DEMOGRAPHIC RESEARCH
A peer-reviewed, open-access journal of population sciences

DEMOGRAPHIC RESEARCH

VOLUME 43, ARTICLE 54, PAGES 1563–1606
PUBLISHED 11 DECEMBER 2020
https://www.demographic-research.org/Volumes/Vol43/54/
DOI: 10.4054/DemRes.2020.43.54

Research Article

A spatial population downscaling model for integrated human-environment analysis in the United States

Hamidreza Zoraghein

Brian C. O'Neill

Contents

Demographic Research: Volume 43, Article 54
Research Article

A spatial population downscaling model for integrated human-environment analysis in the United States

Hamidreza Zoraghein[1]

Brian C. O'Neill[2]

Abstract

BACKGROUND
Spatial population models are important to inform understanding of historical demographic development patterns and to project possible future changes, especially for use in anticipating environmental interactions.

OBJECTIVE
We document, calibrate, and evaluate a high-resolution gravity-based population downscaling model for each US state and interpret its historical urban and rural spatial population change patterns.

METHODS
We estimate two free parameters that govern the spatial population change pattern using the historical population grids of each state. We interpret the resulting parameters in light of the spatial development pattern they represent. We evaluate the model by comparing the resulting total population grid of each state in 2010 against its census-based grid. We also analyze the temporal stability of parameters across the 1990–2000 and 2000–2010 decades.

RESULTS
Our analysis indicates varying levels of performance across states and population types. While our results suggest a consolidated change pattern in urban population across states, rural population change patterns are diverse. We find urban parameters are more stable.

CONCLUSIONS
The model's adaptability, performance, and interpretability indicate its potential for depicting historical state-level spatial population changes. It assigns these changes to different representative categories to assist interpretation.

[1] Population Council, New York, NY, USA. Email: hzoraghein@popcouncil.org.
[2] Pardee Center for International Futures, Josef Korbel School of International Studies, University of Denver, Denver, CO, USA. Currently at Joint Global Change Research Institute, Pacific Northwest National Laboratory, College Park, MD, USA.

CONTRIBUTION

We document and evaluate a gravitational model as well as investigate historical state-level spatial population changes. This research facilitates future work creating projections of the spatial distribution of population at the subnational level, especially those according to the Shared Socioeconomic Pathways (SSPs), widely used scenarios for climate change research.

1. Introduction

Projections of population, both in the form of its aggregate size and spatial distribution, are critical for modeling land-use/land-cover change, urbanization, vulnerability assessment, and sustainable development. For example, spatial population dynamics are a key driver of land-use/land-cover change, which can happen either directly through conversion to residential, industrial, or commercial lands, or indirectly by conversion of different types of land cover to agricultural uses (Bierwagen et al. 2010; Braimoh and Onishi 2007; Gao and O'Neill 2019; Meiyappan et al. 2014; Verburg et al. 2004). In addition, projecting changes in spatial population distribution is an essential element in anticipating future exposure of the population to changes in natural hazards ensuing from climate change, including flooding (Jongman et al. 2015), wildfires (Knorr, Jiang, and Arneth 2016), sea level rise (Hardy and Hauer 2018; McGranahan, Balk, and Anderson 2007; Neumann et al. 2015; Reimann, Merkens, and Vafeidis 2018), heat waves (Georgescu et al. 2014; Jones et al. 2015), and epidemiological events such as outbreaks of vector-borne diseases (Caminade et al. 2014; Hales et al. 2002). Population projection models that can be modified to produce alternative futures consistent with broader socioeconomic scenarios such as the Shared Socioeconomic Pathways (SSPs) (O'Neill et al. 2017) facilitate the integration of population changes with assessments of population vulnerability, which are critical for preparedness and adaptation measures (Rohat 2018).

Well-informed outlooks for future risks and adaptation planning are especially consequential in areas that experience rapid population growth and urbanization. These areas may exert ecological and socioeconomic pressure on their inhabitants and surroundings (Jones and O'Neill 2013), which can be manifested as threats to protected lands and biodiversity (Güneralp and Seto 2013) or elevated energy demands and emissions (Dodman 2009; Ewing and Rong 2008; Raupach, Rayner, and Paget 2010; Zhang et al. 2018), to name a few.

There are several ways to model the spatial distribution of population, with varying levels of complexity. The approaches can be as simple as using population fixed at the

current level (Gasparrini et al. 2017; Hanasaki et al. 2013), or scaling the existing spatial distribution of population proportional to aggregate national projections (Dong et al. 2015; Lehner and Stocker 2015). More complex approaches include those employing gravitational equations (Grübler et al. 2007; Jones and O'Neill 2013) and multivariable intelligent dasymetric modeling (McKee et al. 2015).

Spatially explicit population projections were first generated mostly for regional or city scales (Ballas, Clarke, and Wiemers 2005; Landis 1994; Stimson et al. 2012; Zwick and Carr 2006). However, the emergence of large-scale consistent population datasets such as the Gridded Population of the World (GPW) (CIESIN 2018), Global Rural-Urban Mapping Project (GRUMP) (Balk et al. 2006), LandScan Global (Dobson et al. 2000), and LandScan USA (Bhaduri et al. 2007), as well as remotely sensed land-cover/land-use products such as the National Land-Cover Database (NLCD) (Homer et al. 2015) and Global Human Settlement Layer (GHSL) (Pesaresi et al. 2016) has facilitated the development of large-scale (global/national) spatially explicit population projections (Jones and O'Neill 2013, 2016; McKee et al. 2015), which have been especially useful for large scale environmental issues such as climate change (Jones et al. 2015).

In this paper we document, detail, and evaluate a spatial population model for the United States that produces projections at high resolution (1km) tailored to each state. The model takes state population as an input and produces a projection of spatial distribution within the state that is consistent with the aggregate state total. Model parameters governing the spatial pattern of the development produced by the model can be calibrated (or specified) separately for each state. The combination of flexibility in representing different spatial development patterns in different states, a uniform approach and model structure for all states, and comprehensive national coverage of all population and land area makes the model also well suited to national-level studies that require spatial population projections as one input to integrated analyses. As noted above, examples of such analyses include projections of spatial land use in the United States and its environmental consequences; understanding potential future population exposure and vulnerability to natural hazards, including those related to climate change; and anticipating spatial patterns of energy demand and pollutant emissions. This work fills a gap that exists between large-scale global and national spatial population projection models that lack subnational subtleties, making them too generalized for local analysis (Bengtsson, Shen, and Oki 2006; Jones and O'Neill 2016; van Vuuren, Lucas, and Hilderink 2007), and local population projections that do not have sufficient spatial coverage to be used in studies with national scope (Ballas, Clarke, and Wiemers 2005; Stimson et al. 2012).

Our work is founded on the gravity-based population downscaling model developed by Jones and O'Neill (2013), based on earlier work by Grübler et al. (2007). This model has several advantages. First, it relies on multiple ancillary datasets that make the model

adaptive and able to be improved as new data become available. Second, the well-defined structure of the model makes it easy to adapt to different study areas and requirements. Third, its underlying gravitational form employs two parameters that are estimated based on historical population data. These parameters characterize different patterns of spatial population change, leading to its distinctive and data-driven interpretation in a study area. This is an important feature of this model, allowing it to be modified to reflect different types of spatial population change, and therefore well suited to generate population projections consistent with a range of alternative socioeconomic scenarios in the future. We summarize several new characteristics of the model that distinguish it from the previous works by Jones and O'Neill (2013, 2016).

- It is applied to all 50 states individually, rather than at the national or regional level.
- It downscales aggregate state-level population to higher resolution (1km) grids using updated historical population data based on the highest resolution spatial units disseminated by the US Census.
- The implementation of the model is parallelized to accommodate the parameter calculation at 1km resolution.
- The spatial mask layer is created separately for each state using more accurate national and state-level datasets.
- The method of parameter estimation is modified and includes two steps to thoroughly search for optimal values within a range informed by parameter interpretation.
- An explicit interpretation of parameter values is provided, reflecting implied patterns of spatial population change.
- Model performance is evaluated against observed changes over a recent decade.
- A test of parameter stability over time is provided.

We apply the model separately to the rural and urban population of each state in 2000 and 2010, leading to interpretations of state-specific historical rural and urban population change patterns during this period based on the values of two parameters that are estimated for each state. We evaluate the model on a state-by-state basis by comparing spatial model projections for 2010 with spatial data based on disaggregating census blocks.

This paper seeks to provide the research community with a documented and useful tool that, based on historical experience, can produce high resolution spatial distributions of population for each US state consistent with the IPCC's shared socioeconomic pathways (SSPs). We elucidate several aspects of this tool and evaluate it using historical census data, leaving producing scenario-based state-level population projections for our future work. Therefore, the two primary contributions of this paper are the detailed

specification of the model and its evaluation, in conjunction with the interpretations it provides of historical urban and rural spatial population patterns in each state.

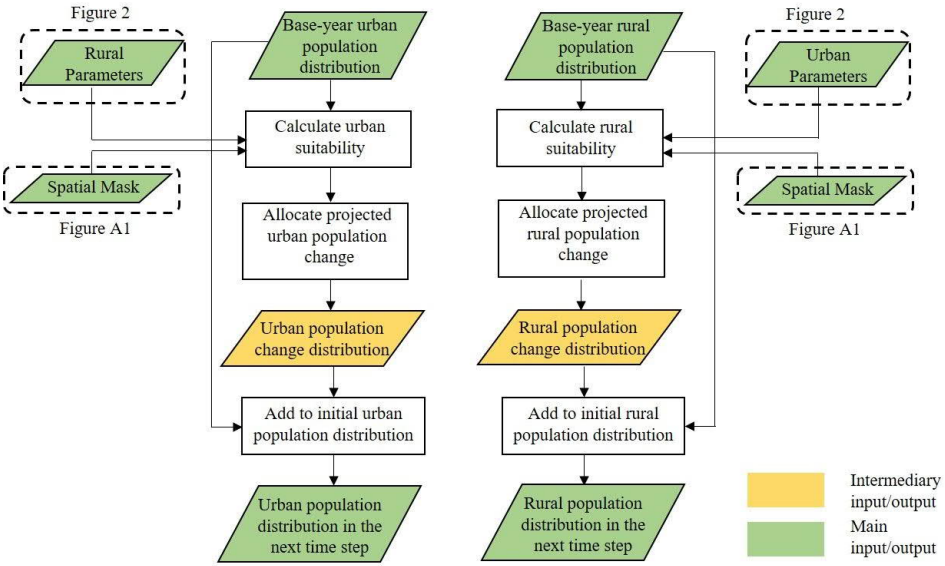
2. Methodology

2.1 Overview

The model takes a spatial downscaling approach that converts aggregate population change of an area over a given time period to changes in the population of grid cells within it. The basic approach was developed in Grübler et al. (2007) and further refined by Jones and O'Neill (2013), who particularly added the estimate of model parameters from historical data. In our application, aggregate population change occurs at the level of US states for two separate populations (rural and urban), the time step is 10 years, and the grid cell resolution is 1km. Rural and urban populations can coexist within a given grid cell. The allocation of the aggregate population changes is based on rural and urban suitability values calculated per cell (discussed below), which determines the attractiveness of the cell to gain population. When population increases, aggregate population change is allocated to grid cells proportional to the suitability of each cell. When it decreases, the aggregate change is allocated proportional to the inverse of the suitability of each cell. This approach is consistent with the assumption that the most attractive cells should gain the most population when it increases and lose the least population when it diminishes. Figure 1 illustrates how the model works.

The central element of the model is the determination of the suitability surface, which is done using a gravity-based, parameterized, negative exponential equation adapted from similar models used in transportation, urban, and economic geography. In addition, a spatial mask is used to modify suitability to account for physical or political constraints on population location. Parameters in the suitability equation are estimated from historical data to determine how the existing population configuration influences the spatial distribution of population in the next time step. In particular, the approach accounts for the role of population agglomerations and their proximity in determining where and to what extent population changes are likely to occur.

Figure 1: High-level illustration of the population downscaling model



The model is composed of calibration and projection components. The first component estimates model parameters based on historical population grids, and the second produces population projection grids. In this paper we focus on the calibration component and apply it to each US state to analyze different types of rural and urban population change patterns. The lessons we learn from this stage provide a basis for applying the model to the development of alternative future scenarios and their corresponding population grids. The input data of the model can be found at Zoraghein and O'Neill (2020), and the code for the calibration component that generated the resulting parameters is available from Zoraghein, O'Neill, and Vernon (2020).

2.2 Suitability

The suitability value of a cell is a numeric proxy for all qualities such as its local amenities, network connectivity, and economic opportunities that make it attractive or repellent to population change. It is assumed that suitability can be modeled as a function of the population in surrounding cells, as well as by accounting for physical or other

constraints on population location. Equation 1 shows the mathematical notation of the gravity-based equation used to derive the suitability value of each cell.[1]

$$v_i = l_i \sum_{j=1}^{n} P_j^{\alpha} \times e^{-\beta d_{ij}} \quad (1)$$

In Equation 1, v_i is the suitability value estimated for the focal cell i (for either urban or rural population, a distinction not represented here); l_i is the mask value modifying the suitability of the focal cell, depending on its topographic and other characteristics (described in the next section); P_j is the total population of the neighboring cell, j; and d_{ij} is the distance between the focal cell and its neighboring cell. The summation over j is performed for n cells contained in the neighborhood within100km of the focal cell, following previous work (Jones and O'Neill 2013, 2016) on representing a distance estimate over which existing amenities are influential in attracting population in the United States (Santos et al. 2011). Consequently, the neighborhood for each cell includes all cells within the state and those in other states that fall within the 100km buffer.

The α and β parameters govern the importance of existing surrounding population concentrations (within the neighborhood defined by n) and their accessibility (a function of distance) in determining the suitability value, respectively. We detail the interpretation of these two parameters in Section 2.5. Although the model calculates the two parameters separately for the rural and urban population, the population element of Equation 1 is the total population. This implies that the suitability of the rural and urban population allocation to a given cell is associated with the presence of both population types in surrounding cells.

2.3 Spatial mask

We derived a mask value for each cell to exogenously constrain population allocation according to physical barriers such as elevation, slope and land-cover, and mandates determined by both federal and state governments in the form of preserved areas. While Jones and O'Neill (2016) employ global datasets to derive a global mask layer, and Jones and O'Neill (2013) use more generalized and coarser resolution datasets to create their mask layer for the coterminous United States, we used high resolution spatial datasets with a more diverse set of categories that constrain population settlement. The combinatory mask value ranges from 0 to 1, with 0 being completely unsuitable for population settlement and 1 indicating no constraints.

[1] The equation used in Jones and O'Neill (2013, 2016) also includes a border adjustment factor (a). However, we decided not to use it as the inclusion of that factor increases the processing time while not changing estimated parameters significantly (B. Jones, pers. comm.).

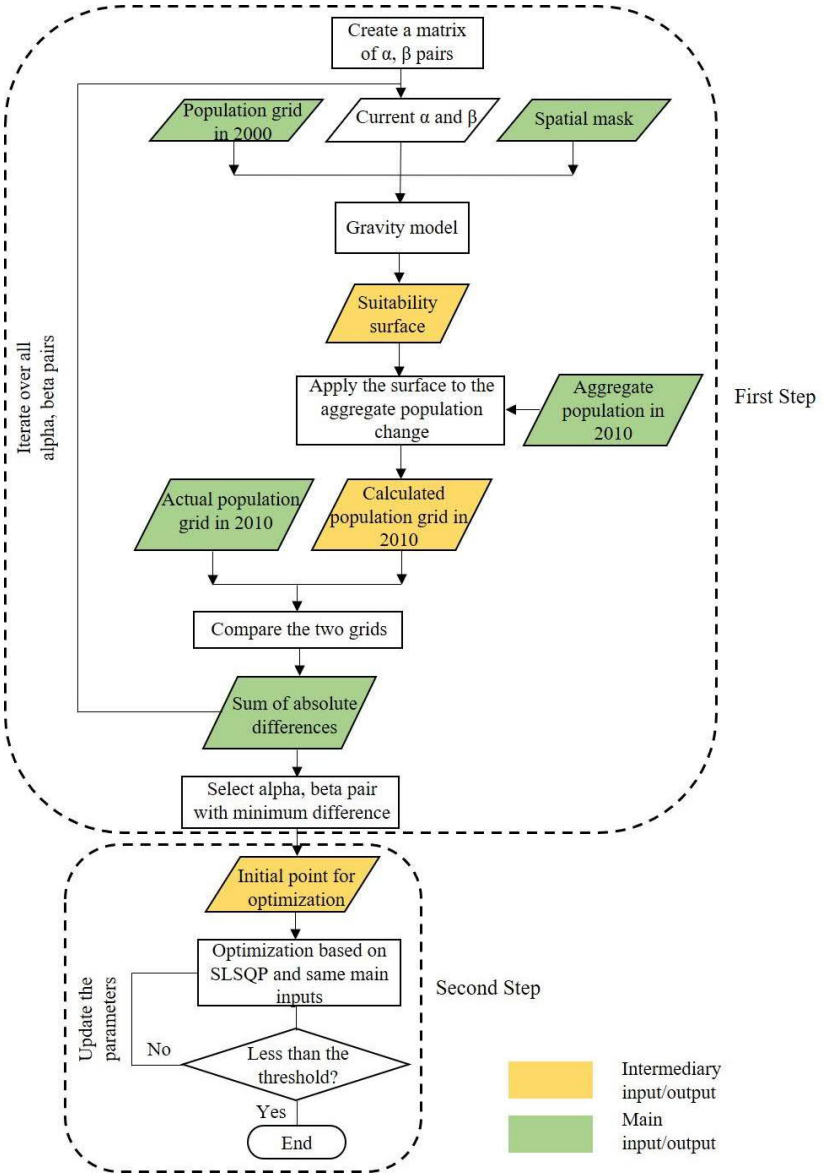
We either established new decision rules or followed the recommendations of previous studies to quantify the influence of each contributing factor on the mask value. For elevation we established a two-threshold approach: if the highest elevation in a state was lower than 1000 meters we set 1000 meters as the elevation threshold for that state. Otherwise, we derived the maximum elevation across populated census blocks from Census 2000 in each state and used that value as its elevation threshold. This two-step decision reflects our distinction between states in which elevation is not a prohibitive factor due to their mild topography and those with topographical barriers to the population allocation that should specifically be addressed.

For slope, we used 25% as the threshold beyond which population allocation is not allowed, following previous work (Jones and O'Neill 2013, 2016). We excluded open water, perennial ice/snow, and wetlands as uninhabitable land-cover types. We incorporated federal land mandates by setting Department of Defense lands, national forests, national wildlife refuges, and national parks as not allowable for population settlement. Finally, we also treated state parks, county parks, airports, golf courses, and cemeteries as uninhabitable. Figures A-1 and A-2 in the Appendix illustrate the steps to create the spatial mask layer and how the combinatory mask value is calculated for a cell, respectively. Moreover, we have listed the specific data sources in Table A-1 of the Appendix.

2.4 Parameter estimation

The calibration component of the model estimates the rural and urban α and β parameters for each state using historical rural and urban population grids in 2000 and 2010. We created these historical grids using census blocks as the smallest set of enumeration units with mutually exclusive rural/urban categories. We describe creating these historical grids in the Appendix. Parameters are estimated at values that, when applied to the population distribution of a given state in 2000, minimize the sum of absolute differences between projected and actual grids for that state in 2010 (Figure 2).

Figure 2: The calibration process of the gravity-based population downscaling model



Previous work (Jones and O'Neill 2013, 2016) used the Generalized Reduced Gradient (GRG) algorithm to solve an unconstrained local optimization problem. We modified this approach in three ways. First, we used the Sequential Least SQuares Programming (SLSQP) (Kraft 1988) method rather than GRG for local optimization because it was faster than other algorithms in reaching a similar pair of optimized parameter values. Second, we treated the problem as a constrained rather than unconstrained optimization, since meaningful limits on parameter values could be set that reduced computational time and improved the ability to find a globally optimal solution. Boundaries for the two parameters (alpha = [–2.0, 2.0] and beta = [–0.5, 2.0]) were identified as thresholds beyond which changes to the parameters were not meaningful given the model structure. For example, high values of β heavily discount the population of surrounding cells, and with values above 2 the effective distance within which surrounding cells matter to the suitability of the focal cell is not more than a single cell (1km, see Figure A-4). Given that beta values equal to or greater than 2 translate to a similar interpretation of the influence of neighboring cells, we set its maximum threshold to 2. On the other hand, large negative values of the parameter exponentially intensify the influence of large distances. Our experiments with incorporating α and β beyond these thresholds resulted in negligible changes in the value of the objective function while increasing the processing time of the optimization considerably.

Third, in order to produce the best initial value for the local optimization, we first generated a matrix of parameter values by dividing [–1.0, 1.0] into 10 intervals for α, and [0, 1.0] into 5 intervals for β. We used these subsets of the final ranges to establish a high-resolution matrix of pairs that were likely to serve as relevant candidate initial points for the second step of the optimization. We calculated the objective function value for each combination of parameters and selected the pair with the lowest objective function value

2.5 Interpretation of parameters

The α parameter indicates the degree to which the population size of surrounding cells translates into the suitability of a focal cell. A positive value indicates that the larger the population that is located within the 100km neighborhood of a focal cell, the more suitable it is for population allocation (while a negative value of alpha would imply a less suitable focal cell). A value of 1 indicates that the contribution of surrounding cells is proportional to their population size. Values greater than 1 indicate that the population of surrounding cells has a strong (more than proportional) effect on the suitability of a focal cell, meaning that new population development would occur predominantly in or very near to already-settled areas. By contrast, values of alpha below 1 indicate that the population of surrounding cells has a weaker (less than proportional) effect on the

suitability of a focal cell, meaning that new population development would occur less strongly in already-settled areas.

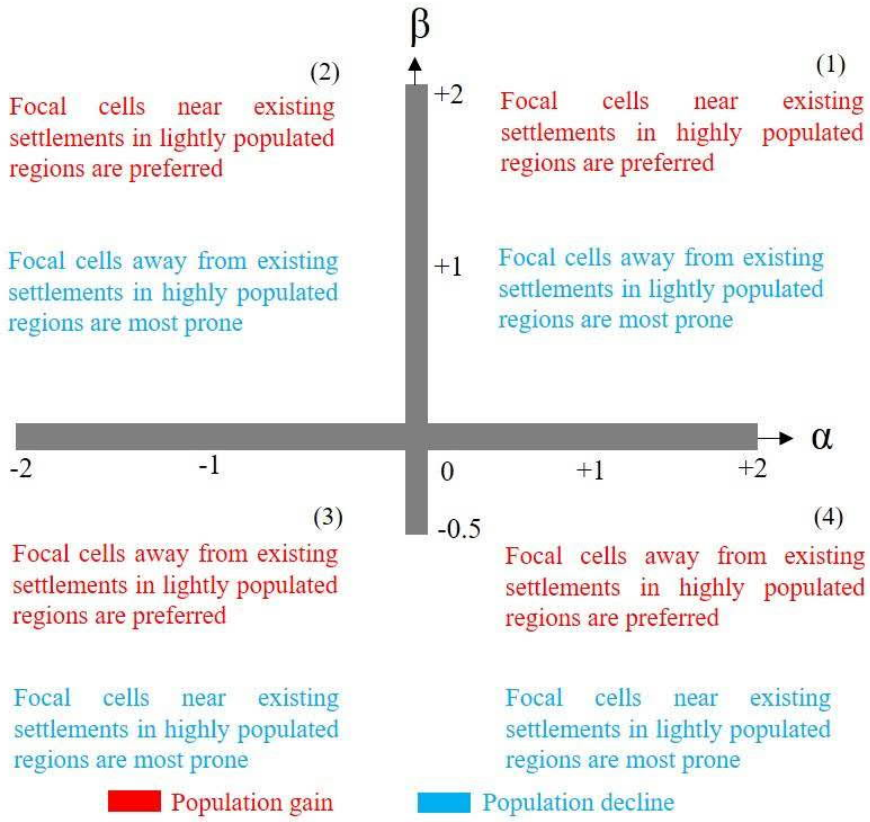
The β parameter reflects the significance of distance to surrounding cells for the suitability of a focal cell. It should be seen within the context of the window or neighborhood structure of the suitability equation: the characteristics of cells more than 100km from the focal cell have no effect on the suitability of that cell. Within 100km, β determines how distance modifies the effect on suitability. Because the exponent in Equation 1 is the negative of β, the higher the positive value of the parameter, the greater the deterrent effect of distance. For higher values of β, local population distributions prevail in determining the suitability of a focal cell and the presence of more distant (but still within 100km) population centers matters less. In other words, between two hypothetical focal cells, if the value of β is high, one with neighboring populous cells would be much more suitable for allocation of new population than one with similarly populous cells located farther away. By contrast, negative values of the parameter imply a lower friction of distance, so that the population characteristics of distant cells matter more than the population of nearer cells to the suitability of a focal cell. When β is 0, it means distance does not matter, and each cell contributes to the suitability of the focal cell proportional to its population raised to α.

Figure 3 illustrates the influence of varying values of α and β on how attractive a focal cell is for population growth or how prone it is to population loss, depending on the direction of change in the aggregate population being downscaled. The quadrant in which estimated parameters of a state fall implies a broad characterization of the historical pattern of spatial population change it has experienced.

When parameters fall in the first (upper right) quadrant, it represents consolidated population growth because population gain tends to concentrate close to areas that are already populated. A positive value of α leads to a preference for cells with large populations in the surrounding region (within their 100km window), and a positive value of β implies that focal cells closer to existing settlements within that region are preferred.

When α is negative while β is still positive, as characterized by the second (upper left) quadrant, a low-density population growth pattern can be conceived. The negative value of α means that focal cells with low population within their 100km windows are preferred, while the positive β value implies that within that region, cells near the highest local population densities are preferred. According to this pattern, new population is more likely to be located close to existing small towns (i.e., the highest density locations within generally low-density areas).

Figure 3: The influence of α and β on the suitability of a focal cell



The third (lower left) quadrant represents rural development or new small settlement growth, which implies the growth of population far from existing population settlements. The negative α means that focal cells in low population regions are preferred, while the negative β implies that within that region, settlement away from existing (small) population centers is preferred.

Finally, the fourth (lower right) quadrant represents population sprawl. Population growth tends to occur in highly populated regions because the positive α favors focal cells with high populations within their 100km window. However, the negative β implies that within that area, focal cells farther away from populous centers are preferred because distance does not act as a prohibitive factor. In contrast to the consolidation pattern, the

sprawling pattern represented by this quadrant still favors growth around populous centers (positive α) within their 100km surrounding region but away from the centers.

We can derive a similar set of categorizations for population decline. If the parameters lie in the first quadrant they imply rural decline, as remote cells with low concentration of population in their surroundings are more prone to population loss. Parameters in the second quadrant indicate a pattern where cells in populated regions but away from populous centers lose more population. Due to the spatial arrangement this pattern creates, it represents consolidation-oriented decline. When parameters fall in the third quadrant they signify a high-density decline, as cells adjacent to populous centers in highly populated regions tend to lose more population. Finally, parameters in the fourth quadrant suggest a low-density decline, as cells close to small towns (population centers in low-density regions) are more prone to population loss.

These interpretations are broad characterizations, each representing the quintessential population settlement pattern associated with a quadrant. However, depending on where the parameters of a state fall within a quadrant, the degree to which its pattern follows these archetypes will differ.

3. Results and discussion

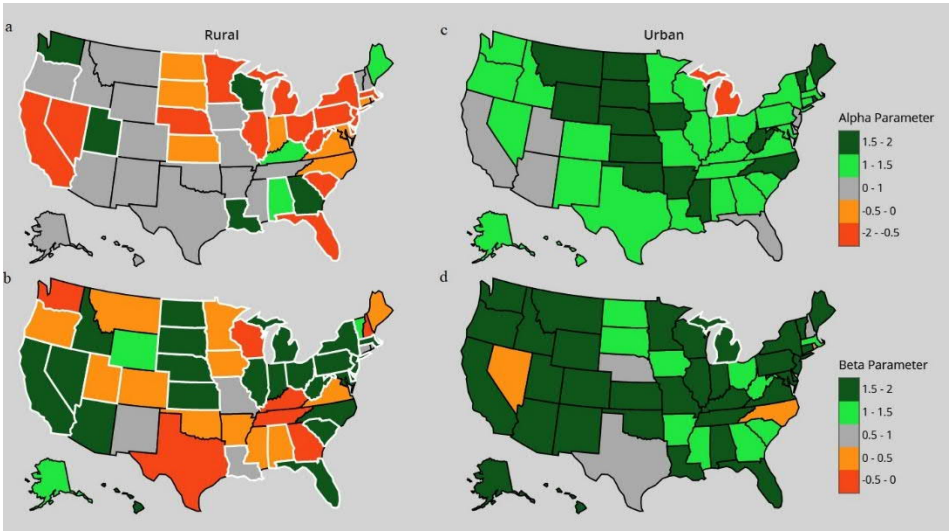
3.1 State-level parameters

Figure 4 shows state-level choropleth maps of the distribution of the estimated α and β parameters for the rural and urban population, while Table A-2 in the Appendix includes the values. For each population type, those states with population decline over the 2000–2010 decade are distinguished with white borders.

Increases and declines in population impose direct effects on the estimation of α and β parameters. Over the 2000–2010 decade, rural population diminished in 24 states, whereas Michigan was the only state with urban population loss. According to the census, there was no rural population in the District of Columbia.

Figures 5 and 6 illustrate scatterplots of resulting urban and rural parameters, respectively. We divided the plots into four quadrants consistent with Figure 3. We also differentiated states based on their sign of population change to reflect the points included in Figure 3.

Figure 4: State-level (a) α for rural population, (b) β for rural population, (c) α for urban population, and (d) β for urban population. Population-declining states have white borders



Both Figures 5 and 6 reveal that in many cases the optimal β value (and to a lesser extent the optimal alpha value) occurs at its maximum allowable limit. Typically, such a pattern implies that the constraint is overly restrictive and is determining the outcome. However, in this case our results indicate that in many instances the influence of the population of surrounding cells on the suitability of a focal cell is strictly limited to adjacent cells (or very remote cells when the inverse of suitability is considered in the case of population decline). This condition is represented by a very large β value, and the result is not sensitive to its precise value as long as it is at or above 2. Our approach limits the computation time spent searching for a precise value, while capturing the fact that a very large beta value is optimal and interpretable.

Figure 5 shows that all states with urban population growth fall in the first quadrant, suggesting a consolidated urban growth pattern. Higher values of α and β in this quadrant indicate stronger consolidation. On the other hand, lower values represent less significant dominance of populous clusters to absorb the growth and more diffuse settlement around clusters.

Figure 5: Scatterplot of state-level urban α and β parameters

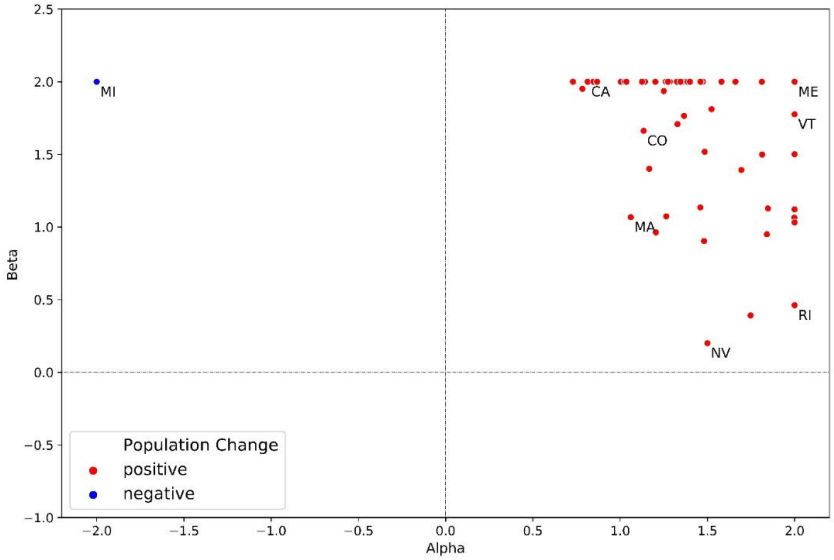


Figure 6: Scatterplot of state-level rural α and β parameters

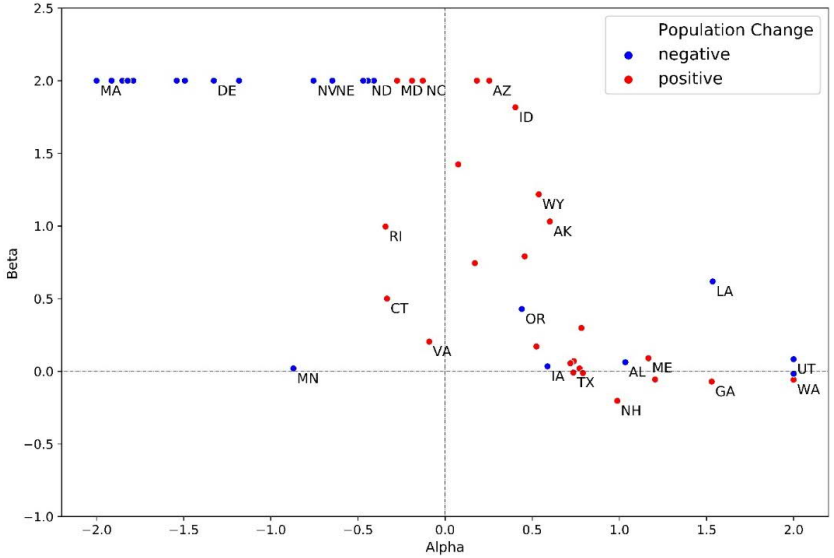
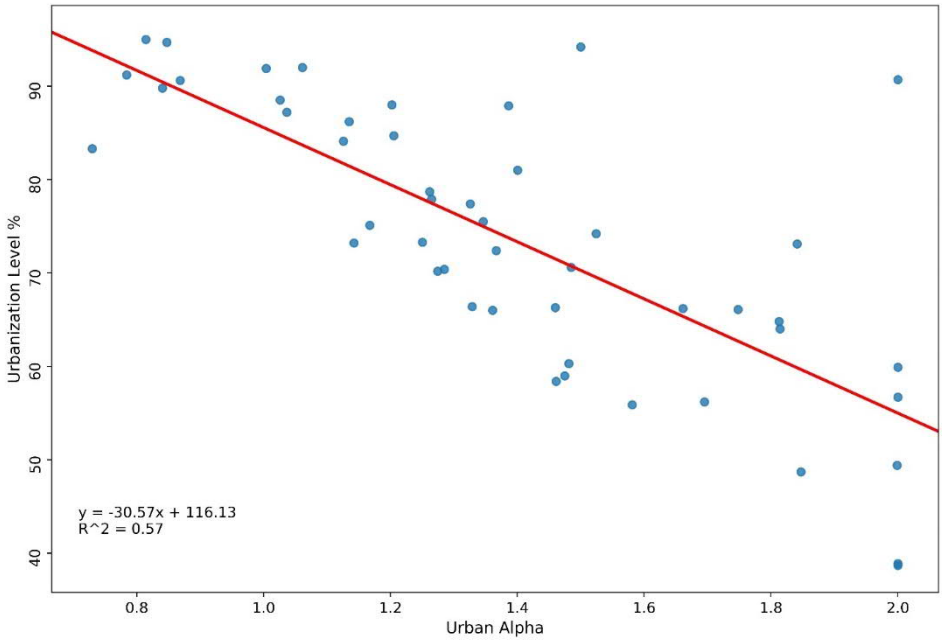


Figure 7 shows a negative association between urban α values and urbanization levels of states with growing urban populations. This suggests increasing returns to scale in states that have lower levels of urbanization, where urban growth occurs most strongly in the most heavily populated regions (100km windows). Conversely, in highly urbanized states, urban population growth is less strongly concentrated in the most highly populated regions. The only exception is Rhode Island with both high α and urbanization, which we suppose results from the small area of the state and its low β, allowing more flexible urban population settlement and high urbanization.

Figure 7: Relation of urbanization levels of US states in 2010 (except Michigan) to their estimated urban α parameters



Michigan, as the only state with urban population decline, lands in the second quadrant, suggesting a consolidation-oriented decline pattern. This is consistent with the historical pathway that Detroit has experienced as the dominant urban center in the state (McDonald 2014).

Figure 6 shows that states with the highest rural population growth – such as Alaska, Wyoming, Arizona, Maine, and Idaho – fall in the first quadrant, following the

consolidated pattern, with rural growth occurring preferentially in highly populated regions, but to varying degrees. Estimated β values also differ, suggesting variation in the importance of proximity to populated centers within these highly populated regions. The states with rural population gain and negative α parameters that follow low-density population growth patterns are Indiana, Maryland, North Carolina, Virginia, Connecticut, and Rhode Island. The low rural population growth in these states tends to settle close to existing population centers in low density regions. Texas, Kentucky, Tennessee, Georgia, New Hampshire, and Washington are six states with growing rural populations with their parameters in the fourth quadrant, pointing to a sprawling development pattern. In general, high negative β in this quadrant could suggest either a high level of accessibility for distant rural areas or urban to rural migration. However, the resulting β values are close to 0, except for New Hampshire, suggesting no significant preference on distance.

Many states with significant rural population decline such as Nevada, Nebraska, North Dakota, Kansas, and Massachusetts are situated in the second quadrant with the maximum allowable β and varying degrees of α. They follow the consolidation-oriented decline pattern, suggesting that rural cells in more highly populated regions but away from populated centers within those regions are most prone to population loss. Oregon, Iowa, Alabama, Louisiana, and Utah are five states with rural population decline that lie in the first quadrant, pointing to a pattern of decline in the most remote, low-density areas. The rural population decline in the majority of these states is small. The very low β for some of these states indicates the insignificant role of where lightly populated areas are in the surroundings of focal cells.

3.2 Spatial autocorrelation

We performed spatial autocorrelation analysis at both global and local scales by deriving the global Moran's I and LISA (Local Indicators of Spatial Association) measures (Anselin 1995) for four parameters, namely rural α, rural β, urban α, and urban β. Our definition of neighbors for each state in this analysis is consistent with that used in Equation 1 to derive its parameters, i.e., in addition to its adjacent states, the set extends to embody those states that intersect a bandwidth of 100km around a given state. Finally, we constructed a binary spatial weights matrix accordingly to formulate the spatial connectivity between states, and row-standardized it. Hawaii and Alaska were not part of the analysis due to lack of neighbors. In the 49 by 49 matrix, each state is represented by a row, with values that are either 0 or 1 divided by the number of its neighbors.

Figure A-5 shows the Moran scatterplots for all parameters. It demonstrates that no parameter has a strong global measure of spatial autocorrelation, indicating no overall spatial clustering for any of the parameters. However, local spatial autocorrelation in the

form of a few hot-spots and cold-spots exist for all parameters except the urban β, shown in Figure A-6.

3.3 Model evaluation

We evaluated the model by comparing the projected population grid in 2010 based on the optimal pair of parameters for each state with the corresponding observed (block-based) grid. We derived cumulative distribution functions (cdf) of errors for the urban, rural, and total population of each state. However, we emphasize total population results here since the primary output of the model is a set of total population grids (sum of urban and rural) to be employed in integrated human-environment analysis. In the resulting cdf plots, the horizontal axis represents the absolute values of percentage differences in 2010, while the vertical axis is the corresponding cumulative population percentage (that is, the percentage of the population in locations with less than a specific level of error). Errors are calculated based on mean population values over 10km by 10km windows to alleviate the spatial mismatch issue typical at the original 1km resolution. Figure 8 shows these plots, based on total population grids, for several states spanning a range of resulting cdfs, while Figure 9 includes those resulting from urban and rural population grids. Table A-3 in the Appendix lists absolute percentage differences corresponding to 50% and 90% of the population for all states. Moreover, Figure A-7 in the Appendix presents block-based and estimated total population grids for two sample states in 2010 as well as their mean population difference maps.

Plots with a narrower distribution indicate better model performance, since lower absolute percentage differences are associated with higher proportions of the state population, whereas wider distributions indicate the contrary. For instance, Figure 8 demonstrates that the model performs better in states such as Connecticut, Massachusetts, New Jersey, and New York than in Arizona, Nevada, and Texas.

Figure 8: CDF plots based on total population grids

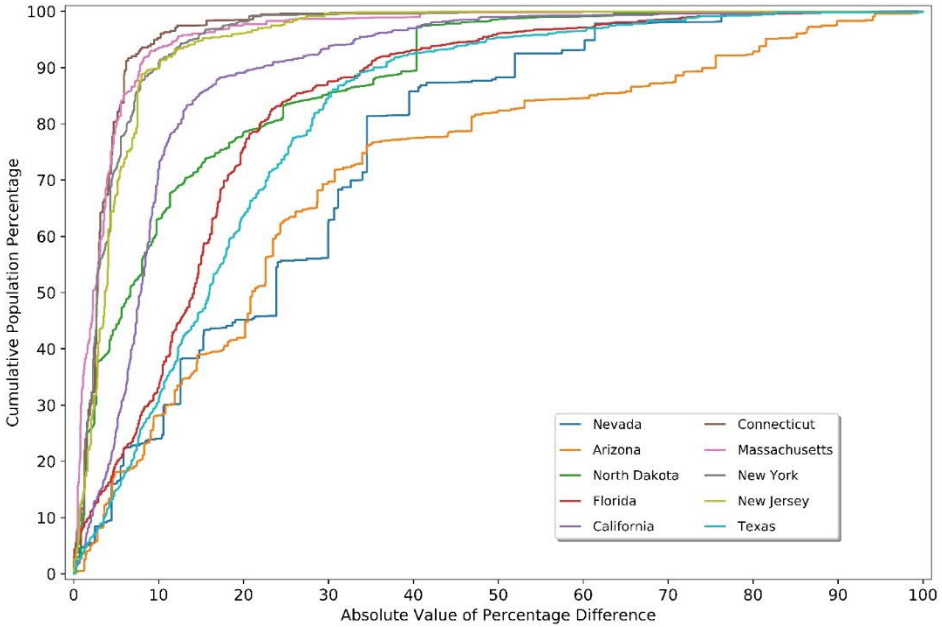
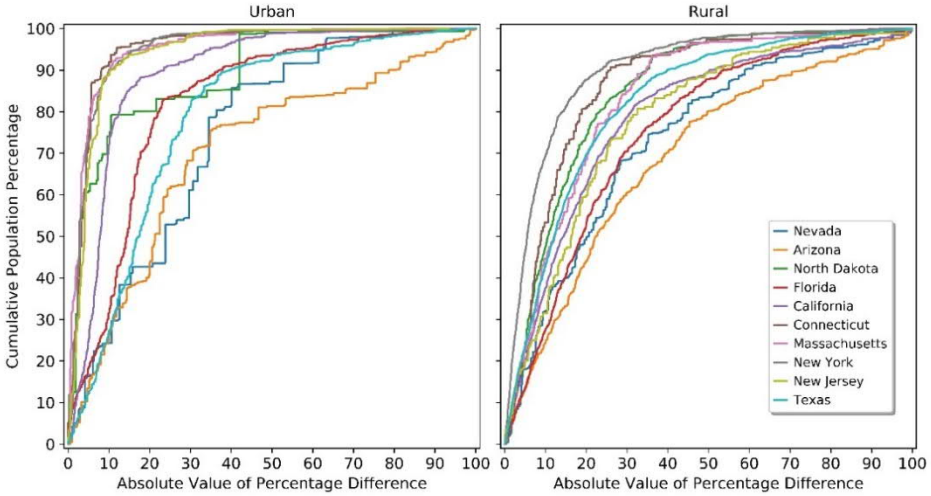


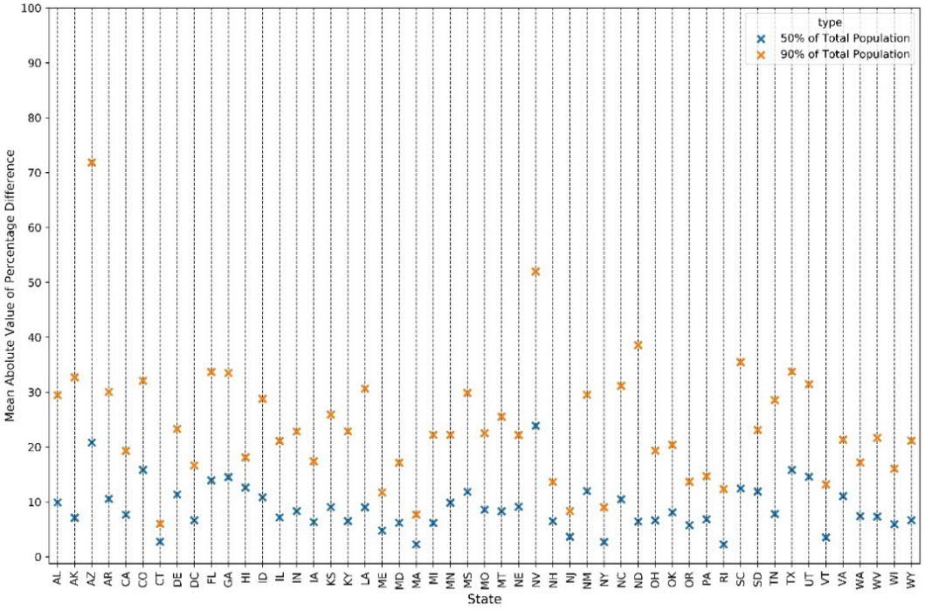
Figure 9: CDF plots based on urban and rural population grids



These plots and Table A-3 point to several observations. First, the model's performance is generally lower for rural populations. Rural blocks are usually larger that urban blocks with fewer residents, leading to higher uncertainty associated with rural population grids. Specifically, this different performance is more striking in states such as Florida, California, New Jersey, Massachusetts, New York, and Connecticut (see also Table A-3 in the Appendix). This may also result from high urbanization levels in these states, where higher absolute values of percentage differences in rural population do not influence the measure in total population. Second, for some states, although absolute percentage differences are mostly low, the existence of a few populated areas with relatively large errors lead to abrupt widening of their cdf. This, for example, explains the shape of the distribution for North Dakota in Figure 8. According to Table A-3, the absolute percentage difference corresponding to 90% of the state's population is six times higher than the value at 50% of the population. Third, the model's performance in Texas, Nevada, and Arizona is lower, which probably results from the model's inability to capture different patterns of spatial population change that had taken place in these expansive states. This contrasts with populous but relatively small states in the Northeast, where performance is generally high.

Overall, the results indicate that many states (plus the District of Columbia) have low absolute percentage differences at 50% and 90% of their population (Figure 10). Particularly, 34 and 49 states out of the total 51 are associated with absolute percentage differences lower than 10% and 20% at 50% of their population, respectively. Furthermore, the value is lower than 20% and 30% at 90% of the population in 18 and 38 states, respectively. This indicates this model's potential, tailored to incorporate US-specific input data, for generating more accurate high-resolution spatial distributions of population projections at the US state level than global models that have not been designed for subnational applications.

Figure 10: Absolute percentage differences at 50% and 90% of the total population in each state



3.4 Temporal stability

To analyze the temporal stability of rural/urban α and β parameters, we also estimated them using population grids in 1990 and 2000. Figure 11 shows the scatterplots of the estimated parameters from both decades with selected states labelled. Because reversal in the sign of the aggregate population change between the two decades inherently leads to disparate sets of parameters representing divergent narratives (see Section 2.5), we excluded such states from this analysis, reducing the number of states to 31 and 48 for the rural and urban population, respectively.

According to Figure 11, parameters are temporally stable for some states. Texas, Arizona, Montana, Alaska, and New Mexico are examples of states with stable rural parameters. Kentucky, New Hampshire, Nevada, North Dakota, and South Carolina represent those with stable urban parameters. On the other hand, the difference between resulting parameters for several states is substantial. For example, the rural parameters of New Jersey, Massachusetts, California, and Colorado, as well as the urban parameters of

North Carolina, Massachusetts, California, and Texas show substantial changes across the two decades.

Figure 11: Scatterplots of rural and urban parameters from both periods with some states labelled

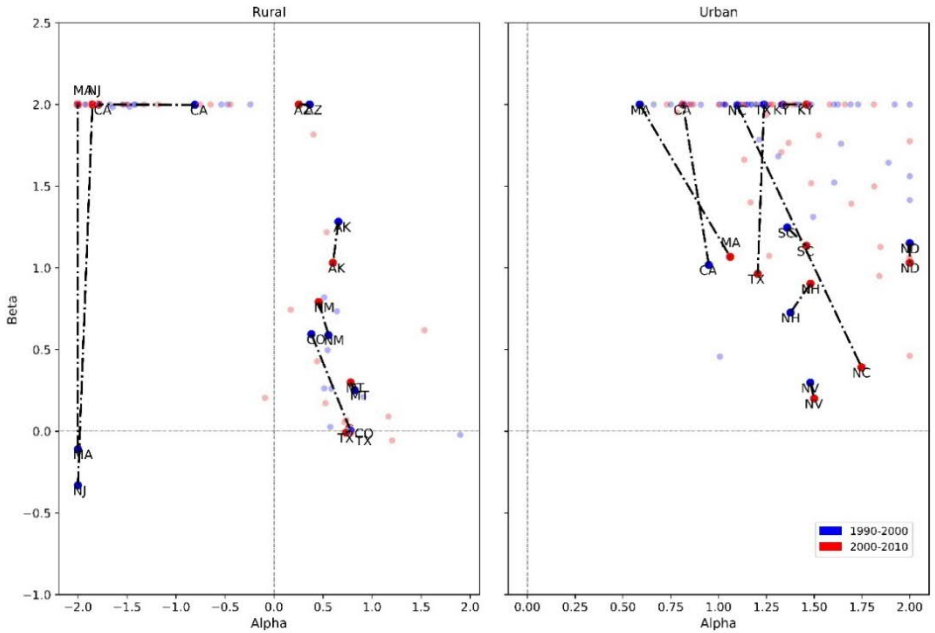


Table 1 summarizes the temporal stability analysis of the participating states. According to Table 1, the stability is higher for urban parameters. Around 88% and 79% of urban α and β values have changed less than 25% across the two decades, while these values are 58% and 45% for rural parameters. Some factors contribute to temporal changes in parameters in addition to the reversal in aggregate population change. First, the criteria for the urban definition employed by the US Census have evolved (Balk et al. 2018). This leads to temporally inconsistent depictions of rural/urban historical population grids in 1990, 2000, and 2010, which in turn might impact the estimation of parameters (Section 7 in the Appendix). This is aggravated for the rural population, as it is treated as the residual after the urban population is determined. Second, even when the sign of the aggregate population change remains intact, the settlement pattern inside the state might have varied, leading to temporal inconsistency in its parameters. For example, Figure 12 shows that the rural population decline in Massachusetts and New Jersey had

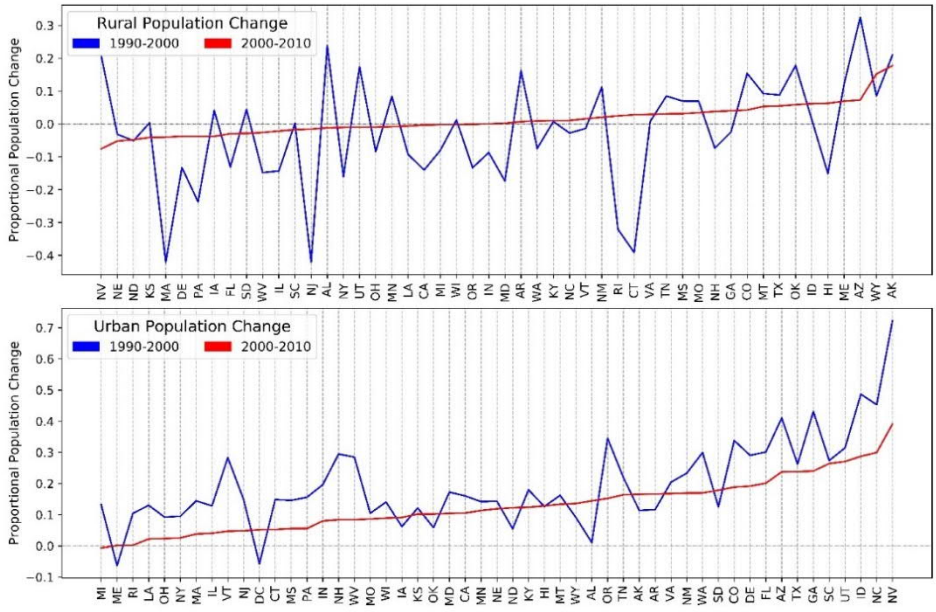
been more severe over the first decade than the second decade. This might have translated to the two separate patterns (high-density and consolidation-oriented decline) determined by the model across the two time periods.

Table 1: The number of states with parameter changes below 10% and 25% between the two decades

Population type	α Change < 10%	β Change < 10%	α Change < 25%	β Change < 25%
Rural[1]	11	12	18	14
Urban[2]	23	27	42	38

Number of participating states: 31.
Number of participating states: 48.

Figure 12: Proportional rural and urban population change of all states during 1990–2000 and 2000–2010 periods



4. Conclusion and future directions

In this paper we documented, calibrated, and evaluated a gravity-based population downscaling model for each US state. We completed this process for both rural and urban population over the 2000–2010 period based on the assumption that splitting the model into two parts would lead to more accurate total population grids, which will be the primary output of the model for integrated human-environment analysis in the United States. We contextualized parameters in terms of the type of spatial population change that has occurred in states, establishing a semantic framework and mapping pairs of state-level rural and urban parameters to distinctive categories of spatial population change such as consolidation, low-density, small-settlement, and sprawl. This interpretation depends on the direction of the aggregate population change.

We evaluated the model's performance in producing the observed spatial distribution of population in 2010. We found that for most states the absolute percentage difference, at 10km by 10km windows, is lower than 20% and 30% at 50% and 90% of their population. We also analyzed the temporal stability of parameters when estimated on data for the decade from 1990 to 2000. We observed higher temporal stability in urban parameters with changes mainly lower than 25% over time.

The evaluation of the model shows different levels of performance across states. There are several factors that impact the performance of the model in some states:

- Although we used the smallest set of census units, deriving historical population grids from these units is error-prone, especially in large rural blocks with low population. Moreover, spatial boundaries of census blocks change over time, resulting in different levels of uncertainty in allocating population to their constituent cells from one census year to another.
- The census's underlying urban/rural definition has not been consistent over time. Therefore, observed urbanization changes can occur through two different processes, one based on the actual urbanization growth and the other resulting from some areas being classified as urban according to the newer set of criteria. This makes it hard for the model to capture the net result of both with available parameters. This is aggravated for rural population as the census classification emphasizes urban areas and treat residuals as rural.
- Possible variation in the spatial population change at the sub-state level that cannot be represented with a single pair of parameters.
- Relationships between population change and other spatial development patterns such as land-use and economic activity that cannot be explained by the population-driven gravity model.

Overall, the model's ability to project a decade of spatial population change, its adaptability, and its interpretability in terms of archetypical patterns of spatial population change all lend confidence to its suitability for applications to generate alternative future spatial development scenarios, that encompass uncertainty in the range of possible future developments. This points to the model's potential implementation in integrated assessment applications such as risk assessment in relation to environmental hazards, resource allocation, and land-use/land-cover change analysis at the subnational scale. The parsimonious structure of the model allows for easy incorporation of different scenarios of spatial population change.

Our future work focuses on potential solutions to improve the model and generate a set of grid-level population projections according to distinctive socioeconomic scenarios such as the SSPs. There are several approaches that could improve the model's performance. First, instead of transferring the population of blocks to grid cells merely based on their overlapping areas, using large-scale ancillary variables such as Historical Settlement Data Compilation for the United States (HISDAC-US) (Leyk and Uhl 2018) and Microsoft building footprint layer[1] will lead to more precise population allocation, especially in large rural blocks. Second, developing consistent data-driven definitions of urban and rural land will alleviate the uncertainty arising from their current temporal incompatibility and provide more reliable historical urban and rural population grids. Third, including exogeneous contributors to the spatial population change, such as land-use, proximity to coastal areas, and economic activity, will enhance the model's performance. However, any modification in that direction should maintain the model's ability to easily adjust to alternative scenarios. Finally, based on this paper's explanatory framework that maps model parameters to different patterns of spatial population change, we will define scenarios and generate projections of the spatial distribution of population accordingly.

5. Acknowledgements

This research was supported by the US Department of Energy, Office of Science, as part of research in Multi-Sector Dynamics, Earth and Environmental System Modeling Program.

[1] https://github.com/microsoft/USBuildingFootprints

References

Anselin, L. (1995). Local indicators of spatial association – LISA. *Geographical Analysis* 27(2): 93–115.

Balk, D., Leyk, S., Jones, B., Montgomery, M.R., and Clark, A. (2018). Understanding urbanization: A study of census and satellite-derived urban classes in the United States, 1990–2010. *PLoS One* 13(12): e0208487. doi:10.1371/journal.pone.0208487.

Balk, D.L., Deichmann, U., Yetman, G., Pozzi, F., Hay, S.I., and Nelson, A. (2006). Determining global population distribution: Methods, applications and data. *Advances in Parasitology* 62: 119–156. doi:10.1016/S0065-308X(05)62004-0.

Ballas, D., Clarke, G.P., and Wiemers, E. (2005). Building a dynamic spatial microsimulation model for Ireland. *Population, Space and Place* 11(3): 157–172. doi:10.1002/psp.359.

Bengtsson, M., Shen, Y., and Oki, T. (2006). A SRES-based gridded global population dataset for 1990–2100. *Population and Environment* 28(2): 113–131. doi:10.1007/s11111-007-0035-8.

Bhaduri, B., Bright, E., Coleman, P., and Urban, M. (2007). LandScan USA: A high-resolution geospatial and temporal modeling approach for population distribution and dynamics. *GeoJournal* 69(1–2): 103–117.

Bierwagen, B.G., Theobald, D.M., Pyke, C.R., Choate, A., Groth, P., Thomas, J. V., and Morefield, P. (2010). National housing and impervious surface scenarios for integrated climate impact assessments. *Proceedings of the National Academy of Sciences* 107(49): 20887–20892. doi:10.1073/pnas.1002096107.

Braimoh, A.K. and Onishi, T. (2007). Spatial determinants of urban land use change in Lagos, Nigeria. *Land Use Policy* 24(2): 502–515. doi:10.1016/j.landusepol.2006.09.001.

Caminade, C., Kovats, S., Rocklov, J., Tompkins, A.M., Morse, A.P., Colón-González, F.J., Stenlund, H., Martens, P., and Lloyd, S.J. (2014). Impact of climate change on global malaria distribution. *Proceedings of the National Academy of Sciences* 111(9): 3286–3291. doi:10.1073/pnas.1302089111.

Columbia University – Center for International Earth Science Information Network – CIESIN (2018). Gridded Population of the World, Version 4 (GPWv4): Population Count, Revision 11.

Dobson, J.E., Bright, E.A., Coleman, P.R., Durfee, R.C., and Worley, B.A. (2000). LandScan: A global population database for estimating populations at risk. *Photogrammetric Engineering and Remote Sensing* 66(7): 849–857.

Dodman, D. (2009). Blaming cities for climate change? An analysis of urban greenhouse gas emissions inventories. *Environment and Urbanization* 21: 185–201. doi:10.1177/0956247809103016.

Dong, W., Liu, Z., Liao, H., Tang, Q., and Li, X. (2015). New climate and socio-economic scenarios for assessing global human health challenges due to heat risk. *Climatic Change* 130(4): 505–518. doi:10.1007/s10584-015-1372-8.

Ewing, R. and Rong, F. (2008). The impact of urban form on U.S. residential energy use. *Housing Policy Debate* 19(1): 1–30. doi:10.1080/10511482.2008.9521624.

Gao, J. and O'Neill, B.C. (2019). Data-driven spatial modeling of global long-term urban land development: The SELECT model. *Environmental Modelling and Software* 119: 458–471. doi:10.1016/j.envsoft.2019.06.015.

Gasparrini, A., Guo, Y., Sera, F., Vicedo-Cabrera, A.M., Huber, V., Tong, S., de Sousa Zanotti Stagliorio Coelho, M., Nascimento Saldiva, P.H., Lavigne, E., Matus Correa, P., Valdes Ortega, N., Kan, H., Osorio, S., Kyselý, J., Urban, A., Jaakkola, J.J.K., Ryti, N.R.I., Pascal, M., Goodman, P.G., Zeka, A., Michelozzi, P., Scortichini, M., Hashizume, M., Honda, Y., Hurtado-Diaz, M., Cesar Cruz, J., Seposo, X., Kim, H., Tobias, A., Iñiguez, C., Forsberg, B., Åström, D.O., Ragettli, M.S., Guo, Y.L., Wu, C. fu, Zanobetti, A., Schwartz, J., Bell, M.L., Dang, T.N., Van, D. Do, Heaviside, C., Vardoulakis, S., Hajat, S., Haines, A., and Armstrong, B. (2017). Projections of temperature-related excess mortality under climate change scenarios. *The Lancet Planetary Health* 1(9): e360–e367. doi:10.1016/S2542-5196(17)30156-0.

Georgescu, M., Morefield, P.E., Bierwagen, B.G., and Weaver, C.P. (2014). Urban adaptation can roll back warming of emerging megapolitan regions. *Proceedings of the National Academy of Sciences* 111(8): 2909–2914. doi:10.1073/pnas.1322280111.

Grübler, A., O'Neill, B., Riahi, K., Chirkov, V., Goujon, A., Kolp, P., Prommer, I., Scherbov, S., and Slentoe, E. (2007). Regional, national, and spatially explicit scenarios of demographic and economic change based on SRES. *Technological Forecasting and Social Change* 74(7): 980–1029. doi:10.1016/j.techfore.2006.05.023.

Güneralp, B. and Seto, K.C. (2013). Futures of global urban expansion: Uncertainties and implications for biodiversity conservation. *Environmental Research Letters* 8(1). doi:10.1088/1748-9326/8/1/014025.

Hales, S., De Wet, N., Maindonald, J., and Woodward, A. (2002). Potential effect of population and climate changes on global distribution of dengue fever: An empirical model. *Lancet* 360(9336): 830–834. doi:10.1016/S0140-6736(02)09964-6.

Hanasaki, N., Fujimori, S., Yamamoto, T., Yoshikawa, S., Masaki, Y., Hijioka, Y., Kainuma, M., Kanamori, Y., Masui, T., Takahashi, K., and Kanae, S. (2013). A global water scarcity assessment under Shared Socio-economic Pathways – Part 2: Water availability and scarcity. *Hydrology and Earth System Sciences* 9(12): 2393–2413. doi:10.5194/hess-17-2393-2013.

Hardy, R.D. and Hauer, M.E. (2018). Social vulnerability projections improve sea-level rise risk assessments. *Applied Geography* 91: 10–20. doi:10.1016/j.apgeog.2017.12.019.

Homer, C.G., Dewitz, J.A., Yang, L., Jin, S., Danielson, P., Xian, G., Coulston, J., Herold, N.D., Wickham, J.D., and Megown, K. (2015). Completion of the 2011 National Land Cover Database for the conterminous United States-Representing a decade of land cover change information. *Photogrammetric Engineering and Remote Sensing* 81(5): 345–354. doi:10.14358/PERS.81.5.345.

Jones, B. and O'Neill, B.C. (2013). Historically grounded spatial population projections for the continental United States. *Environmental Research Letters* 8(4): 044021. doi:10.1088/1748-9326/8/4/044021.

Jones, B. and O'Neill, B.C. (2016). Spatially explicit global population scenarios consistent with the Shared Socioeconomic Pathways. *Environmental Research Letters* 11(8). doi:10.1088/1748-9326/11/8/084003.

Jones, B., O'Neill, B.C., Mcdaniel, L., Mcginnis, S., Mearns, L.O., and Tebaldi, C. (2015). Future population exposure to US heat extremes. *Nature Climate Change* 5(7): 652–655. doi:10.1038/nclimate2631.

Jongman, B., Winsemius, H.C., Aerts, J.C.J.H., Coughlan de Perez, E., van Aalst, M.K., Kron, W., and Ward, P.J. (2015). Declining vulnerability to river floods and the global benefits of adaptation. *Proceedings of the National Academy of Sciences* 112(18): E2271–E2280. doi:10.1073/pnas.1414439112.

Knorr, W., Jiang, L., and Arneth, A. (2016). Climate, CO2 and human population impacts on global wildfire emissions. *Biogeosciences* 13(1): 267–282. doi:10.5194/bg-13-267-2016.

Kraft, D. (1988). A software package for sequential quadratic programming. (Forschungsbericht: Deutsche Forschungs- und Versuchsanstalt fur Luft- und Raumfahrt).

Landis, J.D. (1994). The California Urban Futures Model: A new generation of metropolitan simulation models. *Environment and Planning B: Planning and Design* 21(4): 399–420. doi:10.1068/b210399.

Lehner, F. and Stocker, T.F. (2015). From local perception to global perspective. *Nature Climate Change* 5(8): 731–734. doi:10.1038/nclimate2660.

Leyk, S. and Uhl, J.H. (2018). Data descriptor: HISDAC-US, historical settlement data compilation for the conterminous United States over 200 years. *Scientific Data* 5(180175). doi:10.1038/sdata.2018.175.

McDonald, J.F. (2014). What happened to and in Detroit? *Urban Studies* 51(16): 3309–3329. doi:10.1177/0042098013519505.

McGranahan, G., Balk, D., and Anderson, B. (2007). The rising tide: assessing the risks of climate change and human settlements in low elevation coastal zones. *Environment and Urbanization* 19(1): 17–37. doi:10.1177/0956247807076960.

McKee, J.J., Rose, A.N., Bright, E.A., Huynh, T., and Bhaduri, B.L. (2015). Locally adaptive, spatially explicit projection of US population for 2030 and 2050. *Proceedings of the National Academy of Sciences* 112(5): 1344–1349. doi:10.1073/pnas.1405713112.

Meiyappan, P., Dalton, M., O'Neill, B.C., and Jain, A.K. (2014). Spatial modeling of agricultural land use change at global scale. *Ecological Modelling* 291: 152–174. doi:10.1016/j.ecolmodel.2014.07.027.

Minnesota Population Center (2016). National Historical Geographic Information System: Version 11.0 [Database]. University of Minnesota. doi:10.18128/D050.V11.0.

Neumann, B., Vafeidis, A.T., Zimmermann, J., and Nicholls, R.J. (2015). Future coastal population growth and exposure to sea-level rise and coastal flooding: A global assessment. *PloS One* 10(3): e0118571. doi:10.1371/journal.pone.0118571.

O'Neill, B.C., Kriegler, E., Ebi, K.L., Kemp-Benedict, E., Riahi, K., Rothman, D.S., van Ruijven, B.J., van Vuuren, D.P., Birkmann, J., Kok, K., Levy, M., and Solecki, W. (2017). The roads ahead: Narratives for shared socioeconomic pathways describing world futures in the 21st century. *Global Environmental Change* 42: 169–180. doi:10.1016/j.gloenvcha.2015.01.004.

Pesaresi, M., Ehrlich, D., Ferri, S., Florczyk, A., Freire, S., Halkia, M., Julea, A., Kemper, T., Soille, P., and Syrris, V. (2016). Operating procedure for the production of the global human settlement layer from Landsat data of the epochs 1975, 1990, 2000, and 2014. (JRC Technical Reports). Ispra: Joint Research Centre (IRC). doi:10.2788/253582.

Raupach, M.R., Rayner, P.J., and Paget, M. (2010). Regional variations in spatial structure of nightlights, population density and fossil-fuel CO2 emissions. *Energy Policy* 38(9): 4756–4764. doi:10.1016/j.enpol.2009.08.021.

Reimann, L., Merkens, J.L., and Vafeidis, A.T. (2018). Regionalized Shared Socioeconomic Pathways: narratives and spatial population projections for the Mediterranean coastal zone. *Regional Environmental Change* 18(1): 235–245. doi:10.1007/s10113-017-1189-2.

Rohat, G. (2018). Projecting drivers of human vulnerability under the shared socioeconomic pathways. *International Journal of Environmental Research and Public Health* 15(3): 554. doi:10.3390/ijerph15030554.

Rossiter, K. (2011). What are census blocks? https://www.census.gov/newsroom/blogs/random-samplings/2011/07/what-are-census-blocks.html.

Santos, A., McGuckin, N., Nakamoto, H.Y., Gray, D., and Liss, S. (2011). *Summary of travel trends: 2009 National Household Travel Survey*. Washington, D.C.: U.S. Department of Transportation Federal Highway Administration (FHWA).

Stimson, R., Bell, M., Corcoran, J., and Pullar, D. (2012). Using a large scale urban model to test planning scenarios in the Brisbane-South East Queensland region. *Regional Science Policy and Practice* 4(4): 373–392. doi:10.1111/j.1757-7802.2012.01082.x.

van Vuuren, D.P., Lucas, P.L., and Hilderink, H. (2007). Downscaling drivers of global environmental change: Enabling use of global SRES scenarios at the national and grid levels. *Global Environmental Change* 17: 114–130. doi:10.1016/j.gloenvcha.2006.04.004.

Verburg, P.H., Ritsema van Eck, J.R., de Nijs, T.C.M., Dijst, M.J., and Schot, P. (2004). Determinants of land-use change patterns in the Netherlands. *Environment and Planning B: Planning and Design* 31(1): 125–150. doi:10.1068/b307.

Zhang, G., Ge, R., Lin, T., Ye, H., Li, X., and Huang, N. (2018). Spatial apportionment of urban greenhouse gas emission inventory and its implications for urban planning: A case study of Xiamen, China. *Ecological Indicators* 85: 644–656. doi:10.1016/j.ecolind.2017.10.058.

Zoraghein, H. and O'Neill, B. (2020). Data Supplement: U.S. state-level projections of the spatial distribution of population consistent with Shared Socioeconomic Pathways. (Version v0.1.0) [Data set] [electronic resource]. doi:10.5281/zenodo.3756179.

Zoraghein, H., O'Neill, B.C., and Vernon, C. (2020). Population Gravity Model (Version v0.1.0) [electronic resource]. https://github.com/IMMM-SFA/population_gravity.

Zwick, P.D. and Carr, M.H. (2006). *Florida 2060: A population distribution scenario for the state of Florida*. Gainesville: University of Florida, GeoPlan Center.

Appendix

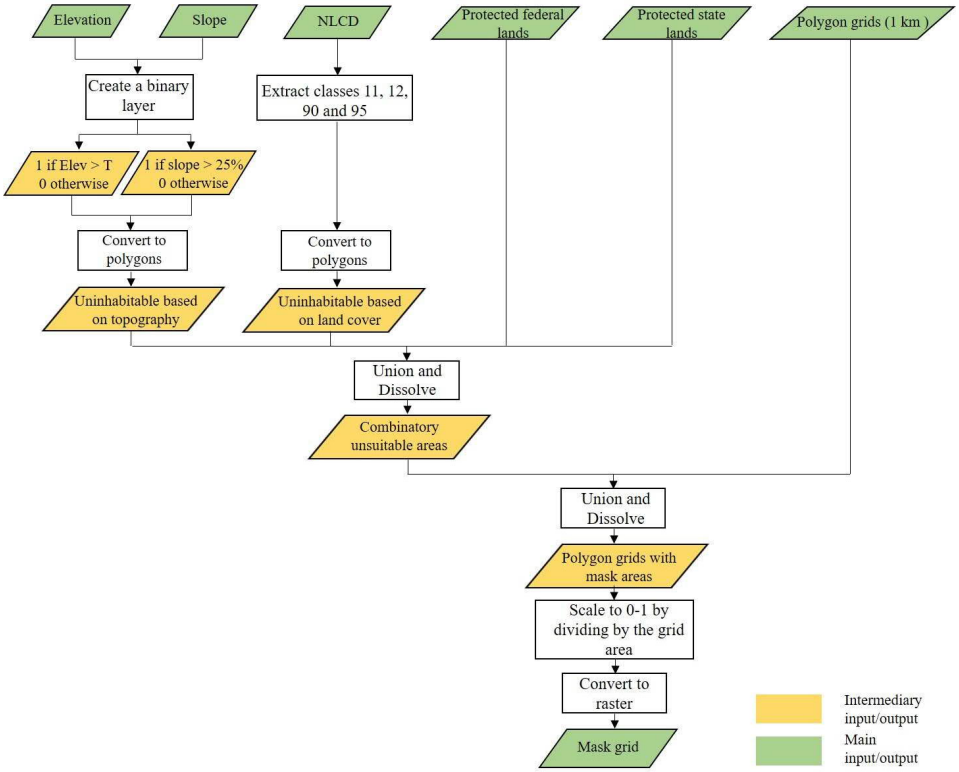
1. Spatial mask

The model considers exogenous factors that influence the suitability value of a cell. These factors are topography, land-use/land-cover, and policy mandates, which have varying effects on the potential settlement of a cell. Therefore, we first collected datasets that represent these determinants in each state. We then applied spatial overlay, aggregation, and polygon-to-raster operations to produce a 1km resolution combinatory spatial mask raster for each state. Table A-1 shows the datasets that were used to create the spatial mask layer, as well as their characteristics and sources. Moreover, Figure A-1 illustrates the process for generating the layer for each state.

Table A-1: Datasets that were used to create the state-level spatial mask layer

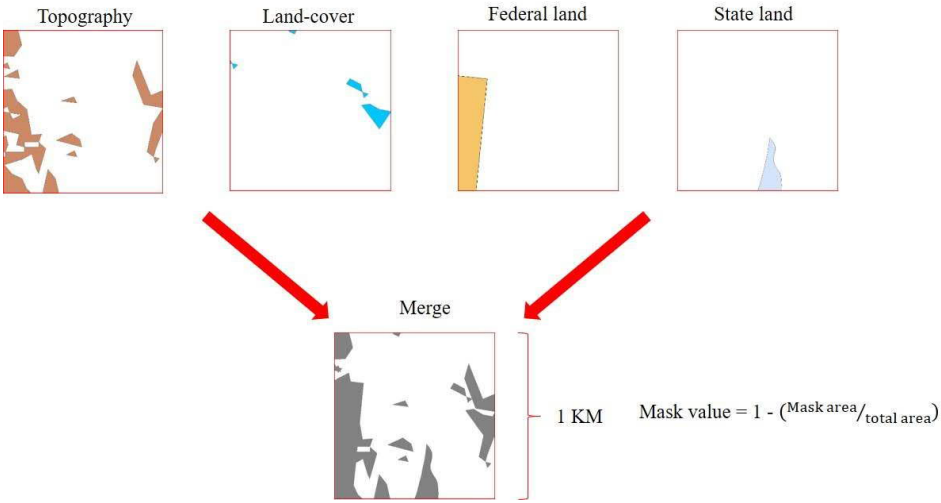
Factor	Dataset	Type	Source
Elevation and Slope	National Elevation Dataset	Raster (30m resolution)	https://datagateway.nrcs.usda.gov/
Land-cover	National Land-Cover Database	Raster (30m resolution)	https://www.mrlc.gov/
Federal mandates	Federal Lands	Polygon	https://hifld-geoplatform.opendata.arcgis.com/datasets/federal-lands
State mandates	State Lands	Polygon	https://www2.census.gov/geo/tiger/TIGER2016/AREALM/

Figure A-1: Steps to create the state-level spatial mask layer



Finally, Figure A-2 demonstrates the estimation of the mask value for a given cell.

Figure A-2: Schematic illustration of the calculation of the spatial mask value for a cell



2. Historical population grids

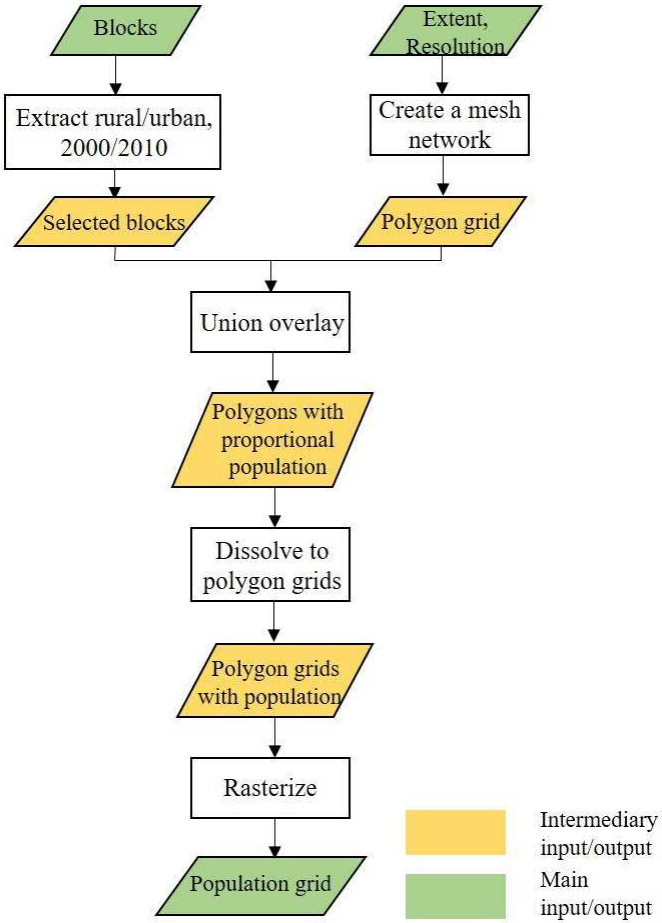
We used blocks – as the smallest aggregation unit disseminated by the US Census – to generate historical population grids at 1km resolution. By contrast, Jones and O'Neill (2013) use census tracts, census populated places, and county-level data, but transferred to 1/8 degree resolution, and Jones and O'Neill (2016) use the GRUMP and GPW V3 data, which are based on census tracts in the United States, also at 1/8 degree resolution. GRUMP takes a different approach to defining rural and urban population than we used in this paper. It employs global nighttime imagery to delineate urban lands, whereas our approach utilizes the US Census criteria applied to blocks.

Census blocks are statistical areas bounded by visible features such as roads, streams, and railroad tracks, as well as by non-visible boundaries such as property lines, city, township, school district, county limits, and short line-of-sight extensions of roads (Rossiter 2011). Although census blocks are generally small, their boundaries vary spatially. In urban areas they typically correspond to a city block bounded by surrounding streets. In rural areas, on the other hand, their spatial extent may be much larger, and in very remote areas they can encompass hundreds of square miles. The total population of

census blocks is also diverse, ranging from 0 in many cases to several hundred or even thousands in blocks located in densely populated cities. The number of census blocks has risen significantly in response to changes in population development and urbanization, increasing from just over 7 million in 1990 to over 11 million in 2010 (Balk et al. 2018).

Blocks are the only census spatial units that are either urban or rural, reflected in their rural/urban classification attribute. Therefore, it is possible to create two mutually exclusive sets of urban and rural blocks for both 2000 and 2010 using the tabular and spatial datasets provided by the Minnesota Population Center (2016). We rasterized these sets of urban and rural blocks to provide 1km resolution population grids, and then disaggregated them by state. Each grid cell can have both rural and urban population values. Figure A-3 demonstrates the process for generating historical rural/urban population grids.

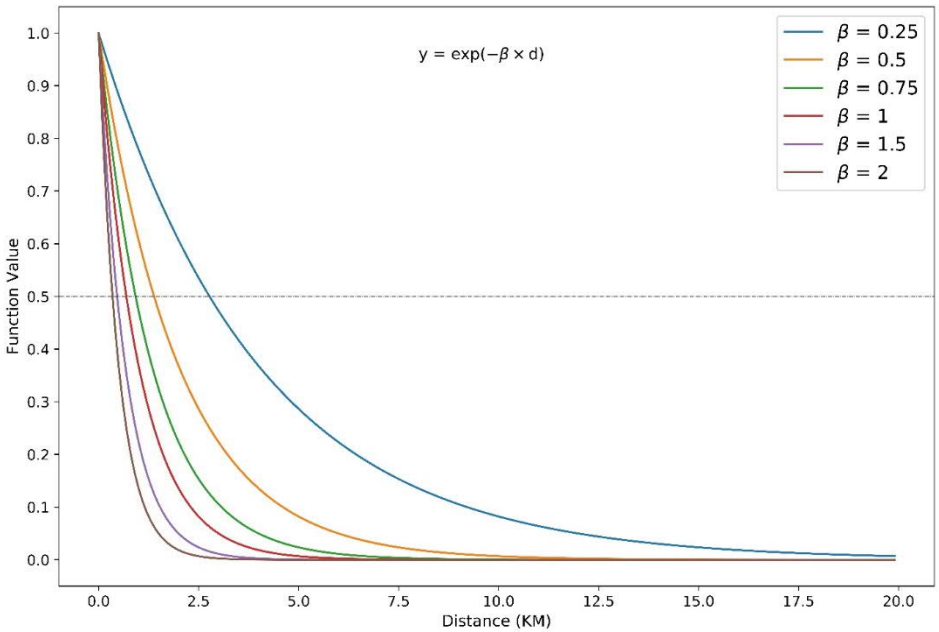
Figure A-3: Steps to create historical rural/urban population grids in 2000 and 2010



3. The Beta effect on suitability

Figure A-4 shows the effect of high beta values on estimating the suitability value of a focal cell. It demonstrates that as the value of β increases, the influence of surrounding cells is more limited to those that are close-by. When β equals 2, contiguous cells represent almost the whole distance effect, overshadowing others.

Figure A-4: The distance decay effect with different β values



4. Spatial autocorrelation analysis

A highly significant global Moran's I for each of the rural α, rural β, urban α, and urban β parameters indicates that the process that has generated its spatial distribution is not random. The LISA measure, on the other hand, focuses on each state, assessing if statistically significant hot spots (states with higher than average values surrounded by neighbors with also higher than average values) or cold spots (states with lower than average values surrounded by neighbors also with lower than average values) exist.

According to Figure A-5, none of the global Moran's I measures is statistically significant at the 0.05 significance level.

Figure A-5: Moran scatterplots for rural/urban α and β parameters

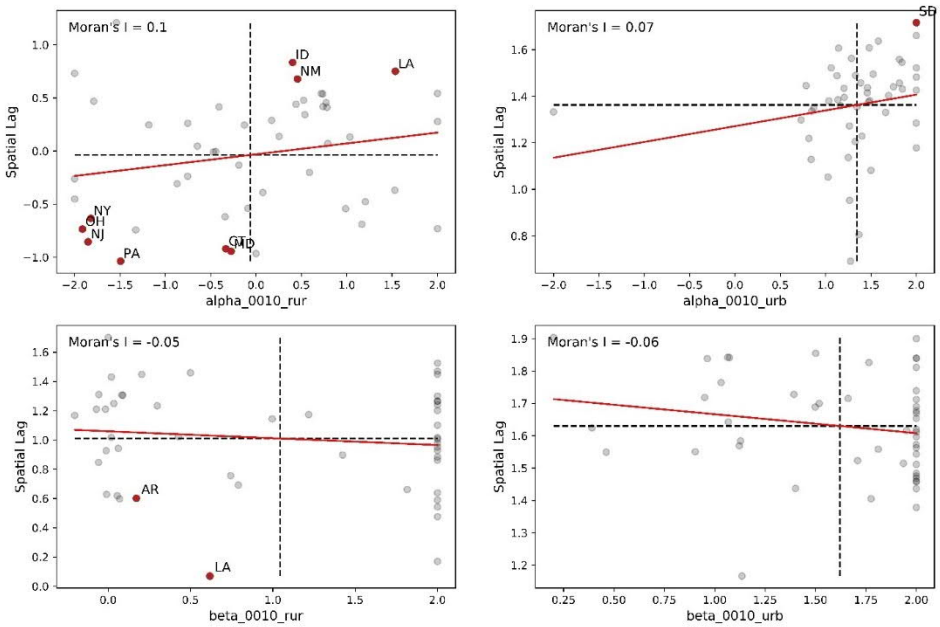
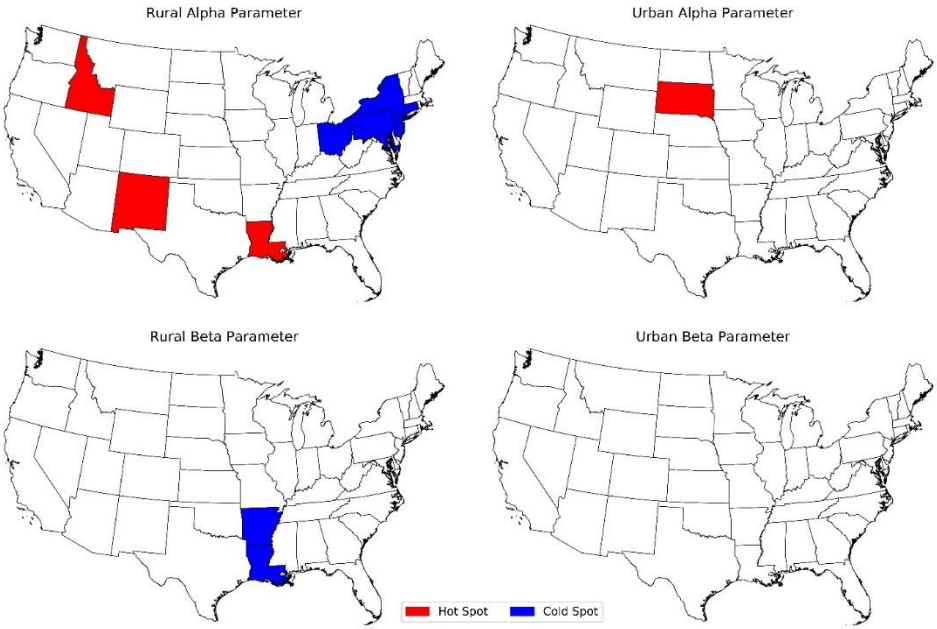


Figure A-6 demonstrates that a few statistically significant local hot spots and cold spots exist for all the parameters except the urban β. Hot spots and cold spots for the rural parameters in Figure 6 only indicate clusters of high and low values, respectively, and their interpretation could be different based on the sign of the rural population change. For example, both Idaho and Louisiana belong to clusters of high rural α, but the interpretation of this parameter is different for Idaho, where the rural population is growing, than for Louisiana, where the rural population is declining. This is also the case within the large northeastern cold-spot cluster for the same parameter in that the interpretation of low rural α is different for a state such as Pennsylvania (with rural population decline) than for Connecticut (with rural population gain). On the other hand, when the sign of the population change is consistent across a region, a single narrative can be assigned. For example, South Dakota and its neighbors in the urban α map represent a region with high dominance of current population agglomerations to attract the growing urban population.

Figure A-6: Hot-spot and cold-spot maps for rural/urban α and β parameters



5. State-level urban and rural parameters

Table A-2: Estimated rural and urban alpha and beta parameters for all states and the District of Columbia

State	Alpha (Rural)	Beta (Rural)	Alpha (Urban)	Beta (Urban)
Alabama	1.03	0.06	1.47	2.00
Alaska	0.60	1.03	1.36	2.00
Arizona	0.25	2.00	0.84	2.00
Arkansas	0.52	0.17	1.69	1.39
California	–1.79	2.00	0.81	2.00
Colorado	0.77	0.02	1.13	1.66
Connecticut	–0.33	0.50	1.20	2.00
Delaware	–1.33	2.00	0.73	2.00
DC	-	-	2.00	1.50
Florida	–1.54	2.00	0.78	1.95
Georgia	1.53	–0.07	1.17	1.40
Hawaii	0.18	2.00	1.00	2.00
Idaho	0.40	1.82	1.48	1.52

Table A-2: (Continued)

State	Alpha (Rural)	Beta (Rural)	Alpha (Urban)	Beta (Urban)
Illinois	–1.18	2.00	1.03	2.00
Indiana	–0.19	2.00	1.37	1.76
Iowa	0.59	0.03	1.81	1.50
Kansas	–0.41	2.00	1.52	1.81
Kentucky	1.20	–0.06	1.46	2.00
Louisiana	1.54	0.62	1.14	2.00
Maine	1.17	0.09	2.00	2.00
Maryland	–0.28	2.00	1.04	2.00
Massachusetts	–2.00	2.00	1.06	1.07
Michigan	–2.00	2.00	–2.00	2.00
Minnesota	–0.87	0.02	1.25	1.94
Mississippi	0.74	0.07	2.00	1.06
Missouri	0.17	0.74	1.28	2.00
Montana	0.78	0.30	1.58	2.00
Nebraska	–0.65	2.00	1.84	0.95
Nevada	–0.75	2.00	1.50	0.20
New Hampshire	0.99	–0.20	1.48	0.90
New Jersey	–1.85	2.00	0.85	2.00
New Mexico	0.46	0.79	1.33	2.00
New York	–1.82	2.00	1.39	2.00
North Carolina	–0.13	2.00	1.75	0.39
North Dakota	–0.44	2.00	2.00	1.03
Ohio	–1.91	2.00	1.26	1.07
Oklahoma	0.72	0.06	1.66	2.00
Oregon	0.44	0.43	1.40	2.00
Pennsylvania	–1.49	2.00	1.26	2.00
Rhode Island	–0.34	1.00	2.00	0.46
South Carolina	–2.00	2.00	1.46	1.13
South Dakota	–0.47	2.00	2.00	1.12
Tennessee	0.79	–0.01	1.33	1.71
Texas	0.74	–0.01	1.21	0.96
Utah	2.00	0.08	0.87	2.00
Vermont	0.08	1.42	2.00	1.78
Virginia	–0.09	0.20	1.35	2.00
Washington	2.00	–0.06	1.13	2.00
West Virginia	–0.76	2.00	1.85	1.13
Wisconsin	2.00	–0.02	1.27	2.00
Wyoming	0.54	1.22	1.81	2.00

6. Absolute values of percentage differences

Table A-3: Absolute values of percentage difference at 50% and 90% of the urban, rural, and total population

State	Urban Error at 50%	Urban Error at 90%	Rural Error at 50%	Rural Error at 90%	Total Error at 50%	Total Error at 90%
Alabama	13.60	45.06	8.87	28.92	9.91	29.44
Alaska	5.20	35.79	16.56	56.67	7.11	32.69
Arizona	21.35	75.42	22.01	70.45	20.82	71.82
Arkansas	13.90	35.38	10.02	25.83	10.54	30.03
California	7.66	21.27	13.67	48.99	7.66	19.29
Colorado	17.47	34.82	11.70	40.11	15.84	32.07
Connecticut	3.25	8.14	8.93	25.06	2.76	6.00
Delaware	11.32	42.29	18.18	39.86	11.35	23.30
DC	6.66	16.64	-	-	6.66	16.64
Florida	14.18	35.90	18.40	52.50	13.91	33.70
Georgia	17.78	44.75	12.54	34.52	14.53	33.49
Hawaii	11.40	19.37	31.70	54.52	12.62	18.11
Idaho	12.72	27.94	9.80	31.93	10.85	28.79
Illinois	7.42	23.08	6.83	25.65	7.17	21.09
Indiana	11.48	26.88	6.51	20.92	8.34	22.83
Iowa	8.13	22.49	6.04	22.09	6.33	17.41
Kansas	9.72	32.12	8.67	22.85	9.07	25.92
Kentucky	8.10	34.34	7.36	21.47	6.49	22.85
Louisiana	10.07	36.56	11.91	35.82	9.02	30.63
Maine	4.56	18.11	5.28	19.44	4.78	11.72
Maryland	7.10	17.12	10.19	30.19	6.20	17.13
Massachusetts	2.93	9.63	12.03	32.58	2.27	7.67
Michigan	6.70	27.73	6.66	23.39	6.18	22.23
Minnesota	10.57	24.64	8.15	25.14	9.86	22.23
Mississippi	17.24	44.37	9.72	32.77	11.84	29.87
Missouri	9.23	28.19	8.09	24.17	8.59	22.51
Montana	9.64	23.33	10.53	35.45	8.29	25.51
Nebraska	12.21	27.56	8.00	24.88	9.11	22.19
Nevada	23.89	52.85	19.61	58.92	23.89	51.97
New Hampshire	7.38	19.35	8.67	25.57	6.48	13.63
New Jersey	4.01	10.17	16.22	49.25	3.66	8.34
New Mexico	13.16	30.75	12.32	43.72	11.97	29.52
New York	2.71	9.97	5.42	21.67	2.71	9.04
North Carolina	16.54	40.75	10.05	27.95	10.49	31.13
North Dakota	2.89	42.00	10.39	33.53	6.44	38.55
Ohio	9.11	23.48	6.22	20.25	6.64	19.34
Oklahoma	8.72	21.18	8.49	26.45	8.11	20.41
Oregon	5.96	14.14	7.60	27.98	5.75	13.67
Pennsylvania	6.93	19.01	7.83	25.66	6.83	14.69
Rhode Island	2.23	12.08	5.54	24.88	2.26	12.33

Table A-3: (Continued)

State	Urban Error at 50%	Urban Error at 90%	Rural Error at 50%	Rural Error at 90%	Total Error at 50%	Total Error at 90%
South Carolina	18.18	45.38	9.70	29.44	12.44	35.48
South Dakota	13.23	27.32	9.41	33.08	11.89	23.09
Tennessee	10.91	34.58	7.95	23.59	7.80	28.55
Texas	16.59	37.41	11.86	37.94	15.83	33.73
Utah	14.55	39.32	13.76	43.31	14.58	31.45
Vermont	3.90	32.83	4.46	19.97	3.53	13.19
Virginia	11.53	25.23	9.66	29.94	11.04	21.33
Washington	7.87	18.73	12.60	36.13	7.41	17.20
West Virginia	11.22	34.30	8.39	25.21	7.33	21.67
Wisconsin	7.34	21.22	6.68	21.83	5.94	16.04
Wyoming	5.42	17.50	12.13	44.57	6.67	21.16

7. US Census inconsistency in defining 'urban'

The US Census has not employed a single set of criteria to classify blocks as either rural or urban. As Table A-4 indicates, the census criteria grow to be more urban-inclusive over time (Balk et al. 2018). Therefore, the reason why some states lose rural population rapidly might be partially due to the census reclassifying some blocks that were rural in 2000 to urban in 2010. One way to disentangle the actual rural population decline from the reclassification-induced part is to establish a consistent set of criteria to define rural and urban, which is not the focus of this current work.

Table A-4: Summary of the US Census definition of 'urban' over time

Spatial Product	Year	Urban Proxy Definition	Population	Spatial Resolution
U.S. Census Blocks	1990	Population density and count dependent.	Variable	Variable based on population
	2000	Population count, density, and proximity dependent.		
	2010	Population count, density, proximity, and urban land-use dependent.		

8. Maps of total population and mean population difference

Figure A-7 shows observed (block-based) and estimated total population grids for New York (a state with relatively low error) and North Carolina (a state with relatively high error) in 2010. It also shows mean population differences (estimated – observed) for these states.

Figure A-7: (a) Block-based total population distribution, (b) estimated total population distribution, and (c) mean population difference for New York in 2010, and (d) Block-based total population distribution, (e) estimated total population distribution, and (f) mean population difference for North Carolina in 2010