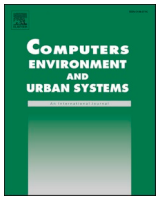
Computers, Environment and Urban Systems 105 (2023) 102024

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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



Areal interpolation of population projections consistent with different SSPs from 1-km resolution to block level based on USA Structures dataset

Heng Wan [a,*], Sumitrra Ganguli [b], Milan Jain [c], David Anderson [d], Narmadha Meenu Mohankumar [e], Kyle Wilson [b]

[a] Earth Systems Predictability & Resiliency Group, Pacific Northwest National Laboratory, Richland, WA 99352, United States
[b] Economics, Policy & Institutional Support Group, Pacific Northwest National Laboratory, Richland, WA 99352, United States
[c] Optimization & Control Group, Pacific Northwest National Laboratory, Richland, WA 99352, United States
[d] Risk & Environmental Assessment Group, Pacific Northwest National Laboratory, Richland, WA 99352, United States
[e] Math, Stats & Data Science, Pacific Northwest National Laboratory, Richland, WA 99352, United States

ARTICLE INFO

Keywords:
Population downscaling
Areal interpolation
USA Structures
Microsoft building footprints

ABSTRACT

Population data are normally collected at various census administrative levels, and areal interpolation of population is often required to transform population to the desired spatial resolution. Building footprint datasets, such as Microsoft building footprints, have proven to be useful in estimating population distribution and can therefore be used for areal interpolation of population. In addition to Microsoft building footprints, the recently released USA Structures dataset provides additional information such as building type and building height for some regions, which may provide valuable information for a better depiction of population distribution and improved population areal interpolation accuracy. In this study, we have conducted areal interpolation of population projections consistent with three different Shared Socioeconomic Pathways (SSP2, SSP3, and SSP5) from 1-km grid cells to block level in Washington state for every ten years from 2020 to 2040 based on USA Structures. We assessed USA Structures-based population downscaling accuracy using U.S. decennial survey data in 2020 under three different downscaling schemes, including population downscaling from census tracts to block groups, from census tracts to blocks, and from block groups to blocks. The resulting accuracies were compared with those based on Microsoft building footprints. The comparison showed that USA Structures achieved higher accuracies across different population density regions and areas with different urbanization extent within our study area.

1. Introduction

Projected population data are an essential data input in numerous research questions that target future predictions, such as future economic modeling, disaster prevention, urban design, environmental modeling, etc. (Brecht, Dasgupta, Laplante, Murray, & Wheeler, 2012; Chen, Paltsev, Reilly, Morris, & Babiker, 2016; Georgescu, Morefield, Bierwagen, & Weaver, 2014; Riordan and Rundel, 2014). Projected population is often produced at national level (Gerland et al., 2014). For example, Samir and Lutz (2017) produced population projections for 195 countries under five different Shared Socioeconomic Pathways (SSPs). These SSPs represent five plausible socioeconomic global change scenarios developed by the climate research community to facilitate comprehensive analysis of future climate impacts, vulnerabilities, adaptation, and mitigation (Riahi et al., 2017). Table 1 shows the summary of each SSPs. SSP1 represents a sustainability scenario guided by environmentally conscious development strategies while SSP2 portrays a middle-of-the-road scenario, reflecting moderate socioeconomic and environmental changes. SSP3 depicts a regional rivalry scenario highlighting geopolitical tensions and fragmented development while SSP4 presents an inequality scenario, focusing on high disparities and limited social cohesion. SSP5 is a fossil-fueled development scenario, characterized by extensive reliance on fossil fuels and limited environmental regulations (O'Neill et al., 2017). The global population projections differ among the SSPs. SSP1 and SSP5 have the lowest projected population of around 7 billion people by 2100, while SSP3 has the highest projection of 12.5 billion by 2100. SSP2 and SSP4 show a medium projected population of around 9.4 billion people (Riahi et al.,

* Corresponding author at: ETB/1342, 902 Battelle Boulevard, Richland, WA 99354, United States.
E-mail address: heng.wan@pnnl.gov (H. Wan).

https://doi.org/10.1016/j.compenvurbsys.2023.102024
Received 7 March 2023; Received in revised form 9 June 2023; Accepted 27 July 2023
Available online 3 August 2023

Table 1
Summary of SSPs.

Scenarios	Summary	Projected population by 2100
SSP1	Sustainability scenario: environmentally conscious development strategy	7 billion
SSP2	Middle of the Road scenario: moderate socioeconomic and environmental changes	9.4 billion
SSP3	Regional Rivalry scenario: geopolitical tensions and fragmented development	12.5 billion
SPP4	Inequality scenario: high disparities and limited social cohesion	9.4 billion
SSP5	Fossil-fueled Development scenario: extensive reliance on fossil fuels and limited environmental regulations	7 billion

2017).

Besides national-level, some subnational population projections also are produced to capture the regional heterogeneity in demographic processes and its corresponding outcomes (Wilson & Bell, 2004; Zoraghein & O'Neill, 2020). In Europe, Eurostat regularly produces regional population projections, and the most recent release is based on 2019 data (Buettner, 2022; Rees, Van Der Gaag, De Beer, & Heins, 2012; Van der Gaag, Van Imhoff, & Van Wissen, 2000). In China, provincial population projections have been estimated under different SSPs from 2010 to 2100 (Chen et al., 2020).

However, in the United States, state level population projections are not routinely produced or updated, even as the demand for sub-national population projections is increasing. While some states produce their own version of state-level population projections, the projected time is relatively short (typically only covering the following 10 to 20 years), and the implemented methods and underlying assumptions vary from state to state (Jiang, O'Neill, Zoraghein, & Dahlke, 2020).

To address the lack of a long-term state-level population projection with a national consensus on method and assumption, Jiang et al. (2020), relied on a cohort-component population projection model extended for multiregional demography to produce the very first set of such population projections for each state across different SSPs. In their method, state-level population projections were updated based on a modified national-level population projection, by considering different migration patterns between states across different gender and age groups. The modified national-level population projection accounts for updated demographic data and a more realistic long-term international migration mechanism consistent with the historical experience in the U. S. Population projections finer than the state level have also been produced for the U.S. For example, Hauer (2019) estimated U.S. population projections at county level by age, sex, and race across different SSPs. Zoraghein and O'Neill (2020) downscaled state-level population projections from Jiang et al. (2020) to 1-km resolution using a gravity-based population downscaling method.

Despite the existence of population data at different spatial resolutions, however, the data may still not match the spatial resolution required by the research (Merkle et al., 2022; Van Vuuren, Smith, & Riahi, 2010). To obtain the population data at the desired spatial resolution, it is imperative to transfer population data from one set of spatial units (source zones) to another (target zones), which is defined as areal interpolation process (Eicher & Brewer, 2001).

The simplest areal interpolation method is simple areal weighting, which directly redistributes population from the source zone to its corresponding target zones in proportion to the areas of the target zones (Goodchild, Anselin, & Deichmann, 1993). The limitation of simple areal weighting is that it assumes the population distribution is uniform across the study area, which is often not accurate (Sadahiro, 1999). Dasymetric mapping is another commonly used areal interpolation method, which disaggregates population to finer spatial units based on ancillary dataset (Eicher & Brewer, 2001; Zandbergen & Ignizio, 2010). Land cover data derived from satellite imagery are the mostly commonly used ancillary data in dasymetric mapping (Cartagena-Colón, Mattei, & Wang, 2022; Mennis, 2009). Other commonly used datasets include imperviousness, road networks, nighttime lights, etc. (Li & Zhou, 2018; Swanwick et al., 2022; Zandbergen & Ignizio, 2010). Due to its flexibility for the selection of ancillary datasets, in recent studies, dasymetric mapping has incorporated some novel data as ancillary datasets. For example, Wan, Yoon, Srikrishnan, Daniel, and Judi (2022) adopted settlement-related U.S. property data as an ancillary dataset and achieved comparatively high population downscaling accuracy when compared with other traditionally used ancillary datasets. Zandbergen (2011) implemented a high-resolution address point dataset for high accuracy dasymetric mapping of the population. Wan, Yoon, Srikrishnan, Daniel, and Judi (2023) explored the use of landscape metrics and found that they regularly outperform other traditionally used ancillary datasets in the dasymetric mapping of the population.

In recent years, dasymetric mapping of population efforts has leveraged building-level datasets because the distribution of residential buildings is directly related to population distribution (Boo et al., 2022). Huang, Wang, Li, and Ning (2021) utilized Microsoft building footprints dataset and Open Street Map (OSM) land use data to downscale population data. In their paper, non-residential buildings in the Microsoft building footprint were first removed based on OSM land use data, and then census tract level American Community Survey (ACS) 5-year population estimates (2013–2017) were redistributed to each corresponding 100-m grid cell in proportion to the Microsoft building area. The downscaled 100-m population grid cells were then re-aggregated to block group level for comparison with actual block group population data from ACS. The accuracy assessment showed that dasymetric mapping of population based on Microsoft building footprints could achieve high population downscaling accuracy when compared with other traditionally used population downscaling methods. One major limitation of using the Microsoft building footprint in dasymetric mapping is that it lacks building type information, making the removal of non-residential buildings a decision guided solely by the OSM land use data, which suffers from incompleteness and relatively low accuracy. Another limitation is that it lacks information on building height, which may not account for the highly urbanized areas dominated by high-rise buildings.

USA Structures is a recently released nationwide building footprint dataset. Compared with Microsoft building footprints, USA Structures provides additional building type and building height information for some regions, which may serve as valuable information for a better depiction of population distribution and thus improving population areal interpolation accuracy. The objective of this paper aims to map Washington state population projections consistent with three different SSPs (SSP2, SSP3, and SSP5) from 1-km resolution to block level for every ten years from 2020 to 2040 based on USA Structures. These block level population projections are important data input for our larger project – Grid Operations, Decarbonization, Environmental and Energy Equity Platform (GODEEEP), which aims at modeling the U.S. energy system interactions across scales under decarbonization and assessing its impact on environmental and energy equity. Specifically, population is a key input in accurately estimating disadvantaged communities (also referred to as DAC). Recently, DOE Justice40 initiatives defined DAC at the census tract level across the USA to determine where benefits of climate and energy investments are or are not currently accruing. By integrating with our spatial disaggregation work, DAC can be estimated at a much higher spatial resolution – at the block level.

As accuracy assessment is not feasible for our mapped future population projections, assessing the accuracy of population downscaling on historical data based on USA Structures can provide insights into the reliability and effectiveness of our mapping approach. Therefore, we utilized USA Structures to downscale U.S. decennial survey population data in 2020 and assessed its accuracy under three different downscaling schemes, including downscaling from census tracts to block groups,

from census tracts to blocks, and from block groups to blocks. Our aim was to showcase the capability of USA Structures in mapping population distributions in our study area, acknowledging the inherent differences in assessing accuracy between historical population downscaling and future projected population downscaling.

2. Method

2.1. Study area

Washington state is selected as the study area. According to the U.S. decennial census in 2020, Washington state has an average population density of 45 persons/km^2, and it consists of both highly urbanized areas with very high population density and rural areas with very low population density. Its broad spectrum of population density could facilitate the evaluation of USA structures on population downscaling across different areas with varying population densities and urbanization extent. Fig. 1 shows the population density at the census tract level for Washington state in 2020. We classify all the census tracts into low (< 250 persons/km^2), medium (250–1000 persons/km^2), and high (> 1000 persons/km^2) population density areas based on classification thresholds defined in Zandbergen and Ignizio (2010). A barplot showing the number of census tracts categorized by different population densities is also included. According to Fig. 1, Washington state is dominated by low population density census tracts in terms of land area, with approximately 97% of total land belonging to low population density census tracts. High population density census tract constitutes approximately 53% of all the census tracts while 27% and 20% of the census tracts belong to the medium and low population density categories, respectively.

2.2. Data

2.2.1. Microsoft building footprints

Microsoft released the Microsoft building footprints dataset in June 2018, which contains >1.29 million building footprint polygon geometries obtained from Bing imagery across the United States. These Bing images are collected from multiple sources with varying image-acquisition dates. Some of these images were acquired between 2019 and 2020, and a large portion of them were captured in previous years, with an averaging year of 2012 approximately. Microsoft building footprints depict U.S. buildings with a good performance, with a commission error of only 0.7% and an omission error of 6.5% (Microsoft, 2019).

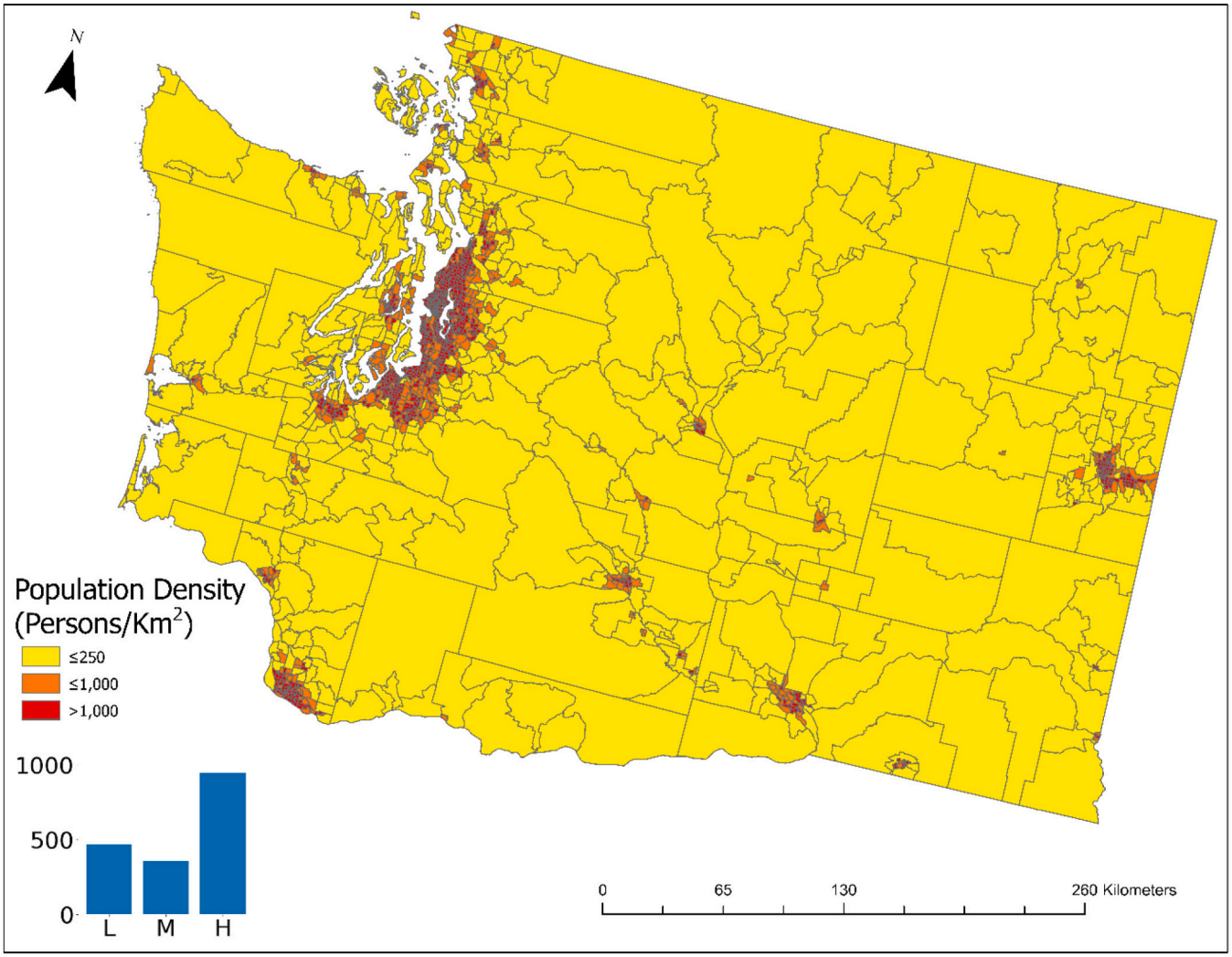


Fig. 1. Washington state population density in 2020 by census tract. The barplot on the lower-left corner shows the number of census tracts by population density, with "L", "M", and "H" indicating low population density (< 250 persons/km^2), medium population density (250–1000 persons/km^2), and high population density (> 1000 persons/km^2) categories.

2.2.2. USA Structures

USA Structures is the nation's first comprehensive inventory of all structures >450 square feet which is jointly developed and recently released by Department of Homeland Security, Federal Insurance and Mitigation Administration, Federal Emergency Management Agency's Response Geospatial Office, Oak Ridge National Laboratory, and the U. S. Geological Survey (USA Structures, 2023). However, its accuracy has not yet been assessed. Compared with Microsoft building footprints generated from Bing imagery, the USA Structures dataset contains building polygons extracted from commercially available satellite imagery including Maxar's Worldview 1, 2, and 3 (USA Structures, 2023). It also provides additional information such as building height derived from LiDAR and building type (e.g., residential, commercial, industrial building) derived from multiple sources including Census Housing Unit data, Homeland Infrastructure Foundation-Level Data, and LightBox parcel data, though the coverage of the additional information is limited (USA Structures, 2022). Fig. 2 shows the building height information coverage for Washington state at the census tract level. Building height information in Washington state is mainly available in highly urbanized areas, such as Seattle, Vancouver, and Spokane. Building type information can assist with excluding non-residential buildings in USA Structures and building height information can assist in delineating population distributions in highly urbanized areas dominated by high-rise buildings. The two of these, together, can potentially contribute to increasing population downscaling accuracy.

2.2.3. Open Street Map building footprint and land use parcel data

OSM building footprint is a crowd-sourced dataset (Brovelli & Zamboni, 2018). It relies on the efforts of volunteers from around the world who contribute to the dataset by utilizing various data sources (Hecht, Kunze, & Hahmann, 2013). OSM building footprints are typically digitized from satellite or aerial imagery, while other data sources such as GPS surveys, government data, and ground-based mapping may also be utilized depending on the specific region or contributor's preference (Neis & Zielstra, 2014). However, it is relatively incomplete in rural areas. The OSM land use parcel dataset depicts the land use attributes of each parcel. Both OSM building footprint and land use parcel data are used to remove non-residential buildings in the building footprint dataset.

2.2.4. Population projections consistent with different SSPs at 1-km resolution

Based on SSP-specific state-level population projections from Jiang et al. (2020), Zoraghein and O'Neill (2020) produced 1-km resolution urban and rural population projections using a gravity-based population downscaling method. This downscaling method redistributes the total state-level population change to each 1-km grid cell in proportion to its suitability value, which reflects the grid cell's potentiality for population growth. This set of 1-km population projections serve as our projected population data source for mapping block level population projections. We have chosen to utilize population projections under SSP2, SSP3, and SSP5 for mapping future block level populations in Washington state. Among the five SSPs, SSP3 predicts the lowest population growth in the U.S. due to low fertility, low international migration, and high mortality while SSP5 produces the highest population growth due to high fertility, high international migration, and low mortality (Samir & Lutz, 2017). SSP2 anticipates a medium level of population growth in the U.S. driven by medium levels of fertility, mortality, and international migration (Samir & Lutz, 2017). Selecting these three SSPs for projected population mapping enables us to capture the full spectrum of potential population dynamics in our study area as their qualitative assumptions would produce the widest possible range of population sizes for our study area (Jiang et al., 2020).

More details about the generation of this 1-km population dataset are discussed in Zoraghein and O'Neill (2020), and below is a brief description about the downscaling process. 1-km grid cell urban and rural population in 2010 for each state is used as the baseline for population projections. The population projection is a stepwise process and is done for every decade. For example, the 1-km population projection in 2020 is achieved by updating the 1-km population in 2010 by downscaling the state-level population change between the year 2010 and 2020 based on suitability value. The calculation of the suitability value for a target cell depends on a series of factors, including total population from neighboring cells, distances between the target cell and neighboring cells, as well as its topographic and land use/land cover characteristics. There are two parameters that govern how the total population from the neighboring cells and the distance between the target cell and the neighboring cells influence the suitability value of the target cell, which determines the spatial pattern of the projected population change. These two parameters are uniquely estimated for each state and each SSP scenario based on the state's historical estimations and the spatial pattern of the projected population change associated with each SSP.

2.3. Data processing

2.3.1. Areal interpolation of population projections based on USA Structures

Fig. 3 shows the flowchart of data processing for areal interpolation of population projections based on the USA Structures dataset, and a picture of areal interpolation is included in the lower left corner. We noted that some buildings had an erroneously low building height (e.g., 0.3 m), and we enforced a minimum building height of 4.2 m, which is the average height of one-story building. All buildings with height <4.2 m were reassigned a height value of 4.2 m.

Before using USA Structures in areal interpolation of population projections, non-residential buildings need to be removed. We used three different sets of ancillary information to exclude non-residential buildings in the dataset. We first removed all non-residential buildings that were labeled by building type information. Due to the incompleteness of building type information, it is not possible to remove non-residential buildings completely. Then USA Structures was overlaid with OSM land use parcel data to mask out all the buildings within non-residential land use parcels. The third removal step overlaid USA Structures with OSM building footprints, and all the buildings labeled as non-residential type based on OSM building attributes were removed.

After removing non-residential buildings, the resulted residential structures were overlaid with the 1-km fishnet grids, resulting in residential structure fragments (in picture of areal interpolation), the residential structure (smaller white polygon) in the middle of the picture was split into two structure fragments after overlying residential structures with 1-km fishnet grids). Population projection was redistributed from each 1-km grid cell to its corresponding residential structure fragment in proportion to the building area of each structure fragment. For grid cells where building height information was available, building volume was used to redistribute the population. If a grid cell contains structure fragments both with and without height information, then a height of 4.2 m was assigned to those structure fragments without height information (structures without height information in USA Structures were mostly one-story), and building volumes were calculated for population redistribution. After this population redistribution process, population was then aggregated from residential structure fragments to block fragment, which was the overlay result between block boundary and 1-km fishnet grids. In picture representation in Fig. 3, the green block was split into four block fragments after overlaid with 1-km fishnet grids. To correct for unusual high population density, an upper population density threshold is normally applied (Eicher & Brewer, 2001). Population density was calculated for each block fragment, and we set the 90th percentile block level population density (3690 persons/km^2) from U.S. decennial survey in 2020 as the upper threshold for population density. For each block fragment with population density greater than the threshold, excess population was redistributed to its

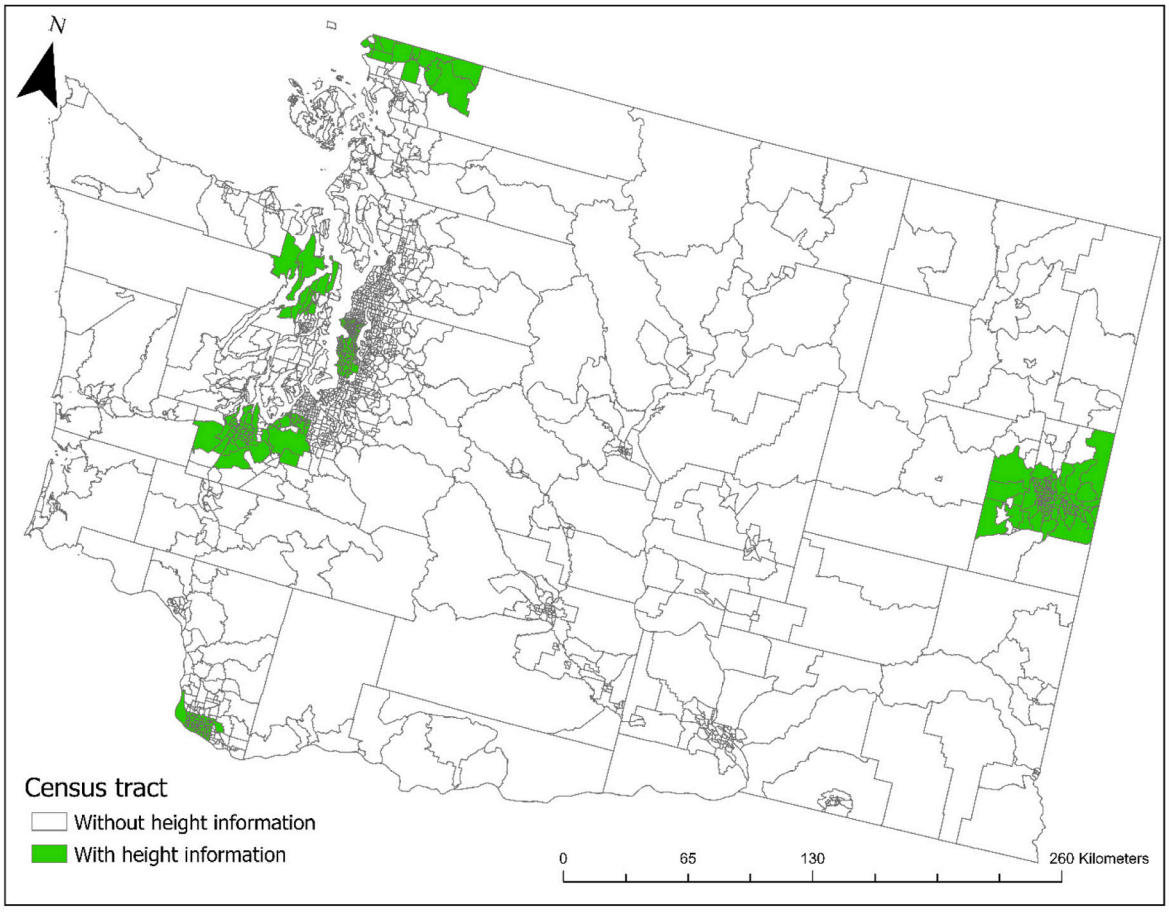


Fig. 2. USA Structures building height information coverage for Washington state at census tract level.

counterparts within the same 1-km fishnet grid. In general, the overflow of population to each block fragment was in proportion to its area. But if the population density threshold were reached for any block fragment receiving the overflow of population, it would stop receiving population and the excess population that was initially designed to flow to this block fragment will be redistributed to the remaining fragments which had not yet reached population density threshold. If all the block fragments have reached the population density, the excess population would remain in the original block fragment. This population adjustment ensured the pycnophylactic property of areal interpolation is met, that the population within each source zone (1-km grid cell) remain the same after population redistribution. Finally, population projections were aggregated from block fragment level to block level.

2.3.2. Population downscaling

Population downscaling based on USA Structures or Microsoft buildings was conducted for three different schemes, including downscaling population from census tracts to block groups, from census tracts to blocks, and from block groups to blocks.

For population downscaling based on USA Structures, building height reclassification and non-residential buildings removal processes followed the same procedures as described in areal interpopulation of population projections. Once completed, population was directly redistributed from source zone to its corresponding target zones in proportion to the total building area of each target zone (For areas with building height information available, redistribution was based on building volume).

For population downscaling based on Microsoft building footprints, non-residential buildings were removed based on OSM land use parcel data and OSM building footprints as described previously. Then population was directly reassigned from source zones to target zones in proportion to the total building area of each target zone.

2.4. Accuracy assessment

Accuracy assessments for projected population areal interpolation results were not possible because we could not obtain true population data for the future, thus we assessed population downscaling accuracies for historical population data based on USA Structures in three different downscaling schemes instead. Accuracies for population downscaling based on Microsoft building footprints were also assessed for comparison purposes.

After population downscaling, Pearson's correlation value (R^2), Root Mean Square Error (RMSE), Percentage of People Placed Incorrectly (PPPI[1]), and Median Absolute Error (MAE) were calculated by comparing the downscaled population with the ground-truth population data from U.S. decennial census survey.

We also assessed population downscaling performance by areas with different population density and impervious cover percentage. We divided target zones into low (< 250 persons/km^2), medium (250–1000 persons/km^2), and high (> 1000 persons/km^2) population density areas according to the population thresholds implemented by Zandbergen and

[1] PPPI is calculated as the percentage of the sum of the population difference between each downscaled population and actual population at the downscaled resolution over the total actual population divided by 2. The division of 2 considers the fact that one incorrectly placed person would result in one person increase in the targeted block where it is incorrectly placed, and one person decrease in the block where it should reside.

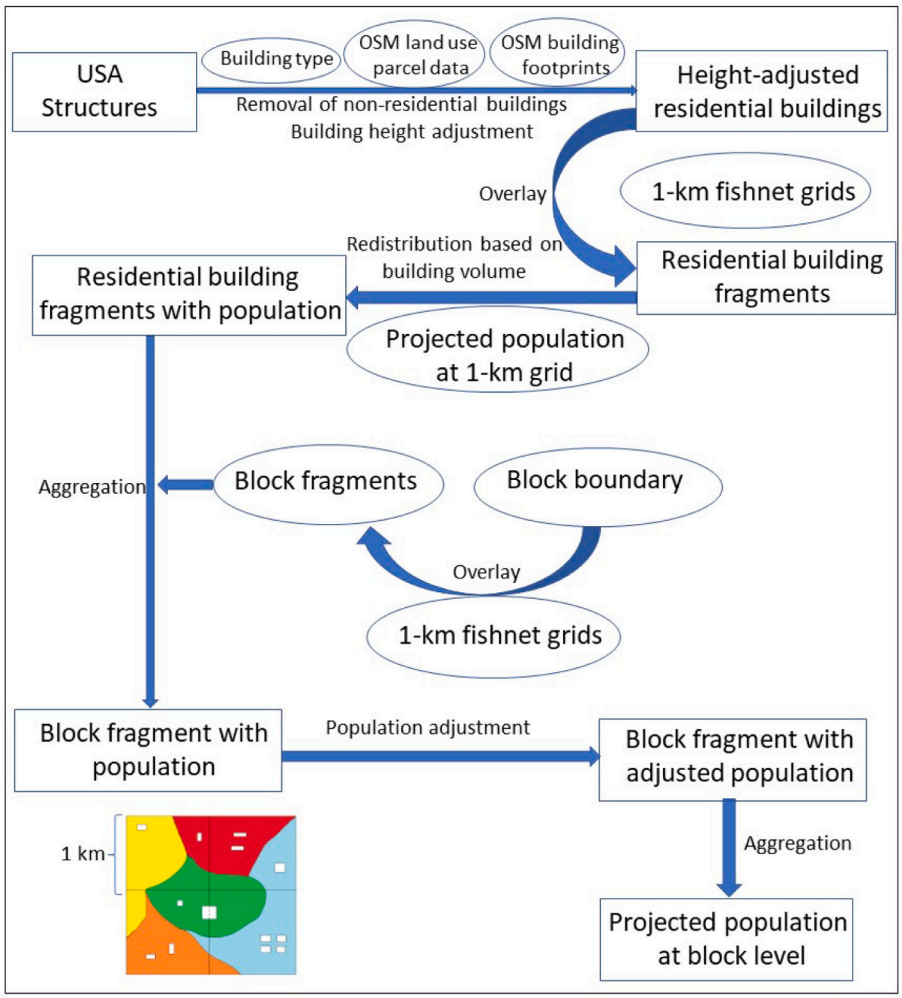


Fig. 3. Flowchart of data processing. A picture of areal interpolation is included in the lower left corner, with different colors of polygons representing different block and smaller white rectangles represent residential structures.

Ignizio (2010). We adopted the NLCD urban category classification strategy with minor adjustment and reclassified each target zone into non-urban (< 20% impervious cover), low-intensity urban (20–49% impervious cover), medium-intensity urban (50–79% impervious cover), and high-intensity urban (≥ 80% impervious cover) categories (Wan et al., 2023).

To explore whether height information contributes to a higher population downscaling accuracy, we also compared accuracy across two groups, one group with height information available in the USA Structures and the other without height information.

3. Results

3.1. Areal interpolation of population projections

Block level urban/rural/total population projections in Washington state under SSP2, SSP3, and SSP5 for every ten years from 2020 to 2040 were areal interpolated from 1-km grid cell population projections based on the USA Structures dataset. The data can be downloaded at doi: https://doi.org/10.5281/zenodo.7406280

3.2. Accuracy assessment for population downscaling

Table 2 shows accuracy assessment results for three different population downscaling schemes, namely census tract level to block group level, census tract level to block level, and block group level to block level, based on USA Structures and Microsoft building footprints, respectively. According to Huang et al. (2021), Microsoft building footprints have a high correlation with population distribution and can therefore be used to achieve high accuracy population downscaling. Based on Table 2, USA Structures dataset achieved better accuracy than Microsoft building footprints for all the three population downscaling schemes within Washington state, though the improvements were not dramatic. This result showed that USA Structures could be used for achieving population downscaling with high accuracy within our specific study area. We noticed that downscaling population from census

Table 2
Population downscaling accuracy results for three different downscaling schemes. Note: PPPI represents Percentage of People Placed Incorrectly; MAPE represents Median Absolute Percentage Error.

	USA Structures			MS buildings		
	R^2	PPPI	MAPE	R^2	PPPI	MAPE
Census tract to block group	0.60	8.30%	12.08%	0.57	8.74%	12.63%
Census tract to block	0.71	19.00%	36.17%	0.68	20.41%	36.54%
Block group to block	0.80	16.66%	33.33%	0.78	17.84%	34.07%

tract level to block group level had the highest accuracy (PPPI value of 8.3%) while census tract level to block level had the lowest accuracy (PPPI value of 19%). This might be because the former scheme has fewer average target zones per source zone than the latter scheme, as one census tract generally has only several block groups while one census tract could contain over one hundred blocks.

Table 3 shows population downscaling accuracy by areas with different population density and impervious cover percentage, and only PPPI is included as the assessing metric for simplicity. Based on Table 3, we observed that both USA Structures and Microsoft building footprints achieved comparably high accuracy across different population density and impervious cover percentage categories when downscaling population from census tract level to block group level. As for population downscaling scheme from census tracts to blocks and from block groups to blocks, both datasets had a much inferior performance in low population density category (< 250 persons/km^2) and low impervious cover percentage category ($< 20\%$), which were dominated by rural areas. This fact coincided with the finding from Zandbergen and Ignizio (2010) that small areal population estimation accuracy generally increases with the increase of population density due to more heterogenous population distribution in rural areas compared to urban areas.

Table 4 shows the PPPI values for groups with and without height information for the two datasets, respectively. We observed that the group without height information outperformed the group with height information across three different population downscaling schemes and two different datasets. Areas with height information were mostly characterized by high-rise buildings, which added complexity to population distribution patterns and thus resulted in a lower accuracy. Additional height information from USA Structures did not provide a large improvement for population downscaling. This may be because the census tracts were designed to be relatively homogeneous in terms of population characteristics and living conditions (Clapp & Wang, 2006). Due to the intrinsic homogeneity of census tract, buildings within the same census tract tend to have similar height values, making the height information less useful when redistributing population based on building volume.

Fig. 4a and b show the state-wide percentage error between the estimated and actual block group population for population downscaling from census tract level to block group level based on USA Structures and Microsoft building footprints, respectively. According to Fig. 4, we found that both datasets achieved high population downscaling accuracy for the majority of the block groups (with error percentage between -10% and 10%). We also noticed that severe over estimation of population happened mainly in rural areas, commercial and industrial areas for both datasets, but USA Structures was less prone to overestimate the population in these areas. One possible explanation is USA Structures depicts buildings more accurately than Microsoft building footprints in Washington state.

Fig. 5 shows the percentage of block group count over the total number of block groups by different categories of error percentage between estimated and actual block group population for population downscaling from census tract to block group. Compared with Microsoft building footprints, USA Structures resulted in less block group count percentage in both severe overestimation ($50\% \sim 100\%$, and $> 100\%$) and underestimation ($-100\% \sim -50\%$, and $-50\% \sim -30\%$) categories while higher block group count percentage in error percentage category of $-10\% \sim 10\%$. This demonstrates that USA Structures outperforms Microsoft building footprints for severe overestimation and underestimation of population in our study area.

Table 4
Percent of People Placed Incorrectly (PPPI) values for three different population downscaling schemes categorized by areas with USA Structures height information and areas without height information (Note: "n" is the number of target zones (i.e., block group, or block) in each category).

	With height	Without height
Census tract to block group	$n = 1280$	$n = 4019$
USA Structures	8.71%	8.16%
Microsoft Buildings	9.00%	8.65%
Census tract to block	$n = 32{,}250$	$n = 121{,}958$
USA Structures	20.48%	18.53%
Microsoft Buildings	21.89%	19.93%
Block group to block	n = 32,250	n = 121,958
USA Structures	18.00%	16.23%
Microsoft Buildings	19.18%	17.40%

4. Novelties and limitations

To the best of our knowledge, this is the first study which explores the possibility of incorporating the recently released building footprint dataset, USA structures, in population areal interpolation. We also assessed whether the additional building height information could help improve population downscaling accuracy, which has not been researched before due to the lack of height information in other nationwide building footprints datasets. This study also contributes to high-resolution (block level) population projections across different SSPs for Washington state for every ten years from 2020 to 2040.

One major limitation of this study is that we used current building footprints to areal interpolate future population projections, which could cause some problems due to time inconsistency between population data and building footprints data. One solution is to model building footprints projection, which is correlated with land use land cover change prediction, to the year consistent with the targeted population projection data, and then use the projected building footprints to conduct population projection areal interpolation. However, this step wouldn't increase accuracy in our implementation of population projection areal interpolation. This is because in this study, the 1-km resolution population projections that served as the source population data were obtained through a gravity-based population downscaling model, which had already considered future land use land cover change information at 1-km resolution. When conducting areal interpolation of population projections from 1-km grid cells to blocks in rural areas, it is

Table 3
Percent of People Placed Incorrectly (PPPI) values for three different population downscaling schemes categorized by different population density and impervious cover percentage (Note: "n" is the number of target zones (i.e., block group, or block) in each category).

	Population density (persons/km^2)			Impervious cover percentage			
	< 250	250–1000	> 1000	< 20%	20% ~ 49%	50% ~ 79%	≥ 80%
Census tract to block group	$n = 1227$	$n = 990$	$n = 3082$	$n = 775$	$n = 675$	$n = 838$	$n = 3011$
USA Structures	8.53%	8.13%	8.27%	8.16%	7.98%	8.03%	8.47%
MS Buildings	9.26%	8.81%	8.52%	8.63%	8.26%	8.26%	9.01%
Census tract to block	$n = 74{,}751$	$n = 16{,}573$	$n = 62{,}884$	$n = 39{,}830$	$n = 14{,}958$	$n = 11{,}245$	$n = 88{,}175$
USA Structures	34.48%	17.86%	15.86%	24.58%	18.87%	18.39%	18.46%
MS Buildings	40.92%	19.95%	16.01%	25.79%	19.26%	19.81%	20.01%
Block group to block	n = 74,751	n = 16,573	n = 62,884	n = 39,830	n = 14,958	n = 11,245	n = 88,175
USA Structures	30.43%	16.00%	13.79%	22.02%	16.61%	15.22%	16.27%
MS Buildings	35.37%	17.88%	13.98%	23.06%	17.00%	16.59%	17.52%

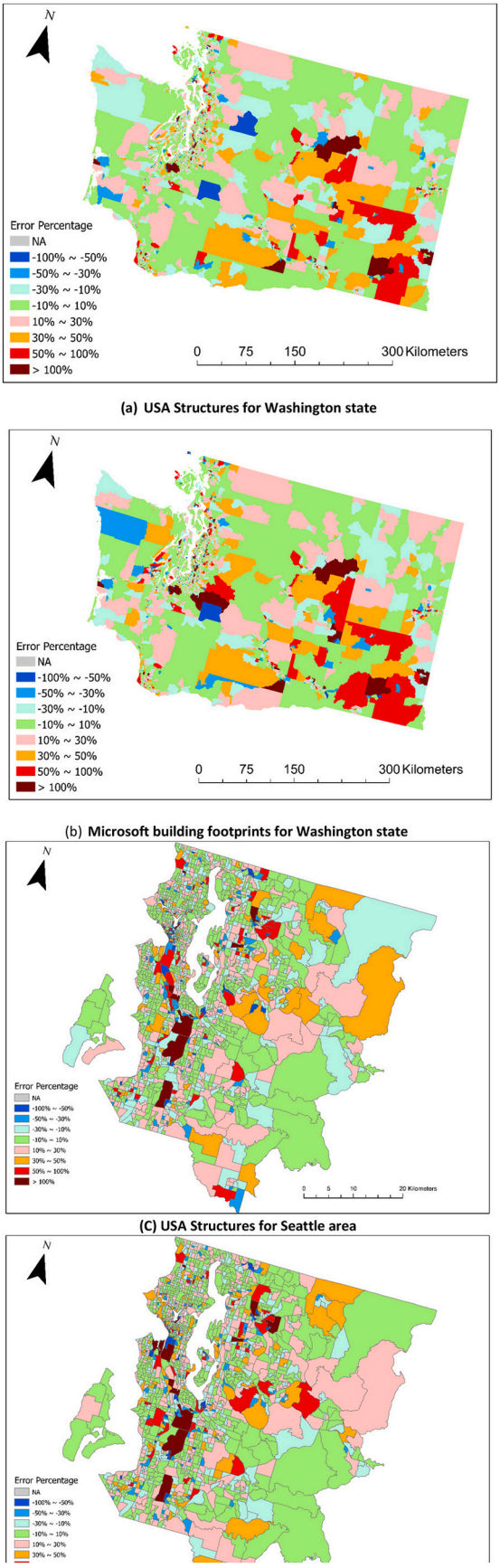


Fig. 4. Error percentage between estimated and actual block group population for population downscaling from census tract level to block group level.

not necessary to consider the newly constructed buildings (or land use land cover change) in the future since rural blocks are much larger than 1-km grid cell in terms of area, thus areal interpolation of population is achieved by simply aggregating populations from all the grid cells within each block. For the urban core areas, blocks are much smaller than 1-km grid cell, and areal interpolation of population projection requires disaggregating source zone population to each building. Meanwhile, these urban center grid cells are highly urbanized, and newly constructed buildings from a non-urban land cover is rare. Therefore, it is reasonable to use the current building footprints distribution to approximate the future building footprints distribution when conducting areal interpolation of population projections. It is those suburb blocks where they are smaller than 1-km grid cell and contain large quantity of open space for the new construction of buildings that require predicting future building constructions to help achieve a better areal interpolation of population projections. But the improvement of accuracy is only under the circumstance of having a high accuracy building footprints projection, which is hard to model and so far, does not exist for the study area. In this study, areal interpolations of population projections for these suburb areas from 1-km grid cells to block level using current building footprint datasets rely on the assumption that future building constructions in each 1-km grid cell are equally distributed. Moreover, as a developed country, U.S. now has a relatively low urbanization rate and is projected to keep this trend in the next couple decades (Chen, Zhang, Liu, & Zhang, 2014; Hsieh, 2014). Thus, we are more confident to use the current building footprints distribution to approximate the distribution of future building footprints projections as our projected years are not far into the future (i.e., 2020–2040). Future research should consider implementing projected building footprints for the areal interpolation of population projections if such dataset is produced with high accuracy.

Another limitation is that USA Structures does not include buildings <450 square feet, thus certain residential building type (i.e., recreational vehicle park) is not represented. As recreational vehicle park typically houses a high proportion of low-income residents, and its importance is increasing in population distribution depiction and population downscaling field, future work should consider incorporating a recreational vehicle park dataset with USA Structures to improve population downscaling accuracy.

5. Conclusions and discussions

We conducted areal interpolations of Washington state population projections under different SSPs for every ten years from 2020 to 2040 from 1-km grid cells to block level based on USA Structures, which serves as U.S.'s first national inventory of structures. To assess the capability of USA Structures in mapping population distributions in our study area, population downscaling accuracy based on USA Structures was assessed based on US decennial survey in 2020 under three different downscaling schemes, namely downscaling population from census tract to block groups, from census tract to block, and from block group to block. Its accuracy was compared with that of Microsoft building footprints dataset, which had been proven to be robust and accurate in population estimation and downscaling. We found that population downscaling performance was inferior in low population density areas for both building footprints datasets across three population downscaling schemes, which could be explained by a more heterogeneous population distribution in rural areas when compared with urban areas. Further comparison between the two datasets revealed that USA Structures outperformed Microsoft building footprints in our study area, which may be due to its more accurate depiction of the US buildings in Washington state and additional building type information used for filtering out non-residential buildings. Compared with Microsoft building footprints, USA Structures contains additional height information for some highly urbanized areas, but our study found that this additional height information did not contribute in improving the accuracy. One

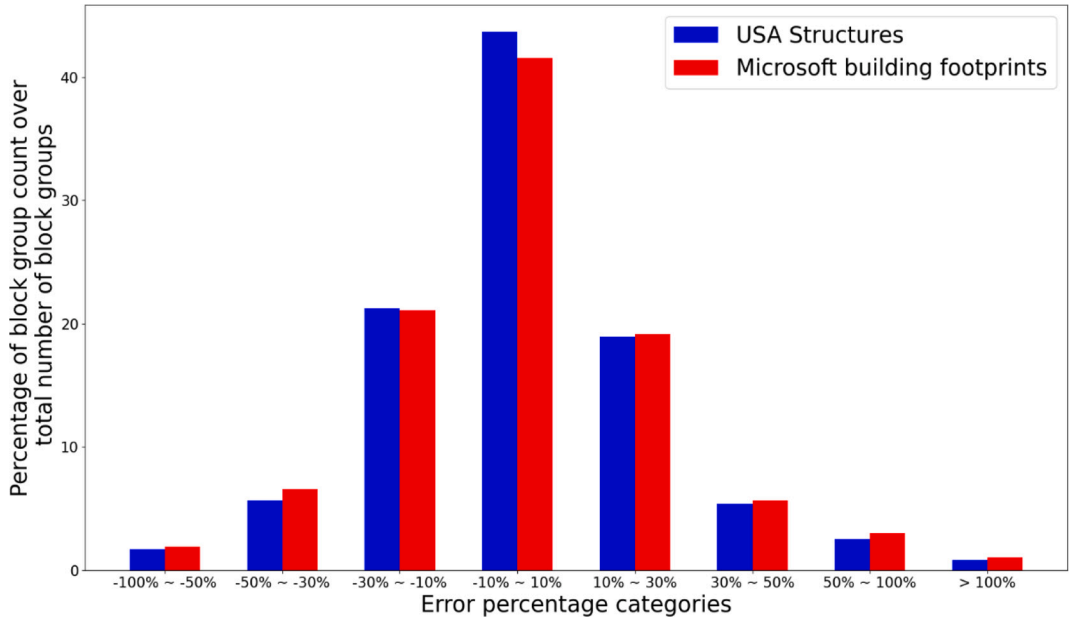


Fig. 5. Bar plot for the percentage of block group count over the total number of block groups by different categories of error percentage between estimated and actual block group population for population downscaling from census tract to block group.

possible explanation could be attributed to the data quality as we are unaware of the potential errors or inaccuracies in the building height data. Conducting a comprehensive accuracy assessment of the building height data in the future would provide valuable insights and help address this question effectively. Another possible explanation is due to the homogeneity of the census tracts in terms of building characteristics. Even though height information did not have a significant impact on all the three population downscaling schemes, we could not conclude it was not useful for mapping population distributions. If the source zone does not follow a census administrative unit boundary, buildings within the source zone would have much less homogeneity, potentially making height information more valuable for population downscaling. For example, in our implementation of areal interpolations of population projections from 1-km resolution to block level, the source zones are 1-km grid cells, and each grid cell could arbitrarily contain buildings with varying heights. Under this circumstance, building height information provides additional valuable information for mapping population distribution and is thus promising to improve areal interpolation accuracy. Future research should compare model accuracy with and without using height information for population downscaling schemes where source zones are more heterogenous in building characteristics.

To improve population downscaling accuracies, OSM land use data and OSM building footprints data were used as ancillary dataset to remove non-residential buildings contained in both USA Structures and Microsoft building footprints. However, due to its crowed-sourced characteristic, OSM data suffer from low accuracy and severe incompleteness. Therefore, the removal of non-residential buildings based on OSM was not complete, and the remaining non-residential buildings could heavily impact population downscaling accuracy. Compared with Microsoft building footprints, USA Structures provides additional building type information, which was further used for the removal of non-residential buildings in this dataset. However, its building type information is also highly incomplete and thus could not fully remove all the non-residential buildings. Future research should incorporate more accurate and complete building type information when using building footprints for the areal interpolation of population.

Our study contributes to high spatial (block level) and temporal (for every ten years) resolution population projections for the state of Washington from 2020 to 2040. These population projections are consistent with three different SSPs (SSP2, SSP3, and SSP5), producing a comprehensive range of population changes within our study area. As a result, they hold significant value in informing various aspects of urban planning, transportation, healthcare, and emergency management. With a more detailed understanding of future population distribution, policymakers could make informed decisions regarding disaster prevention, resource allocation, infrastructure development, and service provision at local scales.

As building footprint datasets incorporate more ancillary information, such as building type and building height, their role in depicting population distribution becomes increasingly important. By utilizing the USA Structures dataset for areal interpolation of population projections and incorporating ancillary data like building type and building height, our study has the potential to inspire researchers to explore similar methodologies for mapping population distribution by integrating building footprints with diverse data sources. Additionally, our approach of integrating ancillary data and building footprints can encourage interdisciplinary collaboration among researchers from fields such as demography, urban planning, data science, and geography. This collaborative effort may result in the development of innovative methodologies, improved modeling techniques, and a more comprehensive understanding of population distribution.

Funding

This work was internally supported by the Laboratory Directed Research and Development (LDRD) program of the Pacific Northwest National Laboratory, a Department of Energy, Office of Science Laboratory.

CRediT authorship contribution statement

Heng Wan: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Sumitrra Ganguli:** Writing – review & editing, Project administration. **Milan Jain:** Writing – review & editing. **David Anderson:** Writing – review & editing, Supervision. **Narmadha Meenu Mohankumar:** Writing – review & editing. **Kyle Wilson:** Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no conflict of interest.

References

Boo, G., Darin, E., Leasure, D. R., Dooley, C. A., Chamberlain, H. R., Lázár, A. N., … Fuller, T. (2022). High-resolution population estimation using household survey data and building footprints. *Nature Communications, 13*, 1–10.

Brecht, H., Dasgupta, S., Laplante, B., Murray, S., & Wheeler, D. (2012). Sea-level rise and storm surges: High stakes for a small number of developing countries. *The Journal of Environment & Development, 21*, 120–138.

Brovelli, M. A., & Zamboni, G. (2018). A new method for the assessment of spatial accuracy and completeness of OpenStreetMap building footprints. *ISPRS International Journal of Geo-Information, 7*, 289.

Buettner, T. (2022). Population projections and population policies. In *International handbook of population policies* (pp. 467–484). Springer.

Cartagena-Colón, M., Mattei, H., & Wang, C. (2022). Dasymetric mapping of population using land cover data in JBNERR, Puerto Rico during 1990–2010. *Land, 11*, 2301.

Chen, M., Zhang, H., Liu, W., & Zhang, W. (2014). The global pattern of urbanization and economic growth: Evidence from the last three decades. *PLoS One, 9*, Article e103799.

Chen, Y., Guo, F., Wang, J., Cai, W., Wang, C., & Wang, K. (2020). Provincial and gridded population projection for China under shared socioeconomic pathways from 2010 to 2100. *Scientific Data, 7*, 83.

Chen, Y.-H. H., Paltsev, S., Reilly, J. M., Morris, J. F., & Babiker, M. H. (2016). Long-term economic modeling for climate change assessment. *Economic Modelling, 52*, 867–883. https://doi.org/10.1016/j.econmod.2015.10.023

Clapp, J. M., & Wang, Y. (2006). Defining neighborhood boundaries: Are census tracts obsolete? *Journal of Urban Economics, 59*, 259–284.

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

Georgescu, M., Morefield, P. E., Bierwagen, B. G., & Weaver, C. P. (2014). Urban adaptation can roll back warming of emerging megapolitan regions. *Proceedings of the National Academy of Sciences, 111*, 2909–2914.

Gerland, P., Raftery, A. E., Ševčíková, H., Li, N., Gu, D., Spoorenberg, T., Alkema, L., Fosdick, B. K., Chunn, J., & Lalic, N. (2014). World population stabilization unlikely this century. *Science, 346*, 234–237.

Goodchild, M. F., Anselin, L., & Deichmann, U. (1993). A framework for the areal interpolation of socioeconomic data. *Environment and Planning A, 25*, 383–397.

Hauer, M. E. (2019). Population projections for US counties by age, sex, and race controlled to shared socioeconomic pathway. *Scientific Data, 6*, 1–15.

Hecht, R., Kunze, C., & Hahmann, S. (2013). Measuring completeness of building footprints in OpenStreetMap over space and time. *ISPRS International Journal of Geo-Information, 2*, 1066–1091.

Hsieh, S.-C. (2014). Analyzing urbanization data using rural–urban interaction model and logistic growth model. *Computers, Environment and Urban Systems, 45*, 89–100.

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

Jiang, L., O'Neill, B. C., Zoraghein, H., & Dahlke, S. (2020). Population scenarios for US states consistent with shared socioeconomic pathways. *Environmental Research Letters, 15*, Article 094097.

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

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

Merkle, M., Alexander, P., Brown, C., Seo, B., Harrison, P. A., Harmáčková, Z. V., Pedde, S., & Rounsevell, M. (2022). Downscaling population and urban land use for socio-economic scenarios in the UK. *Regional Environmental Change, 22*, 1–14.

Microsoft. (2019). *Microsoft/USBuildingFootprints. Github* [accessed 2023 March 04]. https://github.com/Microsoft/USBuildingFootprints.

Neis, P., & Zielstra, D. (2014). Recent developments and future trends in volunteered geographic information research: The case of OpenStreetMap. *Future Internet, 6*, 76–106.

O'Neill, B. C., Kriegler, E., Ebi, K. L., Kemp-Benedict, E., Riahi, K., Rothman, D. S., … Kok, K. (2017). The roads ahead: Narratives for shared socioeconomic pathways describing world futures in the 21st century. *Global Environmental Change, 42*, 169–180.

Rees, P., Van Der Gaag, N., De Beer, J., & Heins, F. (2012). European regional populations: Current trends, future pathways, and policy options. *European Journal of Population, 28*, 385.

Riahi, K., Van Vuuren, D. P., Kriegler, E., Edmonds, J., O'neill, B. C., Fujimori, S., … Fricko, O. (2017). The Shared Socioeconomic Pathways and their energy, land use, and greenhouse gas emissions implications: An overview. *Global Environmental Change, 42*, 153–168.

Riordan, E. C., & Rundel, P. W. (2014). Land use compounds habitat losses under projected climate change in a threatened California ecosystem. *PLoS One, 9*, Article e86487. https://doi.org/10.1371/journal.pone.0086487

Sadahiro, Y. (1999). Accuracy of areal interpolation: A comparison of alternative methods. *Journal of Geographical Systems, 1*, 323–346.

Samir, K. C., & Lutz, W. (2017). The human core of the shared socioeconomic pathways: Population scenarios by age, sex and level of education for all countries to 2100. *Global Environmental Change, 42*, 181–192.

Swanwick, R. H., Read, Q. D., Guinn, S. M., Williamson, M. A., Hondula, K. L., & Elmore, A. J. (2022). Dasymetric population mapping based on US census data and 30-m gridded estimates of impervious surface. *Scientific Data, 9*, 523.

USA Structures [WWW Document] URL https://gis-fema.hub.arcgis.com/pages/usa-structures, (2023) (accessed 12.15.22).

Van der Gaag, N., Van Imhoff, E., & Van Wissen, L. (2000). Internal migration scenarios and regional population projections for the European Union. *International Journal of Population Geography, 6*, 1–19.

Van Vuuren, D. P., Smith, S. J., & Riahi, K. (2010). Downscaling socioeconomic and emissions scenarios for global environmental change research: A review. *Wiley Interdisciplinary Reviews: Climate Change, 1*, 393–404.

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

Wan, H., Yoon, J., Srikrishnan, V., Daniel, B., & Judi, D. (2023). Landscape metrics regularly outperform other traditionally-used ancillary datasets in dasymetric mapping of population. *Computers, Environment and Urban Systems, 99*, Article 101899.

Wilson, T., & Bell, M. (2004). Comparative empirical evaluations of internal migration models in subnational population projections. *Journal of Population Research, 21*, 127–160.

Zandbergen, P. A. (2011). Dasymetric mapping using high resolution address point datasets. *Transactions in GIS, 15*, 5–27.

Zandbergen, P. A., & Ignizio, D. A. (2010). Comparison of dasymetric mapping techniques for small-area population estimates. *Cartography and Geographic Information Science, 37*, 199–214.

Zoraghein, H., & O'Neill, B. C. (2020). US state-level projections of the spatial distribution of population consistent with shared socioeconomic pathways. *Sustainability, 12*, 3374.