LETTER

Empirically based spatial projections of US population age structure consistent with the shared socioeconomic pathways

Erich Striessnig1, Jing Gao2, Brian C O’Neill3 and Leiwen Jiang4,5

1 Wittgenstein Centre for Demography and Global Human Capital (IIASA, VID/OAW, WU), Austria
2 University of Delaware, United States of America
3 University of Denver, United States of America
4 Population Council, United States of America
5 Asian Demographic Research Institute (ADRI), People’s Republic of China

E-mail: striess@iiasa.ac.at

Keywords: spatial population, age structure, population projections, shared socioeconomic pathways

Supplementary material for this article is available online

Abstract

Spatially-explicit population projections by age are increasingly needed for understanding bilateral human–environment interactions. Conventional demographic methods for projecting age structure experience substantial challenges at small spatial scales. In search of a potentially better-performing alternative, we develop an empirically based spatial model of population age structure and test its application in projecting US population age structure over the 21st century under various socioeconomic scenarios (SSPs). The model draws on 40 years of historical data explaining changes in spatial age distribution at the county level. It demonstrates that a very good model fit is achievable even with parsimonious data input, and distinguishes itself from existing methods as a promising approach to spatial age structure modeling at the global level where data availability is often limited. Results suggest that wide variations in the spatial pattern of county-level age structure are plausible, with the possibility of substantial aging clustered in particular parts of the country. Aging is experienced most prominently in thinly populated counties in the Midwest and the Rocky Mountains, while cities and surrounding counties, particularly in California, as well as the southern parts of New England and the Mid-Atlantic region, maintain a younger population age structure with a lower proportion in the most vulnerable 70+ age group. The urban concentration of younger people, as well as the absolute number of vulnerable elderly people can vary strongly by SSP.

Introduction

The spatial pattern of a population’s distribution is a key driver of its vulnerability to many social and environmental stresses. Existing literature has found its influence to be larger than or at least equal to the influence of physical environmental factors, such as heat waves and sea-level rise (Hallegatte et al 2013, Jones et al 2015, Neumann et al 2015, Hauer et al 2016, Lehner et al 2018). Age is another important demographic factor determining people’s vulnerability to all sorts of hazards, however, only few studies have focused on it so far, in particular in conjunction with its distribution across space (Marsha et al 2018). Large-scale (continental to global) long-term (e.g. over the entire 21st century) spatial projections of age-structured populations are scarce. Most such projections are local in scale and short-term in range (Swanson et al 2010, Salvo et al 2013). Many future impact assessments, therefore, take place at the national level (Dong et al 2015).

A key obstacle to the production of large-scale, long-term spatial projections of age structure is the difficulty of applying conventional demographic approaches to such spatial and temporal scales. The dominant method for population projections is the cohort-component approach (Burch 2018), deriving directly from the demographic equation balancing births, deaths, migration, and population growth, with each component differentiated by age. Popularized in the 1920s
and 1930s (Whelpton 1936), the most widely used national-level population projections today all rely on this method (Lutz et al 2019, United Nations 2019). Its application to smaller spatial scales is challenging, because it requires both sufficient data on and future projections of age-specific demographic rates for each spatial unit. In data-poor environments, such information on current conditions is typically not available, and the sheer number of assumptions that must be made for future projections for small regions is an obstacle, especially for migration which can vary greatly over space and time.

Nonetheless, county-level population projections that include age structure have been carried out for the US based on the cohort-component method (Bierwagen et al 2010, US Environmental Protection Agency 2017). However, some input assumptions (fertility and mortality) were not varied at the county level, therefore leaving out some degrees of heterogeneity. To reduce data requirements, Hauer (2019) projected US county-level population and age structure by extrapolating rates of change in each age group, and then scaling the size of all population sub-groups up or down to match a target aggregate national projection. This method does not allow for spatial variation in outcomes between national scenarios, making it perhaps better suited to shorter time horizons. A simpler scaling method has been used for Europe, where Terama et al (2017) downscaled national-level age structure projections from a cohort-component model to sub-national regions, proportionally to their age structures in the base year, a method that does not allow for changes over time in the relative age distribution across regions.

Seeking for a potentially better-performing alternative, we explored empirically based model development methods. We conducted thorough and systematic exploratory data analysis of best available datasets with global coverage. Summary results for US counties presented in SI section 1 is available online at stacks.iop.org/ERL/14/114038/mmedia revealed the following spatial and temporal patterns.

1. In any historical census, US counties can be distinguished by roughly four distinctive age profiles. Similar age profiles appear spatially clustered during earlier census years (1940–1960) but evolve to show dispersed spatial patterns in recent censuses (1970–2010).

2. Over time, most counties switch among different age profiles for various reasons, e.g. aging of the baby boomers or internal migration. Meanwhile, counties with extreme age profiles (e.g. those that are home to large college-aged populations or retirement communities) tend to be more temporally static.

3. The relationships between changes in county-level age structure and its potential predictors are complex and nonlinear.

We developed a new model using regression trees to project changes in county age structure based on current and past county demographic characteristics, reflecting these empirical patterns. Regression trees can model multiple nonlinear relationships simultaneously. During training, the model empirically identifies the best way to model counties with different age profiles and other demographic characteristics as separate cases using different ‘tree branches’. When making projections, at each future decadal time step the model projects county age structure change using whichever branch of the regression tree corresponds to its demographic characteristics at that point in time. This is analogous to the observed pattern that counties switch among different age profiles over time. In addition, because regression trees treat different parts of the data range separately, they are robust to extreme values, which may occur more frequently under certain future scenarios. Altogether, the characteristics inherent to regression trees make them a particularly suitable method for modeling the highly temporally variant phenomenon of age structure change.

We used this model for adding five age groups to existing projections of the spatial distribution of population that contain age structure information only at the national aggregate level. The five age groups satisfy most needs for age structure information by providing outcomes for age groups such as the elderly, children, and the working age population (see Methodology). The existing national and spatial population projections were developed for the shared socioeconomic pathways (SSPs), a set of five alternative scenarios of future societal development. The SSPs have been widely used in climate and other global change research (Striessnig and Loichinger 2016, Jiang and O’Neill 2017, O’Neill et al 2017, Riahi et al 2017). They include national-level projections of population by age and sex (KC and Lutz 2017), and gridded spatial projections of population counts (without age detail) at 1/8-degree and 1 km resolutions (Jones and O’Neill 2016, Gao 2017). Differences between SSPs are described in detail in the Methodology, as well as SI section 2.

In spite of its parsimonious requirements for input variables, the model achieved very good fit when validated over historical data (see SI section 3), a prime prerequisite for its application to future projections in data-poor contexts where the cohort-component approach is not an option. The model’s primary covariates capture the counties’ demographic histories inscribed in their own age structures, with some metrics accounting for potential spatial autocorrelation among neighboring counties. The model’s estimates are stable over long time horizons even without constraints at the national level (SI section 4). Moreover, unrealistically high change rates (implying unreasonable birth, death,
or migration rates) cannot occur using the methodology presented here, as the range of possible outcomes of our regression trees are by definition of the range observed in the past.

Hence, our empirically based approach is a promising candidate for a well-performing global model under data-sparse conditions. Though the work presented here is for the US only, we are currently exploring the usability of our approach for other countries, and plan to integrate the lessons learnt to develop a global model eventually. The US is a good first test case because here different SSPs lead to very different future trajectories of population growth, population aging, and spatial distributions of population (SI section 2).

Methodology

Exploratory analysis

Before setting up our model, we used a range of exploratory data analysis and visualization techniques to understand the patterns in the US decadal census data since 1940, which is the first US census available with sufficient age and spatial detail (Manson et al. 2018). Most notably, we applied hierarchical cluster analysis to county-level age profiles from every census, respectively. Our findings suggested that while in the 1940s and 50s, age structure across the US was characterized by marked spatial autocorrelation at the county level, the pattern observed since the 1970s followed a different regime in which spatial autocorrelation of age structure is far less pronounced (SI section 1). This is consistent with existing literature confirming the profound sociodemographic changes that the US underwent up until the 1970s that substantially altered the spatial patterns of age structure (Jackson 1987): with the onset of the baby boom, average family sizes increased markedly, leading to an increased demand for larger homes outside America’s urban cores. The mobility revolution facilitated longer commutes and the move to the suburbs, while the gradual westward shift of the population was still ongoing.

Considering these patterns, we trained our empirical model on the census data from the more recent decades only (1980–2010) during which the observed spatial patterns in age structure are consistent and to keep the 1970 census wave for validation purposes. For future projections, it is unlikely that the data generating process will return to the pre-1970s regime, given that it has already been stable for more than four decades. Note, however, that although spatial autocorrelation of age structure has declined, variation in age structure across counties remains high (see figure S1), driven mainly by persistent urban-rural differences (for details see figure S3 of the SI appendix).

Model design

To avoid small numbers in thinly populated rural counties and to obtain more robust modeling results over long time horizons, we use five aggregate age groups (0–19, 20–29, 30–54, 55–69, 70+). This choice of age ranges reflects earlier findings from studies of age-specific mobility patterns observed in the US in the second half of the 20th century (Johnson et al. 2005, Johnson and Winkler 2015) while trying to account for differences in household consumption behavior over the lifecycle (Ando and Modigliani 1963, Lee and Mason 2011, Mason and Lee 2013), as well as age-specific vulnerability patterns (Jonkman et al. 2009, Zagheni et al. 2016). Modeling each age group separately helps keeping the structure of our trees simple and interpretable. The regression trees are fitted using the rpart package in R Statistical Software (Therneau et al. 2015). The process of pruning the trees, as well as the detailed structures of all five trees are described in SI section 2.

Our dependent variable, is the change in the relative shares (Δrs), i.e. the difference in the population share of each age group in the total county population over a decade (i.e. between two census waves) relative to the same difference observed at the national level:

\[
\Delta r_s = \frac{\text{share}_c - \text{share}_{c-1}}{\text{share}_n - \text{share}_{n-1}}.
\]

Thus, we are modeling county-level deviations in the change in the share over time from the same change at the national level. This approach allows us to automatically account for the national-level change that is occurring in the aggregate scenario for which we are producing a spatial age structure outcome, and to focus the model on the differences across counties relative to the national trend. We tested various alternative specifications of the dependent variable, including the raw age-specific shares of population or differences in those shares compared to the national shares. The best fit to the observed shares was achieved by the change in the relative share variable in equation (1).

In choosing independent variables, care was taken to use variables that would be available in the context of projections (see full list in table S3). Lacking information on the drivers of population change at the county level, temporally lagged shares of all five age groups in the total county population up to three census waves in the past allow us to study the influence of historic demographic shifts on future developments. The use of lagged shares of all age groups to project the future changes in a single age group provides a cohort perspective that mimics the behavior of a cohort-component model. If, for example, a baby boom produces an especially large share of population in the 0–19 age group in one time period, older age groups should expect to grow in future time periods as the baby boomers age. The lagged structure of the full set of age groups allows for this dynamic to be represented.
in the model. To rule out the possibility of linear dependence among our predictors, we conducted a thorough collinearity investigation (see SI section 3.1 for details).

**Training and validation**

The trees use lagged information on age structure from the previous three census waves to project changes over a given decade. We trained the model by closely reproducing observed change over the 2000–2010 decade using patterns and trends observed in the census waves of 1980, 1990 and 2000. While training the model on one decade of change may seem limiting for supporting century-scale projections, additional confidence in the applicability of the model comes from (1) the use of three preceding decades of data in estimating the model, (2) a validation test run on a different decade, and (3) the fact that the 2000–2010 decade is not unusual with respect to trends since at least 1970 (see figure S1).

As shown in section 3.2 of the SI, the goodness of fit for all five regression trees is confirmed by values of $R^2$ close to and above 0.90. To evaluate the model’s generalizability, we applied it to estimate changes over the 1990–2000 period, using data from the census waves of 1970, 1980, and 1990. The performance continues to be high with values of $R^2$ well above 0.80 for all five trees. In addition, no spatial pattern could be identified in the residuals, supporting the quality of the model for short-term modeling.

Longer-term performance of the model would be supported by successfully capturing demographic processes related to fertility, mortality or migration. To examine whether this is the case, we tested the trees’ sensitivity to county-level migration and crude birth rates, even though information on those would not actually be available for use in future projections. Neither variable added to the trees’ explanatory power, as the trees did not pick them and were able to find appropriate surrogate variables, implying that the informational content of these omitted variables is already captured by lagged age structures and neighborhood characteristics, and that our projection model is reasonable without them.

**National population projections**

The main input to our future projections are the SSPs. At the national aggregate level, population aging is the dominant trend in all five of them, but the speed of aging varies between them (figure S5) and so do the drivers of population change. While under SSP5 (‘Fossil-fueled development’) rich OECD countries, including the US, are assumed to experience high fertility and migration combined with low mortality, which limits the rate of aging, the opposite is true under SSP3 (‘Regional rivalry’), leading to faster aging of the US population. Under SSP2 (‘Middle of the road’), all three of these drivers are assumed to be at medium levels, yet the resulting national share of elderly population in 2100 is relatively large, almost the same as in SSP3, albeit for very different reasons. While under SSP3 population aging is driven by low fertility and low migration, under SSP2 it is the higher net in-migration in the first half of the century that leads to increased aging once migration starts to decline in the second half of the century, as is the assumption with all SSPs. Despite all SSPs sharing the same assumptions for the US with regard to urbanization (Jiang and O’Neill 2017), they lead to substantially different spatial population outcomes at the sub-national scale (Jones and O’Neill 2016).

Our statistical model is capable of distinguishing between those different national-level pathways, providing different sub-national distributions of aging depending on which demographic forces drive them.

**Projection**

Given the SSPs as input, we then used the five regression trees to project county-level population shares of the five age groups independently from each other in ten-year steps covering the 2010–2100 period. At each ten-year step, the regression trees determine which branches are used to make projections for any given county. Depending on the evolution of each county’s age related characteristics, different tree branches may be used for the same county at different times. To ensure that (1) the five age-group shares of each county sum to one, and (2) the adjusted outcomes are consistent with the national aggregate age structure given by the SSPs, we renormalized the counties’ age-group shares estimated by the trees using iterative proportional fitting (IPF). Running this renormalization at every time step avoids potentially larger errors caused by the accumulation of small mismatches over time. It is worth noting that the regression trees performed highly, and the sum of their unconstrained estimates within each county is already close to one. As shown in SI section 4, IPF leads only to minor changes to the age structures predicted for individual counties.

The outcomes of our county age structure model depend on the existing national and spatial population models in two main ways. First, national population size acts as a constraint on the spatial population model, and that model produces population size and density by county that enters the county age structure model as independent variables. Second, national age structure is used as a constraint on the county age structure model as part of the IPF. Therefore, national-level population outcomes, such as high growth rates and a young age structure from a high fertility scenario, and spatial population outcomes, such as concentrated growth in urban areas, will both influence the results from the county age structure model. These factors interact with the current and lagged demographic conditions across counties to produce future outcomes. What would not be reflected in the model are changes in spatial patterns.
of demographic change relative to historical patterns. For example, if domestic or international migration patterns strongly shift their spatial pattern, so that new cities and states become preferred destinations (or origins in the case of domestic migration), this will not be reflected in the spatial population model and therefore also not in the county age structure model. If the age structure of rural-urban migration changes substantially relative to historical trends, this will change the nature of rural-urban differences in age structure away from historical patterns and not be reflected in the model.

Results

We applied the regression tree methodology described above to project county-level age structure for the conterminous US consistent with all five of the SSPs. We show results for the Middle of the Road scenario (SSP2) as well as for SSPs 3 and 5, which bound future assumptions on the drivers of population change in the US; results for the remaining SSPs are reported in the SI appendix section 5.

Figure 1 shows observed (1940–2010) and projected (2020–2100) proportions of total county population in five different age groups. In accordance with the general trend of population aging predicted for the US under all five SSPs, the proportions of children under the age of 20, as well as the population of working age (30–54), decrease over time. In contrast, there is a steady increase in the population share above age 70, an age group particularly vulnerable to the consequences of climate change, such as increased frequency and intensity of heat waves. While at the national level this proportion is predicted to increase from around 9% in 2010 to between 26% (SSP5) and 34% (SSP1) in 2100, results show that in 25% of counties, that proportion can rise to between 33% and 40% by the end of the century. In places like Charlotte County, FL, where the proportion 70+ is already high in 2010 (24%), it is projected to reach between 40% and 50% in 2100. This would be twice as high as observed in any single county in 2010 (27.6% in Sumter County, FL). The spatial distribution of these subnational differences in aging are explored below.

We also find a large increase in the number of counties with very little population under the age of 20. While currently there are no counties with less than 5% of their population under age 20, in SSP2 and SSP5 this figure grows to 10% of all counties, and in SSP3 it reaches 35% of all counties by the end of the century. This result is driven by (1) fertility decline at the national level, most strongly in the population decline future described by SSP3 where the total fertility rate (TFR) for the US in 2095–2100 is assumed to reach 1.47 (compared to 1.87 today; CIA 2017), corresponding to a decline in the proportion aged 0–19 at the national level from 27% in 2010 to 15.3% in 2100; (2) negative population momentum, as the proportion of potential parents is also going down.
steadily; (3) the increase in the proportion of elderly population above the age of 70; and (4) a spatial model that keeps the variance in proportions of total national population across counties approximately constant over time (for details see Jones and O’Neill 2016). Only under SSP5 is national-level TFR expected to rise to above replacement level (2.29) by the end of the century, limiting the number of counties with very small proportions of children.

The substantial aging of the population even in the Middle of the Road scenario is reflected spatially throughout the entire conterminous US (figure 2). Yet counties with large cities, as well as their high-density neighboring counties, which attract large amounts of working age population and their children, maintain a relatively high proportion of youth and conversely, a lower proportion of elderly people. This is most visible in California, the East Coast and in the Chicago area. The thinly populated counties in the Midwest and the Rocky Mountains experience aging most drastically.

Figure 2. Observed (2010) and projected (2050, 2100) proportion 0–19 (left) and 70+ (right) under SSP2.

Although similar at the national level, patterns of aging by the end of the century differ substantially between SSP2 and SSP3 at the sub-national level. These two scenarios have similar shares of national population in the 70+ age group late in the century, but most counties have a larger fraction of population in this age group in SSP3 than in SSP2 (figure 3), while urban counties, where most of the population resides, are affected by aging to a lesser extent in SSP3. The more intense urban concentration of younger people in SSP3 is due to the differences in the drivers of population change at the national level, as well as their manifestation at the county level. As the total US population starts to shrink in the second half of the century under SSP3, a large proportion of counties are left with almost no population under the age of 20 due to sustained low fertility. Yet the empirical model allocates more youth to places with both high population density and growth that are surrounded by counties with similar characteristics. Positive neighborhood
effects, resulting from the gravitational force of large agglomerations, diminish with distance from the urban cores and lead to the pattern that urban areas maintain higher fractions of children than rural areas despite population decline at the national level.

A similar effect is at play in SSP5, which also produces a more intense concentration of the younger population in urban areas than in SSP2 (figure 3). In this case, however, it is the more rapid population growth (rather than a shrinking population) that favors concentration of younger age groups in and around cities. In addition, the higher population density plays a role in creating more extensive urban and suburban areas of relatively young populations. Aggregated results for all five SSPs in rural and urban counties across four macro-regions of the conterminous US can be found in figure S15 of the SI appendix. They support the finding of more rapid aging in rural America described here, particularly in the Western parts of the country.

While results in terms of proportions of population by age give a clearer indication of age structure changes separate from overall population growth, total numbers of people in each age group are also an important indicator of vulnerable populations. For example, the spatial distribution of the total numbers of people above the age of 70 by county (see figure S16) shows that SSP5, which has a younger population overall than SSP2, actually has a larger number of people in the 70+ category in most counties due to the strong population growth in that scenario. Conversely, SSP3 has an older age structure but fewer people in the oldest age group due to slower overall population growth. Across SSPs, the spatial distribution of the number of people above age 70 does not shift in dramatic ways, in contrast to the substantial differences in the spatial distribution of population shares above age 70.

**Discussion**

Our projections provide plausible county-level age distributions that are consistent with existing national and county-level population projections based on each of the SSPs.

Spatial projections of age-structured populations are critical to understanding risks posed by hazards from climate and other environmental changes, as well as to consumption behavior that may drive those changes. Our model provides a new category of
demographic information to existing projections of US population distribution that makes those projections much more relevant to risk analysis. We find the possibility of wide variation in age structure outcomes at the county level, as well as substantial aging clustered in particular parts of the country: across all SSPs, cities and surrounding counties maintain a younger population age structure with a lower proportion in the most vulnerable 70+ age group. The largest of these clusters can be found in California, as well in the southern parts of New England and the Mid-Atlantic region, ranging roughly from Albany to Richmond. Remote rural counties, on the other hand, tend to age more rapidly. According to the US National Climate Assessment’s recent Climate Science Special Report (2017), climate change impacts will also vary regionally. Northern regions of the US are expected to see the largest increases in the intensity of heat waves, while chronic, long-duration drought will become increasingly likely in the Southwest. Combining societal projections with projections of environmental hazards such as heat waves, droughts, or floods can yield improved estimates of potential impacts on the most vulnerable segments of society with the potential for improving the rigor of intervention efforts and lowering their cost.

The model also represents a novel method in demography for projecting future age structure that is especially well suited to regions with data limitations. Although it differs from traditional cohort-component methods that treat components of population change directly, it takes these into consideration indirectly through proxy variables whose ability to implicitly capture the effects of demographic rates has been successfully tested. We anticipate that the model will be useful for projecting spatial age structure in places where the data requirements of conventional models cannot easily be satisfied.

Data limitations for supporting empirically based models used to make long-term (almost century scale) projections pose a common challenge across many fields, and the work presented here is no exception. The use of changes observed over one decade to project changes over a century can be problematic, as the changes in age structure shares observed over that decade can be anomalous. However, our analysis draws on 40 years of data (1970s through 2000s), substantially more than many projection models in common use in the global change field, and our exploratory data analysis (see SI section 1 for details) confirms that this period is in fact characterized by a data generating process that is both stable and markedly different from the one observed in the earlier (1940–1960) period.

Future work is planned to improve the model in the following aspects: (1) Investigate possibilities to generate future scenarios beyond historically observed patterns, e.g. using empirical models trained for different countries as representations of different sociodemographic trends. (2) Incorporate a larger number of age groups. This requires dealing with the trade-off between model performance and the number of age groups tracked simultaneously. (3) Improve validation to better assess the model’s performance in long-term projections. (4) Allow city size and distance to city to vary dynamically across time and SSP. (5) Integrate further important dimensions of population heterogeneity besides age, e.g. gender, race or educational attainment.

Data availability statement

The data that support the findings of this study are openly available at https://doi.org/10.22022/pop/10-2019.54.

Code availability statement

Codes and materials used in this work are available upon requests.

Acknowledgments

IIASA has made funds available for publishing of this paper. IIASA encourages and actively supports its researchers to publish their research in journal articles or books that are made available for free to all users (gold open access).

ORCID iDs

Erich Striessnig @ https://orcid.org/0000-0001-5419-9498
Jing Gao @ https://orcid.org/0000-0003-1778-8909
Brian C O’Neill @ https://orcid.org/0000-0001-7505-8897
Leiwen Jiang @ https://orcid.org/0000-0002-2073-6440

References

Ando A and Modigliani F 1963 The ‘Life Cycle’ hypothesis of saving: aggregate implications and tests Am. Econ. Rev. 53 55–84
Bierwagen B G, David M T, Pyke C R, Chooate A, Groth P, Thomas J V and Morefield P 2010 National housing and impervious surface scenarios for integrated climate impact assessments Proc. Natl. Acad. Sci. 107 20887–92
Burch T K 2018 The cohort-component population projection: a strange attractor for demographers Model-Based Demography: Essays on Integrating Data, Technique and Theory (Demographic Research Monographs) ed K Burch Thomas (Cham: Springer International Publishing) pp 135–51
CIA 2017 The World Factbook, 2017 (https://cia.gov/library/publications/resources/the-world-factbook/index.html)
Dong W, Liu Z, Liao H, Tang Q and Li X en 2015 New climate and socio-economic scenarios for assessing global human health challenges due to heat risk Clim. Change 130 505–18
Gao J 2017 Downscaling Global Spatial Population Projections from 1/8-degree to 1 km Grid Cells NCAR Library (https://doi.org/10.5065/D60Z721H)
