, Liu D, Wang F, Wang J, Cai J, Li X, Xu Y, He Q, Wang H, Xu Q, Song Y, 2021. Observed inequality in urban greenspace exposure in China. Environ. Int 156, 106778 10.1016/j.envint.2021.106778. [PubMed: 34425646]

Stevens FR, Gaughan AE, Linard C, Tatem AJ, 2015. Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data. PLoS One 10 (2), e0107042. [PubMed: 25689585]

Sudmeier-Rieux K, Arce-Mojica T, Boehmer HJ, Doswald N, Emerton L, Friess DA, Galvin S, Hagenlocher M, James H, Laban P, Lacambra C, Lange W, McAdoo BG, Moos C, Mysiak J, Narvaez L, Nehren U, Peduzzi P, Renaud FG, et al. , 2021. Scientific evidence for ecosystem-based disaster risk reduction. Nat. Sustain 4 (9), 803–810. 10.1038/s41893-021-00732-4.

Tatem AJ, 2017. WorldPop, open data for spatial demography. Sci. Data 4 (1), 170004. 10.1038/sdata.2017.4. [PubMed: 28140397]

Tran KB, Lang JJ, Compton K, Xu R, Acheson AR, Henrikson HJ, Kocarnik JM, Penberthy L, Aali A, Abbas Q, Abbasi B, Abbasi-Kangevari M, Abbasi-Kangevari Z, Abbastabar H, Abdelmasseh M, Abd-Elsalam S, Abdelwahab AA, Abdoli G, Abdulkadir HA, Murray CJL, 2022. The global burden of cancer attributable to risk factors, 2010–19: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 400 (10352), 563–591. 10.1016/S0140-6736(22)01438-6. [PubMed: 35988567]

Tsou MH, Zhang H, Nara A, & Han SY (2018). Estimating hourly population distribution change at high spatiotemporal resolution in urban areas using geotagged tweets, land use data, and dasymetric maps. arXiv preprint arXiv: 1810.06554. Doi: 10.48550/arXiv.1810.06554.

Tu W, Liu Z, Du Y, Yi J, Liang F, Wang N, Qian J, Huang S, Wang H, 2022. An ensemble method to generate high-resolution gridded population data for China from digital footprint and ancillary geospatial data. Int. J. Appl. Earth Observ. Geoinform 107, 102709 10.1016/j.jag.2022.102709.

Wang L, Wang S, Zhou Y, Liu W, Hou Y, Zhu J, Wang F, 2018. Mapping population density in China between 1990 and 2010 using remote sensing. Remote Sens. Environ 210, 269–281. 10.1016/j.rse.2018.03.007.

Wei S, Lin Y, Zhang H, Wan L, Lin H, Wu Z, 2021. Estimating Chinese residential populations from analysis of impervious surfaces derived from satellite images. Int. J. Remote Sens 42 (6), 2303–2326. 10.1080/01431161.2020.1841322.

Xu Y, Song Y, Cai J, Zhu H, 2021. Population mapping in China with Tencent social user and remote sensing data. Appl. Geogr 130, 102450 10.1016/j.apgeog.2021.102450.

Zhu D, Huang Z, Shi L, Wu L, Liu Y, 2018. Inferring spatial interaction patterns from sequential snapshots of spatial distributions. Int. J. Geogr. Inf. Sci 32 (4), 783–805. 10.1080/13658816.2017.1413192.

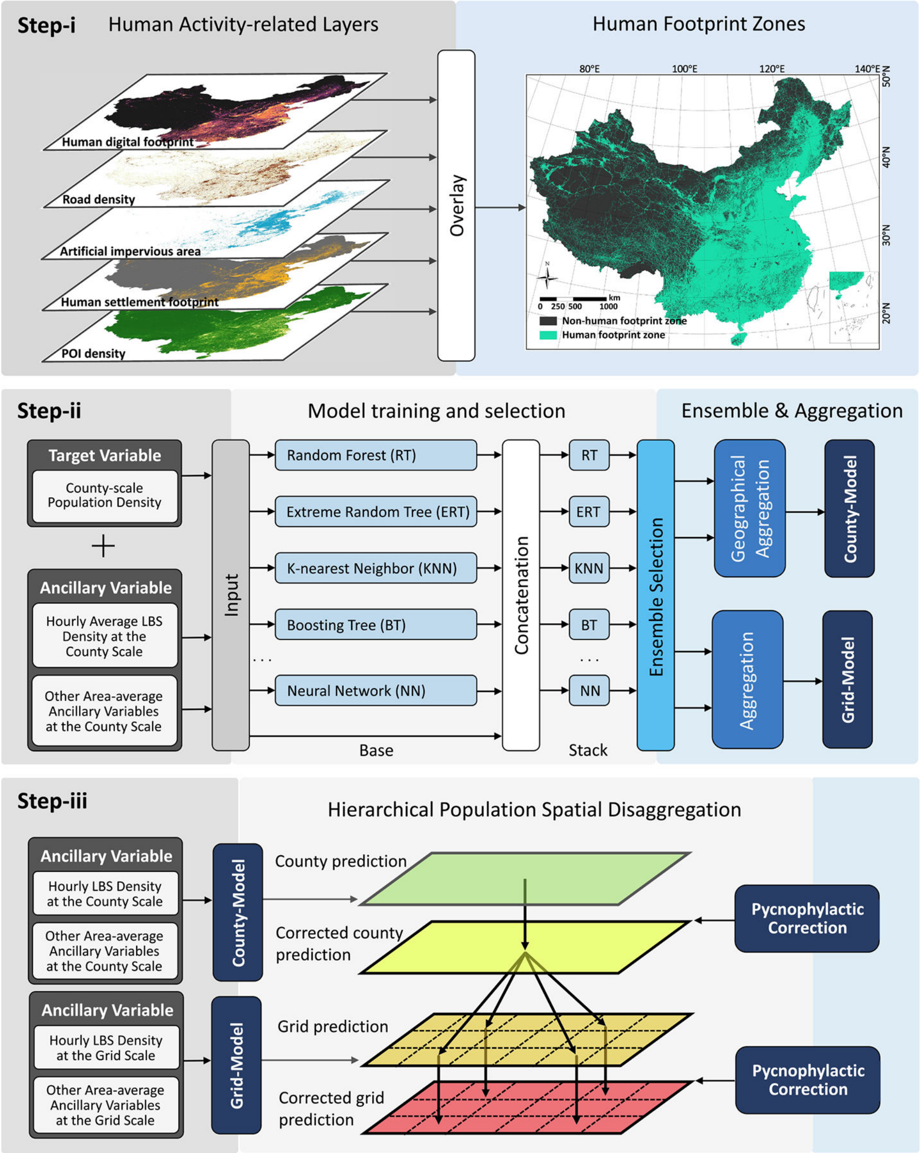
Step-i
Human Activity-related Layers
Human digital footprint
Road density
Artificial impervious area
Human settlement footprint
POI density
Overlay
Human Footprint Zones
80°E
100°E
120°E
140°E
50°N
40°N
30°N
20°N
N
0 250 500 1000 km
Non-human footprint zone
Human footprint zone
Step-ii
Model training and selection
Ensemble & Aggregation
Target Variable
County-scale Population Density
+
Ancillary Variable
Hourly Average LBS Density at the County Scale
Other Area-average Ancillary Variables at the County Scale
Input
Random Forest (RT)
Extreme Random Tree (ERT)
K-nearest Neighbor (KNN)
Boosting Tree (BT)
Neural Network (NN)
Base
Concatenation
RT
ERT
KNN
BT
NN
Stack
Ensemble Selection
Geographical Aggregation
County-Model
Aggregation
Grid-Model
Step-iii
Hierarchical Population Spatial Disaggregation
Ancillary Variable
Hourly LBS Density at the County Scale
Other Area-average Ancillary Variables at the County Scale
County-Model
County prediction
Corrected county prediction
Pycnophylactic Correction
Ancillary Variable
Hourly LBS Density at the Grid Scale
Other Area-average Ancillary Variables at the Grid Scale
Grid-Model
Grid prediction
Corrected grid prediction
Pycnophylactic Correction

Fig. 1.

Flowchart of population spatialization framework. Step-i: Delineation of human footprint zones; Step-ii: Construction of muliti-scale population prediction models; Step-iii: Hierarchical population spatial disaggregation.

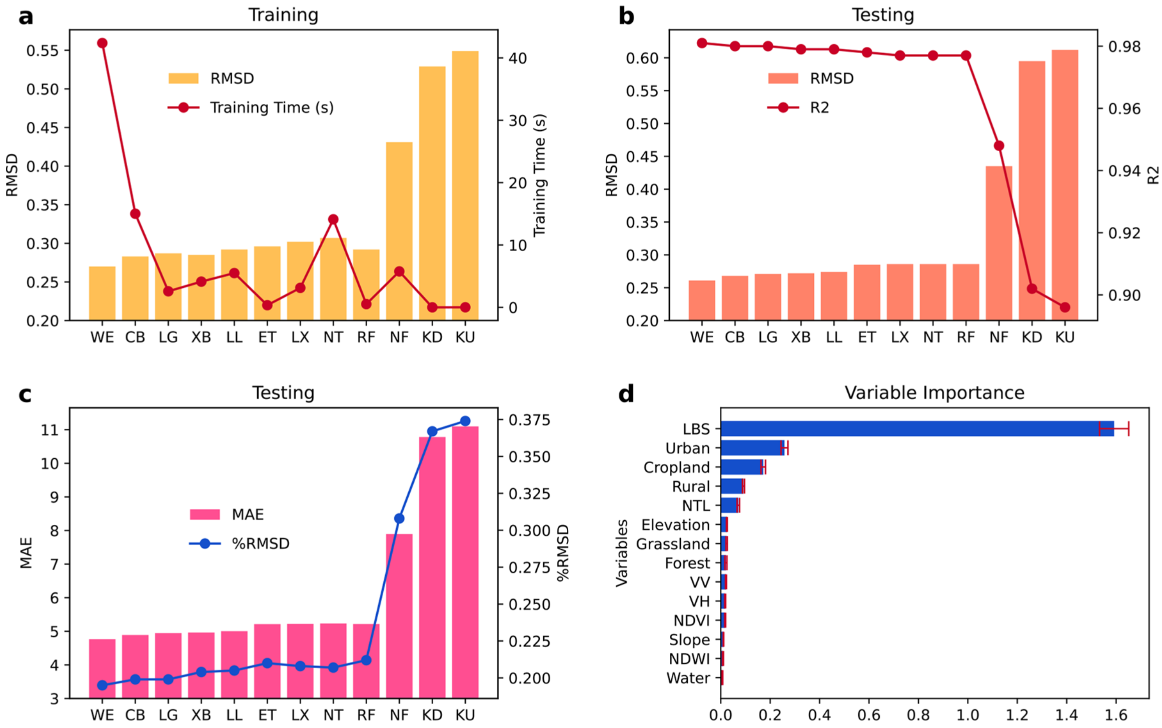

Fig. 2.

Model performance and feature importance: (a) Training performance measured by RMSD and training time; (b) Testing performance measured by RMSD and R^2; (c) Testing performance measured by MAE and %RMSD; (d) Importance score of features. Models evaluated include Ensemble Model (WE), CatBoost (CB), LightGBM (LG), XGBoost (XB), LightGBMLarge (LL), ExtraTreesMSE (ET), LightGBMXT (LX), NeuralNetTorch (NT), RandomForestMSE (RF), NeuralNetFastAI (NF), KNeighborsDist (KD), and KNeighborsUnif (KU).

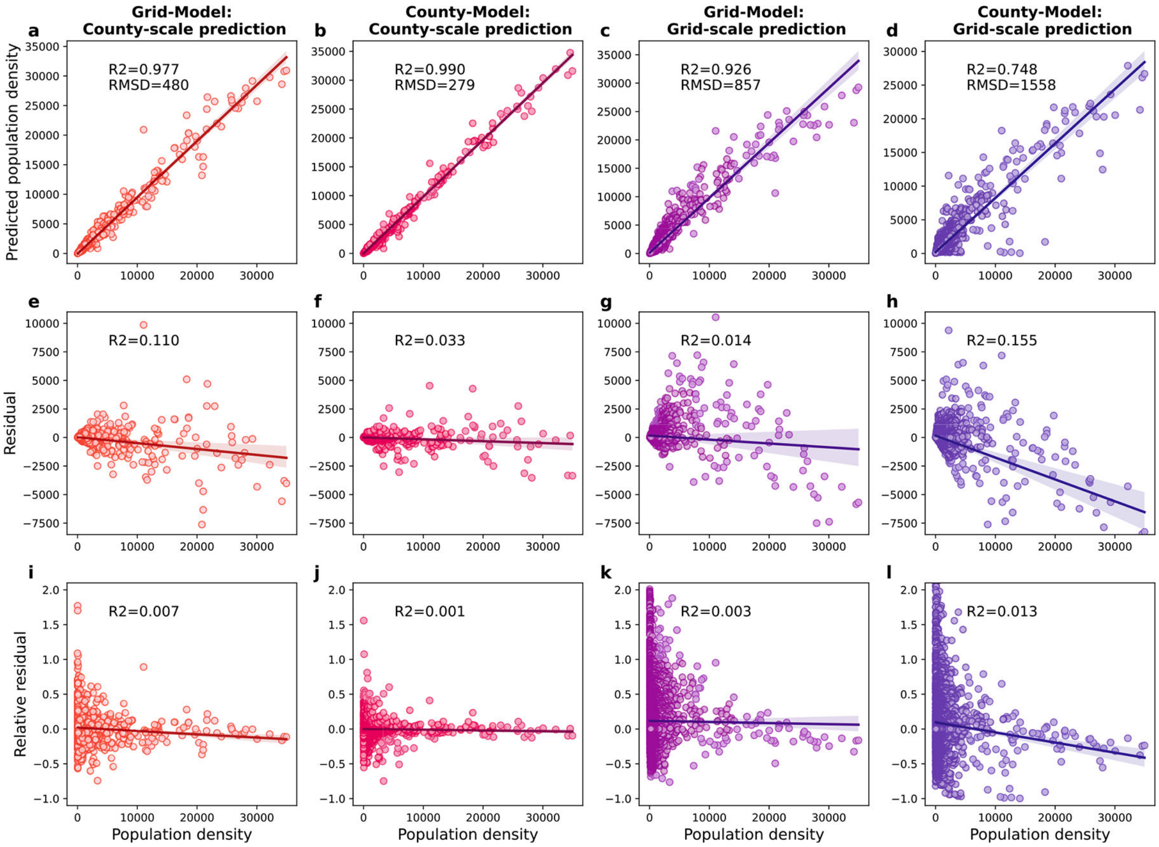

Fig. 3.

Comparison of the Grid-Model and County-Model in county-scale and grid-scale population density predictions: (a-d) Scatter plots of predicted population density versus actual population density; (e-h) Scatter plots of residuals versus population density; (i-l) Scatter plots of relative residuals versus population density.

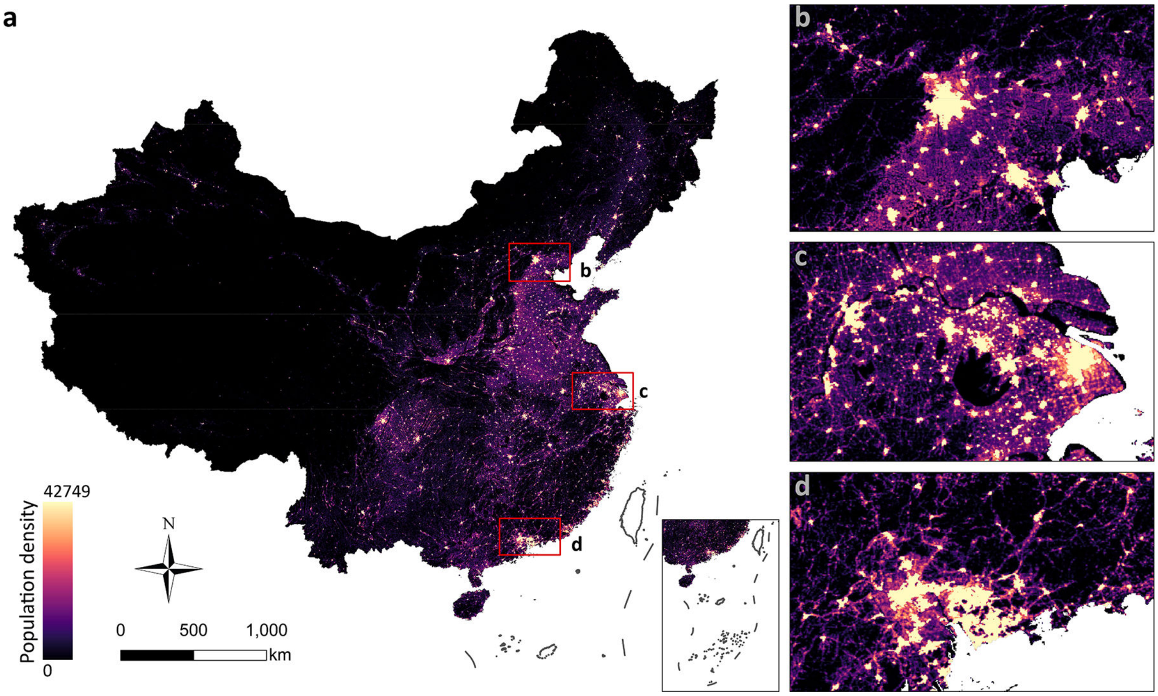

Fig. 4.

Annual average population distribution in 2016 across (a) China, (b) the Beijing-Tianjin-Hebei urban agglomeration, (c) the Yangtze River Delta urban agglomeration, and (d) the Pearl River Delta urban agglomerations.

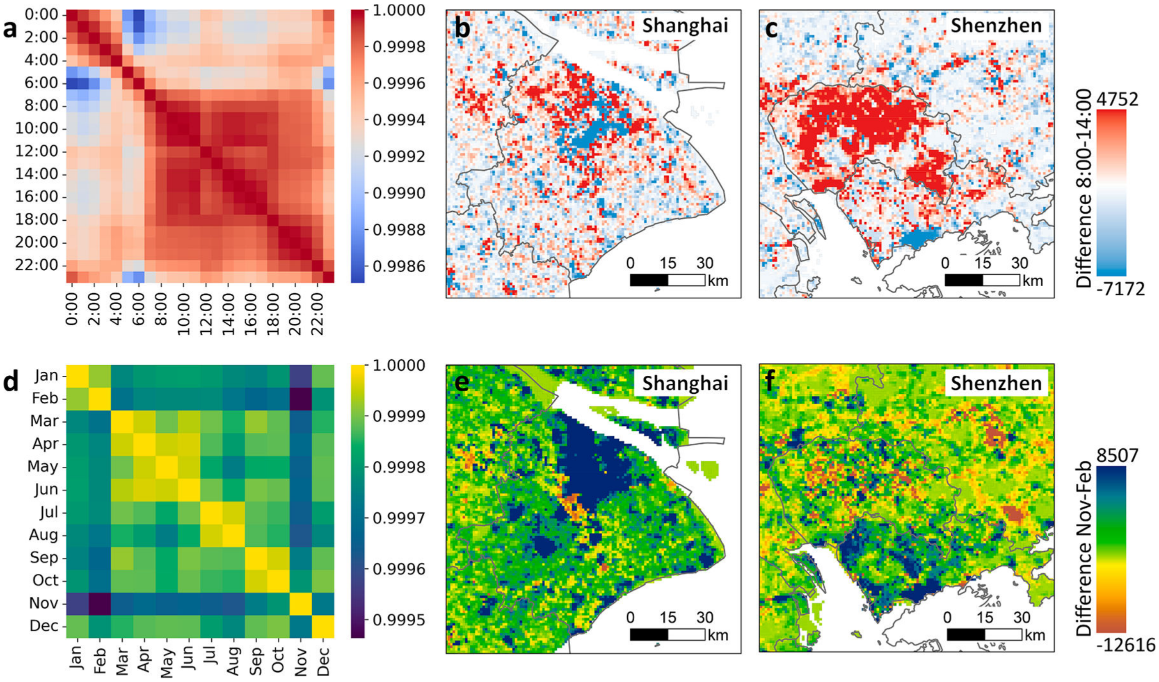

Fig. 5.

Spatiotemporal dynamics of population distribution in China: (a) Correlation matrix of hourly population density at county scale for November 7 (Monday); (b-c) Variation in population distribution between 8:00 and 14:00 on November 7 (Monday) in and around Shanghai and Shenzhen, respectively; (d) Correlation matrix of monthly population density at county scale; (e-f) Variation in population distribution between November and February in and around Shanghai and Shenzhen, respectively.

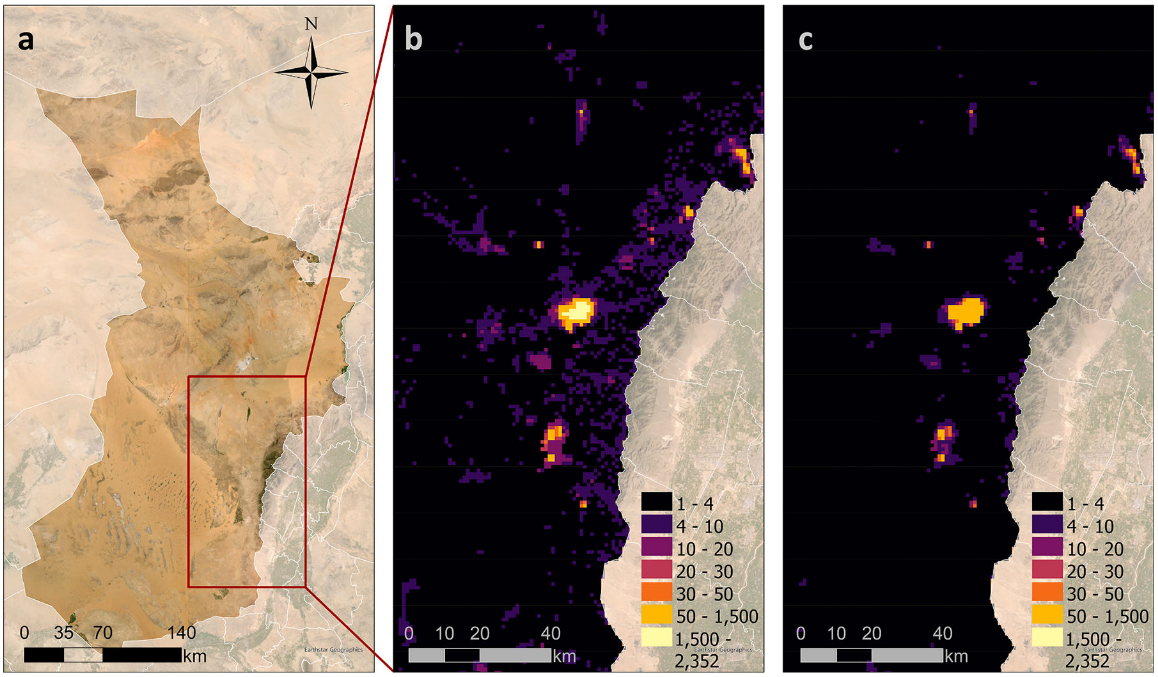

Fig. 6.

Population spatialization adopting and neglecting human footprint zones in the county of Alxa Left Banner, Inner Mongolia: (a) True-color satellite imagery; Population spatialization (b) adopting and (c) neglecting human footprint zones.

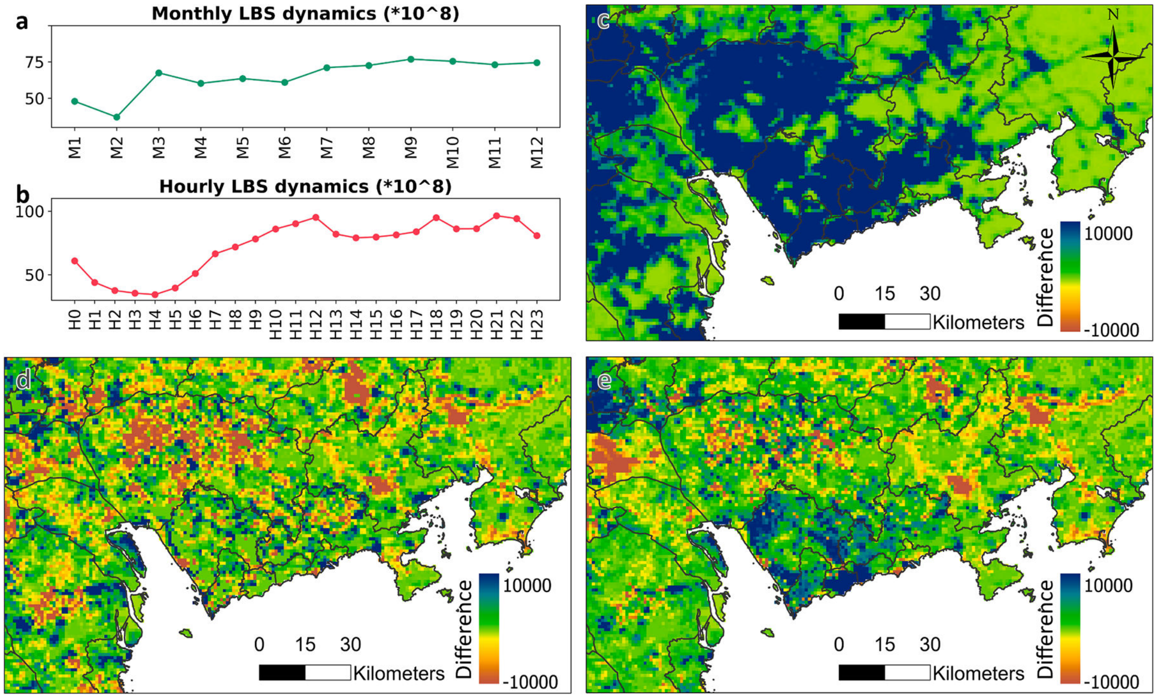

Fig. 7.

Dynamics of LBS record magnitude: (a) Monthly variations throughout 2016; and (b) Average hourly variations in November 2016. Population distribution changes (November – February) in and around Shenzhen city obtained using different spatialization strategies: (c) Using raw LBS data; (d) Using corrected LBS data; and (e) Using hierarchical spatialization strategy with pycnophylactic correction.

Table 1

List of datasets included in the spatial data layers library.

Step[*]	Category	Dataset & Source	Format	Spatial Resolution/ Scale	Temporal Resolution	Year
1	Human Settlement	World Settlement Footprint 2015	Grid	10 m	Annual	2015
1	Artificial impervious surface	Global Artificial Impervious Area	Grid	30 m	Annual	2015
1	Road network density	OpenStreetMap	Polyline	N/A	N/A	2018
1	POI density	Amap point of interest (POI)	Point	N/A	N/A	2018
1&2	Social sensing	Tencent location-based service (LBS) data	Grid	30 arc-second	5 min	2016
2&3	Demographics	Census data at county level (1 % population sample survey)	Polygon	County level	N/A	2015
2	Nighttime light data	NPP-VIIRS nighttime light (NTL) data	Grid	15 arc-second	Annual	2016
2	SAR	Sentinel-1 GRD	Grid	10 m	6 days	2016
2	MR multispectral data	Landsat-8 Operational Land Imager (OLT)	Grid	30 m	16 days	2016
2	Land-use	Land-use Status Remote Sensing Monitoring Database of China	Grid	30 m	Annual	2015
2	Topography	SRTM V4 digital elevation data (DEM) and slope	Grid	1 arc-second	N/A	2000
3	Gridded Population data	WorldPop	Grid	100 m	Annual	2015
3		LandScan	Grid	30 arc-second	Annual	2015
3		Gridded Population of the World (GPW)	Grid	30 arc-second	Annual	2015
3		Global Human Settlement Layer (GHSL)	Grid	1 km	Annual	2015

[*] The step(s) in which the dataset is used in Section 3.

Table 2

List of variables for regression model construction.

Data source	Feature	Variables
Census data	Population density	Pop_den
Tencent LBS data	LBS density	LBS
Landsat-8 OLT	Mean of NDVI	NDVI
	Mean of NDWI	NDWI
NPP-VIIRS NTL	Mean of nighttime light	NTL
Sentinel-1 GRD	Mean of VV	VV
	Mean of VH	VH
Land-cover data	Urban coverage	Urban
	Rural coverage	Rural
	Water coverage	Water
	Forest coverage	Forest
	Grassland coverage	Grassland
	Cropland coverage	Cropland
DEM	Mean of elevation	Elevation
	Mean of slope	Slope