U.S. County level probabilistic population projections by age, sex, and race for the period 2012-2100 consisent with the Shared Socioeconomic Pathways *

Mathew E. Hauer [1*] *Florida State University*

Small area and subnational population projections are important for understanding long-term demographic changes and typically take the form of a cohort-component model. Cohort-component relies on oftentimes difficult or even impossible to obtain subnational components of change due to data suppression for privacy reasons, small-cell sizes, or are simply unavailable. Cohort-Change Ratios (CCRs) are one approach that overcomes these data limitations but tend to produce unrealistic projected populations due to exponential compounding. I present a simple, parsimonious projection technique based on a variation of CCRs I call cohort-change differences (CCDs). Using ex-post facto analysis for the period 2000-2015 for 3,136 U.S. counties in temporally rectified county boundaries, eighteen five-year age groups (0-85+), two sex groups (Male and Female), and three race-groups (White, Black, Other) using CCDs in a Bayesian structural time series for the period 1969-2000, I show that CCDs produce reduced errors compared to CCRs. I then provide county-level population projections by age, sex, and race in five-year intervals for the period 2020-2100, using Bayesian structural time series, consistent with the Shared Socioeconomic Pathways. These data and methods have numerous potential uses and can serve as inputs for addressing questions involving sub-national demographic change in the United States.

Keywords: Population projections; subnational; demographic change; cohort-change ratios

* Corresponding author. hauer@uga.edu. p: 706-542-9369.

[1] Department of Sociology, Florida State University.

[*] The data and code that supports this analysis are available in the supplementary materials.

BACKGROUND & SUMMARY

Population projections have a long history in the social and physical sciences as a means of examining demographic change, planning for the future, and to inform decision making in a variety of applications (Smith, Tayman and Swanson, 2006; Passel and Cohn, 2008; Hebert et al., 2003; Hales et al., 2002; Hauer, Evans and Mishra, 2016; Gerland et al., 2014; Colby and Ortman, 2017). Small-area population projections tend to be less robust than projections produced for larger geographies (Swanson, Schlottmann and Schmidt, 2010; Smith, Tayman and Swanson, 2006; Alexander, Zagheni and Barbieri, 2017) yet there is an increasing demand for small-area demographic analysis (Alexander, Zagheni and Barbieri, 2017; Chi, 2009; Smith, Tayman and Swanson, 2013; Raymer, Abel and Rogers, 2012; Tatem et al., 2012).

Cohort-component, the typical demographic projection methodology employed across all scales, requires data on each component process disaggregated by the dimensionality of the population to be projected and oftentimes elements of these data are difficult to obtain at subnational geographies. For example, in the united States, there is no comprehensive data set of gross migration estimates by age, sex, and race for all U.S. counties, precluding the use of a gross-migration model of a cohort-component. Additionally, the National Center of Health Statistics (NCHS) suppress birth data in counties with fewer than 100k persons and data data are only available for cells with more than 10 deaths (Tiwari, Beyer and Rushton, 2014). Even when these data are available, at the county-level, vital event data are not released as the sub-county level making cohort-component extremely difficult, if not impossible, to implement.

The Hamilton-Perry method (Hamilton and Perry, 1962; Swanson, Schlottmann and Schmidt, 2010) is a simple, parsimonious technique for producing population projections directly from multiple age-sex distributions through the use of cohort-change ratios (CCRs) (Baker et al., 2017). The minimal data requirements to produce CCRs, the ability to implement the same method across scale, the ability to implement CCRs in Leslie matrices (Sprague, 2012), and the general ease of implementation make CCRs attractive in the production of small-area demographic projections. However, CCRs suffer from two major disadvantages over the use of cohort-component: 1) short-term rapid population growth can create impossibly explosive growth in long-range projections. Consider McKenzie county North Dakota. In 2010 McKenzie had a population of 6,360 that had ballooned to 12,792 by 2015, with a CCR for the 20-24 year old population of 2.46 (416 to 1,027). Implementing a 50-year population projection using that CCR would create a projected population that is approximately 8,000 times larger (2.46^{10}) – clearly an improbable number given the small, rural nature of its population – yielding a theoretical 20-24 year old population of approximately 8 million. 2) Small cell sizes can create impossibly large CCRs with very small numeric change. A population changing from 2 to 6 persons has a net-change of just 4 persons but a CCR of 3.0.

In this paper I demonstrate an alternative to CCRs, which I call cohort-change differences (CCDs), that create linear rather than exponential growth. This technique has all of the advantages of CCRs by remaining just as simple and parsimonious with minimal data requirements, ease of implementation, leveraging the relationship between CCRs/CCDs and the fundamental demographic equation, while producing projected populations without impossibly explosive growth. I begin by outlining the methodology and describe how such a method can be easily implemented within a stochastic framework. To evaluate the technique, I use ex-post facto evaluations of a population projection by age/sex/race in five-year intervals for the period 2000-2015 for 3,136 temporally rectified county-boundaries, with eighteen five-year age groups (0-85+), two sex groups (Male and Female), and three race groups (White, Black, and Other). Overall, the errors reported here are on par with or better than deterministic cohort-component population projection models undertaken at the sub-national level and with Bayesian population projection models undertaken at various national and sub-national scales (Smith and Tayman, 2003; Wilson, 2016; Smith, 1997; Rayer, 2008; Wilson and Rees, 2005; Booth, 2006; Wilson, 2012; Raftery et al., 2012; Boyle et al., 2010; Daponte, Kadane and Wolfson, 1997; Lutz, Sanderson and Scherbov, 1996).

Producing high-quality, highly-detailed population projections is a challenging endeavor and no rigorous set of U.S. sub-national projections by age, sex, and race currently exists. With such a large need for sub-national projections and to better understand the changing demographics of the U.S. population, I produce such a set of high-quality, highly-detailed projections and make both the *R* code and subsequent projections available for dissemination to a wide audience. Here, I present age-sex-race specific population projections for all U.S. counties controlled to the Shared Socioeconomic Pathways (O'Neill et al., 2014). I generate these projections using a historic time series of population estimates for the period 1990-2016 (Wonder, 2016) in temporally rectified county boundaries and race groupings using Leslie matrices populated by CCRs and CCDs (Baker et al., 2017), projected through the use of a Bayesian structural time series model (Harvey, 1990).

These projections, like all projections, involve the use of assumptions about future events that may or may not occur. Users of these projections should be aware that although the projections have been prepared with the use of standard methodologies, documentation of their creation, open-source computer code, and extensive evaluations of their accuracy and uncertainty, they may not accurately project the future population of a state, county, age, sex, or race group. The projections are based on historical trends and current estimates. These projections should be used only with full awareness of the inherent limitations of population projections in general and with knowledge of the procedures and assumptions described in this document.

METHODS

The cohort-component method is the most accepted methodology to produce population projections (Smith, Tayman and Swanson, 2006; Preston, Heuveline and Guillot, 2000). The method makes use of all three population component processes (fertility, mortality, and migration) and applies them across varying population cohorts to arrive at a future population. Equation 1 outlines the basic structure of a cohort-component model.

$$P_{t+1} = P_t + B_t - D_t + M_{t,in} - M_{t,out} \tag{1}$$

Where P_t is the population at time t, B_t is the births at time t, D_t is the deaths at time t, and $M_{t,in/out}$ refers to in- or out-migration at time t.

Cohort-component requires data on each component process disaggregated by the dimensionality of the population to be projected. To produce detailed projections by age, sex, and race, detailed data age, sex, and race. Certain elements of these data can be difficult to obtain for complete national coverage of sub-national geographies. There is no comprehensive data set of gross migration estimates by age, sex, and race for all U.S. counties. Birth and death data are typically obtained through the National Center of Health Statistics (NCHS) vital events registration databases (Martin et al., 2018). Birth data, however, are only available for counties with populations greater than 100k and Death data are only available for cells with more than 10 deaths (Tiwari, Beyer and

Rushton, 2014). These limitations surrounding fertility, mortality, and migration render a universal county-level population projection difficult, if not impossible, to complete using publicly available data sets.

An alternative to cohort-component is the Hamilton-Perry method (Swanson, Schlottmann and Schmidt, 2010; Baker et al., 2017,), which uses cohort-change ratios (CCRs) in place of components to project populations. The basic CCR equation is found in equation 2.

$$
\begin{aligned}
CCR_t &= \frac{{}_nP_{x,t}}{{}_nP_{x-y,t-1}} &(2)\\
{}_nP_{x+t} &= CCR_t \cdot {}_nP_{x-y,t} &(3)
\end{aligned}
$$

Where ${}_nP_{x,t}$ is the population aged x to $x+n$ in time t and ${}_nP_{x-y,t}$ is the population aged x to $x+n-y$ in time t where y refers to the time difference between time periods. These CCRs are calculated for each age group a, for each sex group s, for each race group r, in each time period t, in county c. Thus to find the population of ten to fourteen year olds (${}_5P_{10}$) in five years ($t+1$), we multiply the ratio of the population aged 10-14 in time t (${}_5P_{10,t}$) to the population aged 5-9 five-years prior in time $t-1$ (${}_5P_{10-5,t-1}$) to the population aged 0-4 in time t (${}_5P_{10-5,t}$). ie, if we have 100 5-9 year olds five years ago and we now have 125 10-14 year olds and 90 5-9 year olds, we can expect the number of 10-14 year olds in 5 years to be ($125/100 \cdot 90 = 112.5$).

CCRs offer several advantages and disadvantages over the use of a cohort-component model. CCRs are considerably more parsimonious than cohort-component. Calculation of CCRs for use in population projections requires data as minimal as an age-sex distributions at two time periods – data ubiquitous across multiple scales, countries, and time periods. However, this parsimony comes at a relatively steep price: CCRs can lead to impossibly explosive growth in 1) long-range projections due to the natural compounding of the ratios and 2) in small cell sizes with impossibly large CCRs due to a small numeric change in population. As outlined above, consider the growth currently occurring in McKenzie County, North Dakota (FIPS=38053) driven by the Shale oil boom. In 2010 McKenzie had a population of 6,360 that had ballooned to 12,792 by 2015, according to the Vintage 2016 population estimates from the US Census Bureau, with a CCR for the 20-24 year old population of 2.46 (416 to 1,027). Implementing a 50-year population projection using that CCR would create a projected population that is approximately 8,000 times larger (2.46^{10}) – clearly an improbable number given the small, rural nature of its population. Kalawao County, Hawaii (FIPS= 15005) has 2017 estimated population of just 88 persons. Numeric change in any given age-group could lead to impossibly large CCRs in a county as sparsely population as Kalawao County.

Cohort Change Differences

The implementation of CCRs naturally implies a multiplicative model, typically utilizing Leslie matrices. It is possible, however, to implement an **additive** model by using the *difference* in population rather than the *ratio* of population.

$$\begin{aligned} CCD_t &= {}_nP_{x,t} - {}_nP_{x-y,t-1} \\ {}_nP_{x+t} &= CCD_t + {}_nP_{x-y,t} \end{aligned} \tag{4}$$

Where ${}_nP_{x,t}$ is the population aged x to $x+n$ in time t and ${}_nP_{x-y,t}$ is the population aged x to $x+n-y$ in time t where y refers to the time difference between time periods. These CCDs are calculated for each age group a, for each sex group s, for each race group r, in each time period t, in county c. Thus to find the population of ten to fourteen year olds (${}_5P_{10}$) in five years ($t+1$), we add the difference of the population aged 10-14 in time t (${}_5P_{10,t}$) to the population aged 5-9 five-years prior in time $t-1$ (${}_5P_{10-5,t-1}$) to the population aged 0-4 in time t (${}_5P_{10-5,t}$). ie, if we have 100 5-9 year olds five years ago and we now have 125 10-14 year olds and 90 5-9 year olds, we can expect the number of 10-14 year olds in 5 years to be (125-100 + 90 = 115).

CCDs are just as parsimonious as CCRs but have the additional advantage of producing linear growth rather than exponential growth. However, for areas experiencing population declines, CCDs have the potential of creating impossible negative populations through linear decline. A blended approach, using CCDs in areas projected to increase and CCRs in areas projected to decrease creates more utility in the projections and previous research has shown blended linear/exponential population projections outperform both linear and exponential models, respectively (Wilson, 2016).

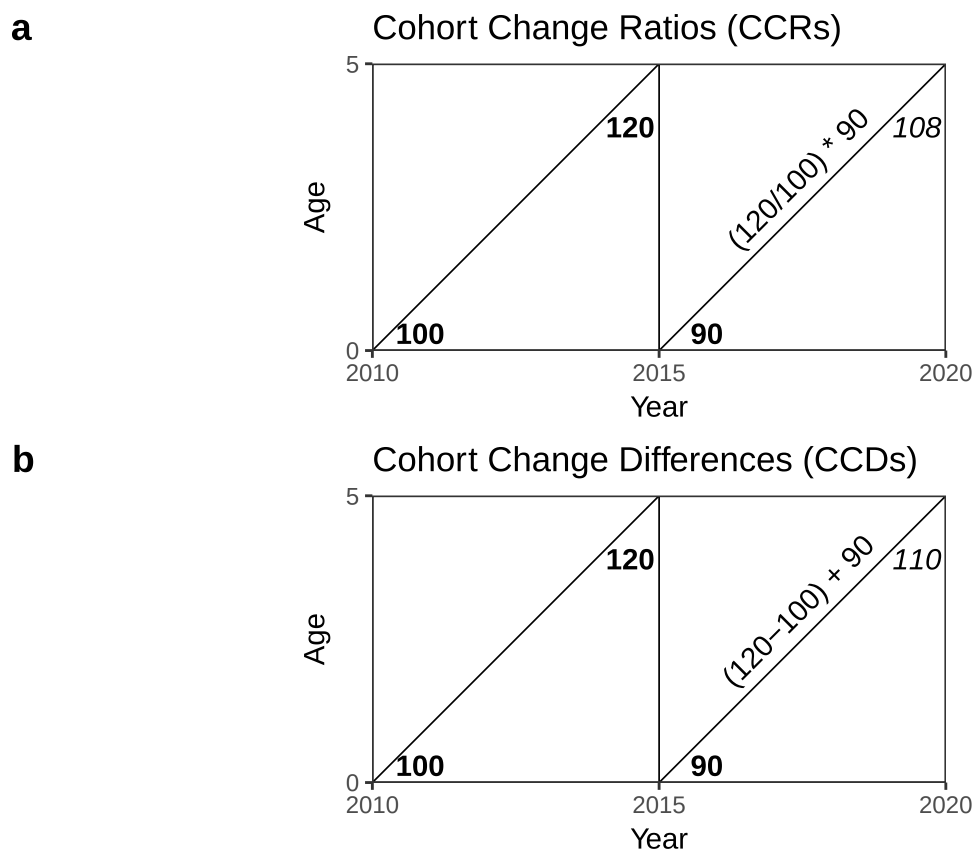

Figure 1: **Lexis Diagrams for CCRs and CCDs.** (a) demonstrates the general framework for Cohort-change ratios and (b) the general framework for cohort-change differences. The observed populations are in bold while the projected populations are italicized.

Projecting CCRs and CCDs

It is unlikely that CCRs/CCDs will remain unchanged over the projection horizon. To account for possible changes in CCRs/CCDs, I employ the use of an Unobserved Components Model (UCM) for forecasting equally spaced univariate time series data (Harvey, 1990). UCMs decompose a time series into components such as trends, seasons, cycles, and regression effects and are designed to capture the features of the series that explain and predict its behavior. UCMs are similar to dynamic models in Bayesian time series forecasting(West, 1996). All projections were undertaken in R using the RUCM package.

The basic structural model (BSM) is the sum of its stochastic components. Here I use an irregular, level, and a random error component and it can be described as:

$$y_t = \mu_t + \sum_{j=1}^{m} \beta_j x_{jt} + \epsilon_t \tag{5}$$
$$\varepsilon_t \sim i.i.d.\ N(0, \theta_\epsilon^2)$$

Each of the model components are modeled separately with the random error ε_t modeled as a sequence of independent, identically distributed zero-mean Gaussian random variables. $\sum_{j=1}^{m} \beta_j x_{jt}$ provides the contribution of the autoregressive component.

The level component is defined as:

$$\mu_t = \mu_{t-1} + \xi_t \tag{6}$$
$$\xi_t \sim i.i.d.\ N(0, \theta_\xi^2)$$

These equations specify a trend where the level μ_t vary over time, governed by the variance of the disturbance term ξ_t in their equations. Here all individual CCRs/CCDs (CCR_{asrc}) over all series are modeled (n=336744) in individual UCMs.

The projected CCRs and CCDs are then input into Leslie matrices to create projected populations (Caswell, 2001).

Equation 7 describes the Leslie matrices for CCRs and Equation 8 describes the Leslie matrices for CCDs.

$$\begin{bmatrix} n_0 \\ n_1 \\ \vdots \\ n_{18} \end{bmatrix}_{t+1} = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ CCR_0 & 0 & 0 & \dots & 0 & 0 \\ 0 & CCR_1 & 0 & \dots & 0 & 0 \\ 0 & 0 & CCR_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 & 0 \\ 0 & 0 & 0 & \dots & CCR_{16} & CCR_{17} \end{bmatrix} \cdot \begin{bmatrix} n_0 \\ n_1 \\ \vdots \\ n_{17} \end{bmatrix}_{t} \tag{7}$$

$$\mathbf{T} = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ CCD_0 & 0 & 0 & \dots & 0 & 0 \\ 0 & CCD_1 & 0 & \dots & 0 & 0 \\ 0 & 0 & CCD_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 & 0 \\ 0 & 0 & 0 & \dots & CCD_{16} & CCD_{17} \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ n_0 & 0 & 0 & \dots & 0 & 0 \\ 0 & n_1 & 0 & \dots & 0 & 0 \\ 0 & 0 & n_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 & 0 \\ 0 & 0 & 0 & \dots & n_{16} & n_{17} \end{bmatrix} \tag{8}$$

$$\begin{bmatrix} n_0 \\ n_1 \\ \vdots \\ n_{18} \end{bmatrix}_{t+1} = \begin{bmatrix} \sum \mathbf{T_{1j}} \\ \sum \mathbf{T_{2j}} \\ \vdots \\ \sum \mathbf{T_{17j}} \end{bmatrix}$$

Equation 7 and Equation 8 both require special consideration for two specific age groups: the population aged 0-4 (${}_5P_0$) and the population comprising the open-ended interval (${}_\infty P_{85}$; CCR_{17} and CCD_{17}). The population aged 0-4 (${}_5P_0$) and 85+ (${}_\infty P_{85}$) must have special consideration since the preceding/proceeding age groups do not exist for these age groups.

To project 0-4 year olds, I use the child-woman ratio (CWR)

$$
\begin{aligned}
CWR_t &= \frac{{}_5P_{0,t}}{{}_{45}W_{15,t}} \qquad (9)\\
{}_nP_{x+t} &= CWR_t \cdot {}_{45}W_{15,t+1}
\end{aligned}
$$

Where ${}_{45}W_{15}$ is the population of women in child-bearing ages 15-50. I use the state/race-specific CWRs for member counties.

The population aged 0-4 in time $t+1$ are projected by applying a 1.05 sex ratio at birth (SRB) to the projected children born of women of childbearing age $[15, 50)$ in time $t+1$.

To calculate the CCR/CCD for the open-ended age group,

$$
\begin{aligned}
{}_\infty CCR_{85,t} &= \frac{{}_\infty P_{85,t}}{{}_\infty P_{85-y,t-1}} \qquad (10)\\
{}_\infty P_{85+t} &= {}_\infty CCR_{85,t} \cdot {}_\infty P_{85-y,t}\\
{}_\infty CCD_{85,t} &= {}_\infty P_{85,t} - {}_\infty P_{85-y,t-1} \qquad (11)\\
{}_\infty P_{85+t} &= {}_\infty CCD_{85,t} + {}_\infty P_{85-y,t}
\end{aligned}
$$

If a given race/county combination is projected to increase, I use CCDs and if a given race/county combination is projected to decline, I use CCRs.

DATA

Data used to project the populations consist of a single primary data source: the National Vital Statistics System U.S. Census Populations with Bridged Race Categories data set[1]. These data harmonize racial classifications across disparate time periods to allow population estimates to be sufficiently comparable across space and time. All county boundaries are generally rectified as well. The National Center for Health Statistics bridge the 31 race categories used in Census 2000 and 2010 with the four race categories used in the 1977 Office of Management and Budget standards.

There are two primary bridged-race data sets. The first covers the time period 1969-2016 and utilizes three race groups: White, Black, and Other. The second covers the time period 1990-

[1]Data can be downloaded here: https://seer.cancer.gov/popdata/download.html

2016 and uses four race groups (White, Black, American Indian/Alaska Native, and Asian/Pacific Islander) as well as two origin groups (Hispanic and Non-Hispanic). Evaluation of the population projections makes use of the three race group data set covering 1969-2016.

Extra considerations

Group quarters: All *resident* populations are projected in this modelling scheme such that the populations at launch year are equal to the total population minus the group quarters population. Group quarters populations at time t are then added back into the resident population at time $t + 1$.

Miscellaneous In the event a UCM contained NA or infinite values or produced covariance matrices with values larger than 10,000,000, the projections were set to 0. Upper and Lower bounds of failed UCMs were set to 0. Any infinite, NA, or NAN CCR, CCD, or CWR was set to 0. Any projected negative populations are also set to 0.

Projection Controls

As shown below, any set of population projections are likely to produce higher than expected projections (see Table 1). To prevent runaway population growth, I control the projected output to the Shared Socioeconomic Pathways (SSPs) (O'Neill et al., 2014). The SSPs are socio-economic scenarios that derive emissions scenarios coupled with climate policies. They are designed to evaluate both climate change impacts and adaptation measures in harmony with the Representative Concentration Pathways (RCPs) for emission scenarios. Scholars have downscaled the SSPs to gridded population projections (Jones and O'Neill, 2016), while these projections are incredibly useful, they lack detailed demographic characteristics.

The five SSPs are colloquially named SSP1 (Sustainability), SSP2 (Middle of the Road), SSP3 (Regional Rivalry), SSP4 (Inequality), and SSP5 (Fossil-fueled Development) (O'Neill et al., 2017). These five SSPs cover potential futures involving various growth policies, fossil-fuel usage, mitigation policies, adaptation policies, and population change (Samir and Lutz, 2017).

Each SSP contains projected population information in five-year increments for 5-year age groups (0-100+) and two sex groups (Male and Female) for the period 2020-2100 and I truncate the open-ended interval from 100+ to 85+ to be consistent with US Census Bureau population estimates. I control my projected age/sex/race/county projections to the the SSPs by using

$$P_t = \frac{p_{asrc}}{p_{as}} \cdot P_{as,SSP} \tag{12}$$

where p_{asrc} refers to the age/sex/race/county specific population projected as outlined above, p_{as} refers to the age/sex specific population projection, and $P_{as,SSP}$ refers to the age/sex specific population projection for each SSP.

EVALUATIONS

To evaluate the projection accuracy, I use the base period 1969-2000 to project the population for eighteen age groups, two sexes, three races (White, Black, Other), and 3134 counties for the projection period 2000-2015. I utilize an ex-post facto analysis at periods 2005, 2010, and 2015 using a pure CCD model, a pure CCR model, and blended model CCR/CCD). The CCR/CCD model utilizes CCDs if a county is projected to grow and CCRs if it is projected to decline. Blended models have been shown to outperform both purely linear or purely exponential models in simple extrapolation approaches to population projections (Wilson, 2016).

In keeping with demographic tradition (Hauer, Baker and Brown, 2013; Booth, 2006), I evaluate the projections using three primary statistics. To determine the overall accuracy of the projections, I use Absolute Percent Errors (APE), to determine the bias of the projections I use the Algebraic Percent Error (ALPE), and to determine the accuracy of the uncertainty interval I evaluate the percentage of actual counts within the 80th percentage projection interval. In some places I have substituted a Symmetric Absolute Percent Error (SAPE) (Shcherbakov et al., 2013).

Equations 13 describe the equations used to evaluate errors. P_i refers to the projected value and A_i refers to the actual, observed value.

$$APE = |\frac{P_i}{A_i}| \tag{13}$$
$$ALPE = \frac{P_i}{A_i} \tag{14}$$
$$SAPE = \frac{|(P_i - A_i)|}{(P_i + A_i)} \tag{15}$$

Overall Errors

Table 1 reports the overall errors for the sum of the population for the whole US. Overall the pure CCD model outperformed the purely CCR model, suggesting CCDs in this model could produce more accurate results compared to CCRs. It should also be noted that all model variants (CCD, CCR, and CCR/CCD) have a tendency to over-project the overall population in the United States.

Table 1: **Evaluation of overall total errors for the entire United States.**

TYPE	YEAR	POPULATION	PRED	APE
CCD	2005	292,555,450	297,152,440	1.57%
CCD	2010	306,397,813	313,792,252	2.41%
CCD	2015	317,731,270	330,800,311	4.11%
CCD/CCR	2005	292,555,398	297,450,239	1.67%
CCD/CCR	2010	306,397,691	314,409,947	2.61%
CCD/CCR	2015	317,731,063	331,726,190	4.40%
CCR	2005	292,555,450	298,969,168	2.19%
CCR	2010	306,397,813	320,306,395	4.54%
CCR	2015	317,731,270	349,770,119	10.08%

?? reports the overall errors for the sum of the population in each of the counties. Here we can see that for the average county, the CCD and CCR/CCD models produce similar Apes but the CCR/CCD model tends to produce slightly lower Apes when compared to the purely CCD model.

In all cases, the errors associated with the CCR model are greater than the CCD or CCR/CCD varieties.

Table 2: **Evaluation of overall errors for each county.**

TYPE	n	EVAL	2005	2010	2015
CCD	3134	Median APE	2.4964%	5.089%	8.30%
CCD/CCR	3134	Median APE	2.5122%	4.951%	7.87%
CCR	3134	Median APE	2.5896%	5.300%	9.05%
CCD	3134	Median ALPE	1.140%	1.486%	3.67%
CCD/CCR	3134	Median ALPE	1.285%	1.755%	4.09%
CCR	3134	Median ALPE	1.314%	2.312%	5.25%
CCD	3134	In 80th percentile	74.28%	74.1%	78.8%
CCD/CCR	3134	In 80th percentile	72.85%	70.8%	74.0%
CCR	3134	In 80th percentile	82.04%	82.6%	86.3%

Figure 2 shows the absolute percent errors associated with the total population for the CCR/CCD model in U.S. counties in 2015. Most states and counties see relatively low errors with the median APE of just 8.29% by 2015, however some isolated pockets of high errors do exist randomly distributed throughout the United States. Additionally, 94.8% of counties had observed population totals within the 80th percentile prediction interval with the CCR/CCD model by 2015. This large number of counties could suggest that prediction bands at at the scale of population totals in counties could be too wide.

Age Structure Error

?? reports the overall errors for age groups at the county level. All three models produce similar Apes. For any given county, the average error is approximately 11%. Similar to the overall errors, the bias tends to be for over-projection of age groups as all of the ALPEs are positive.

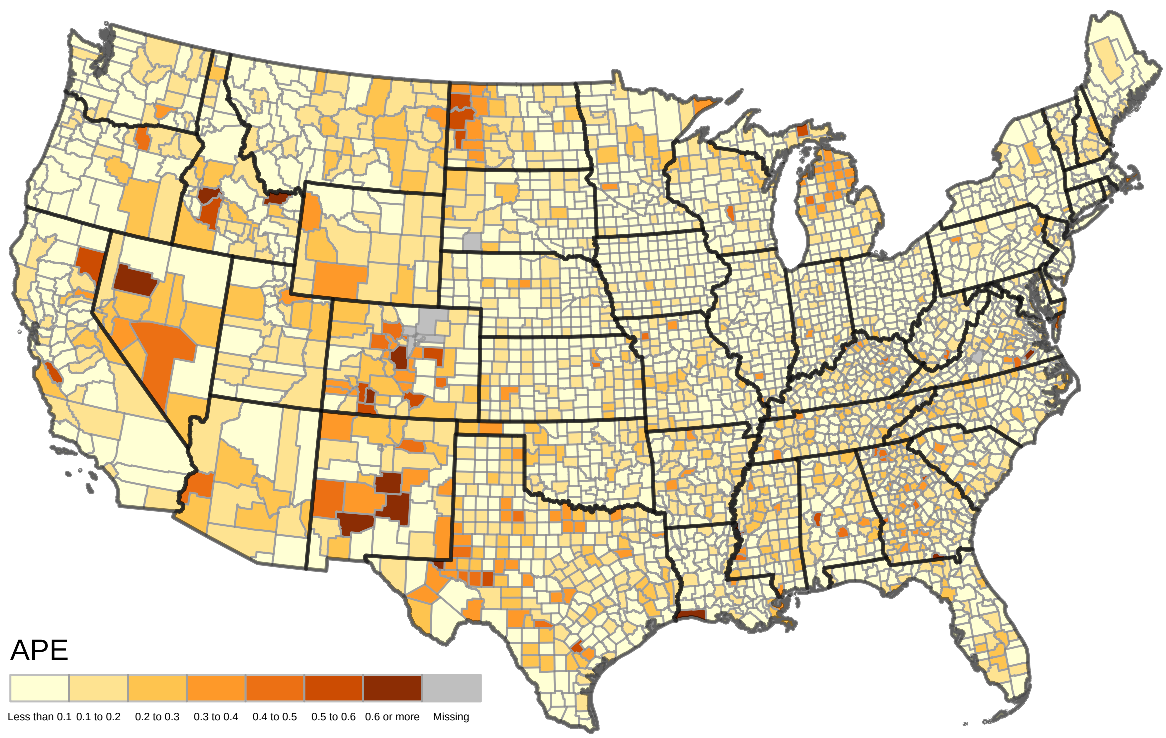

Figure 2: **Map of county errors of the total population in 2015 using the CCR/CCD model.** Here I show the geographic distribution of absolute percent errors. Most states and counties have low error rates of the total population with isolated pockets of large errors.

Table 3: **Evaluation of Age Group Errors.**

TYPE	n	EVAL	2005	2010	2015
CCD	56412	Median APE	5.3940%	8.253%	11.537%
CCD/CCR	56412	Median APE	5.3288%	7.989%	10.869%
CCR	56412	Median APE	5.4050%	8.159%	11.323%
CCD	56412	Median ALPE	1.000%	1.122%	3.220%
CCD/CCR	56412	Median ALPE	1.212%	1.490%	3.608%
CCR	56412	Median ALPE	1.321%	1.568%	3.372%
CCD	56412	In 80th percentile	64.79%	71.25%	69.10%
CCD/CCR	56412	In 80th percentile	62.93%	69.87%	68.24%
CCR	56412	In 80th percentile	68.46%	75.19%	73.63%

Figure 3 shows projected age structures in nine samples counties across three county types – college counties, suburban counties, and retirement counties. In all three county types the age structures are preserved in the projections. All three county types exhibit differing age structures with important considerations. For college counties, the college-age population (those aged 15-24) do not age in place within those communities. The large population peaks in those counties show great in-migration at the college ages and then great out-migration afterwards. In suburban counties, a "double hump" age structure is typically present with large numbers of both adolescents and middle-aged adults. Most twenty-somethings either cannot afford to live in affluent suburban areas, move away for school or work, or do not have the family reasons for living there. Retirement communities are often identified by the large numbers of populations over the age of 55.

Figure 4 shows the Algebraic Percent Errors by age group for all three evaluation periods. For every age below 60, the CCD and CCR/CCD models produce lower ALPEs, but for age groups over 60, the CCR model produces lower ALPEs. This could be reflective of mortality being the dominant population process for older age populations and is especially pronounced in the 85+ age group. Here, the 50th percentile ALPE for CCD and CCR/CCD is approximately 20% while the

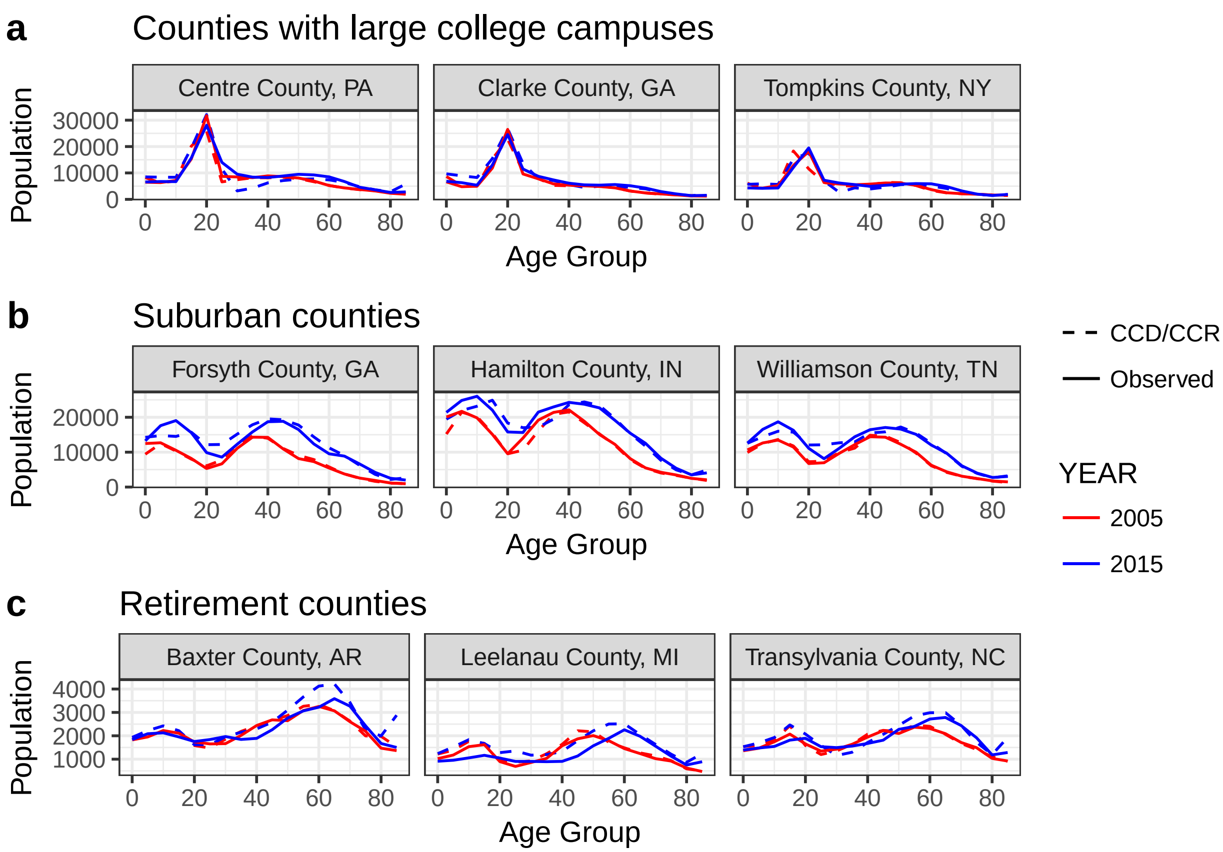

Figure 3: **Age structures of various county types.** I compare the projected age structures to the observed age structures in nine counties across three county types using the CCR/CCD model. (a) demonstrates counties with major universities, (b) demonstrates sample suburban counties, and (c) demonstrates sample retirement counties. All three county types have age structures largely preserved despite widely different age structures.

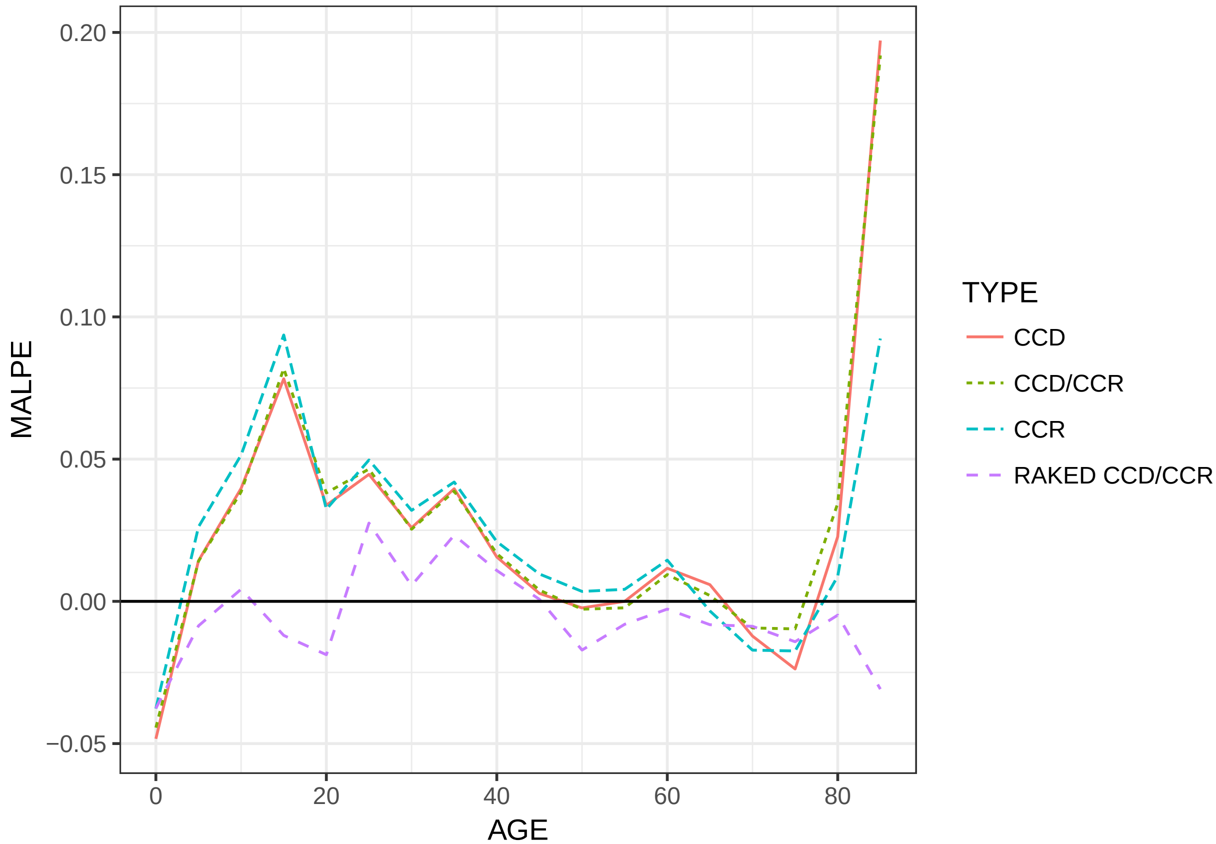

Figure 4: **Algebraic Percent Errors by age group.** I plot the 50th percentile Algebraic Percent Error (ALPE) by age group.

Mult model is just under 10%. This bias is virtually eliminated when controlling the populations to the Age/Sex total of the United States.

Race Errors

Figure 5 reports the ALPE and the APE distribution by race group for all counties. The White race group tends to have the lowest errors associated with the projections, followed by Black, and then Other.

Age, Sex, Race joint errors

Finally, I show the joint errors associated with all possible Age/Sex/Race/County combinations. Here the average error for any given ASRC combination (such as Black Females aged 20-24 in Lincoln County NV) are approximately 11-12% for all three methods after 15 years. In contrast

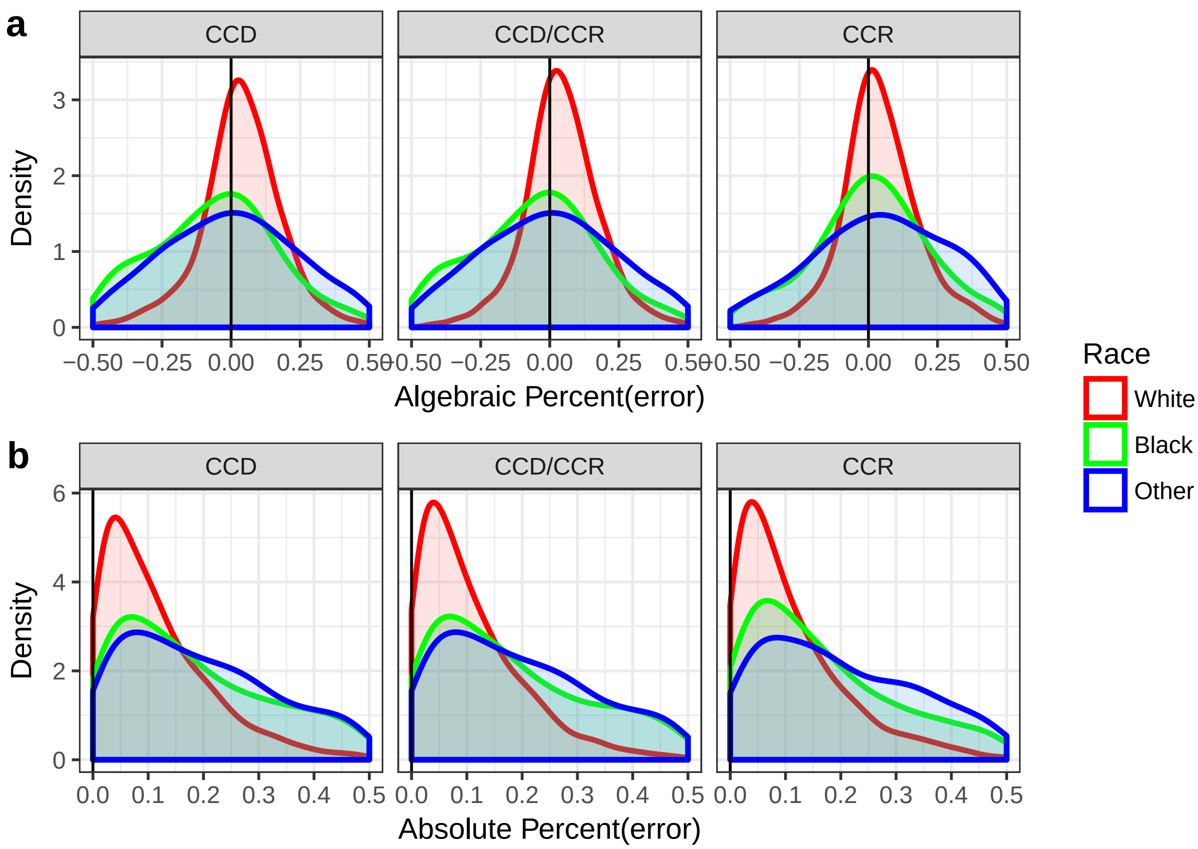

Figure 5: **Race group errors. (a) shows the Algebraic Percent Errors for all three methods and (b) shows the APE distribution of errors.**

to the confidence bounds being too wide when discussing the overall total populations in counties, when examining any given ASRC combination it appears that the projection intervals are too narrow for all three methods. Between two-thirds and three-fourths of observed populations fall within the 80th percentile.

Table 4: **Evaluation of Age/Sex/Race/County joint Errors.**

TYPE	num	EVAL	2005	2010	2015
CCD	336744	Median SAPE	6.3291%	8.869%	11.60%
CCD/CCR	336240	Median SAPE	6.2490%	8.597%	11.11%
CCR	336744	Median SAPE	6.2421%	8.929%	12.67%
CCD	336744	In 80th percentile	44.47%	52.74%	53.35%
CCD/CCR	336240	In 80th percentile	43.62%	51.94%	52.63%
CCR	336744	In 80th percentile	53.29%	60.25%	58.92%

PROJECTIONS

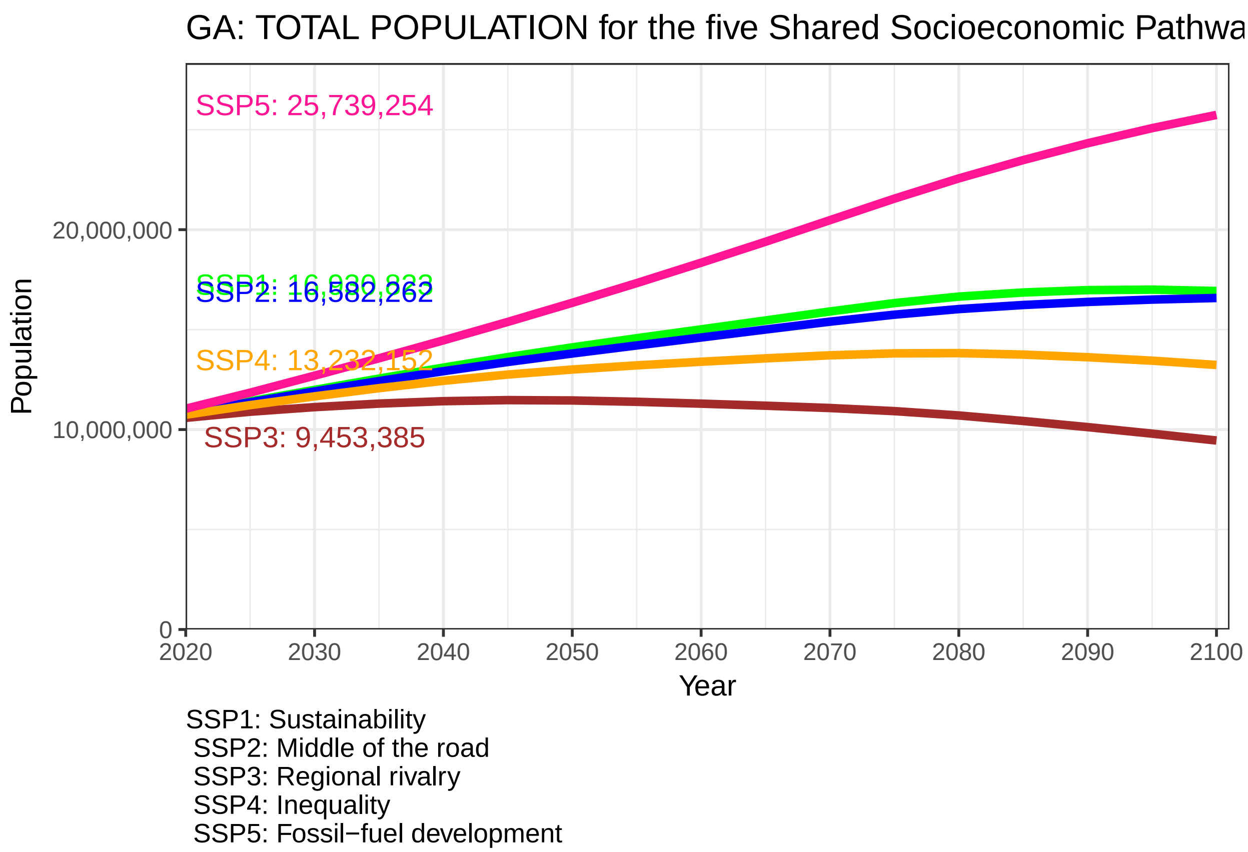

Figure 6: **Sample projected populations for Georgia.** Projections are aggregated to the state total for each five SSPs.

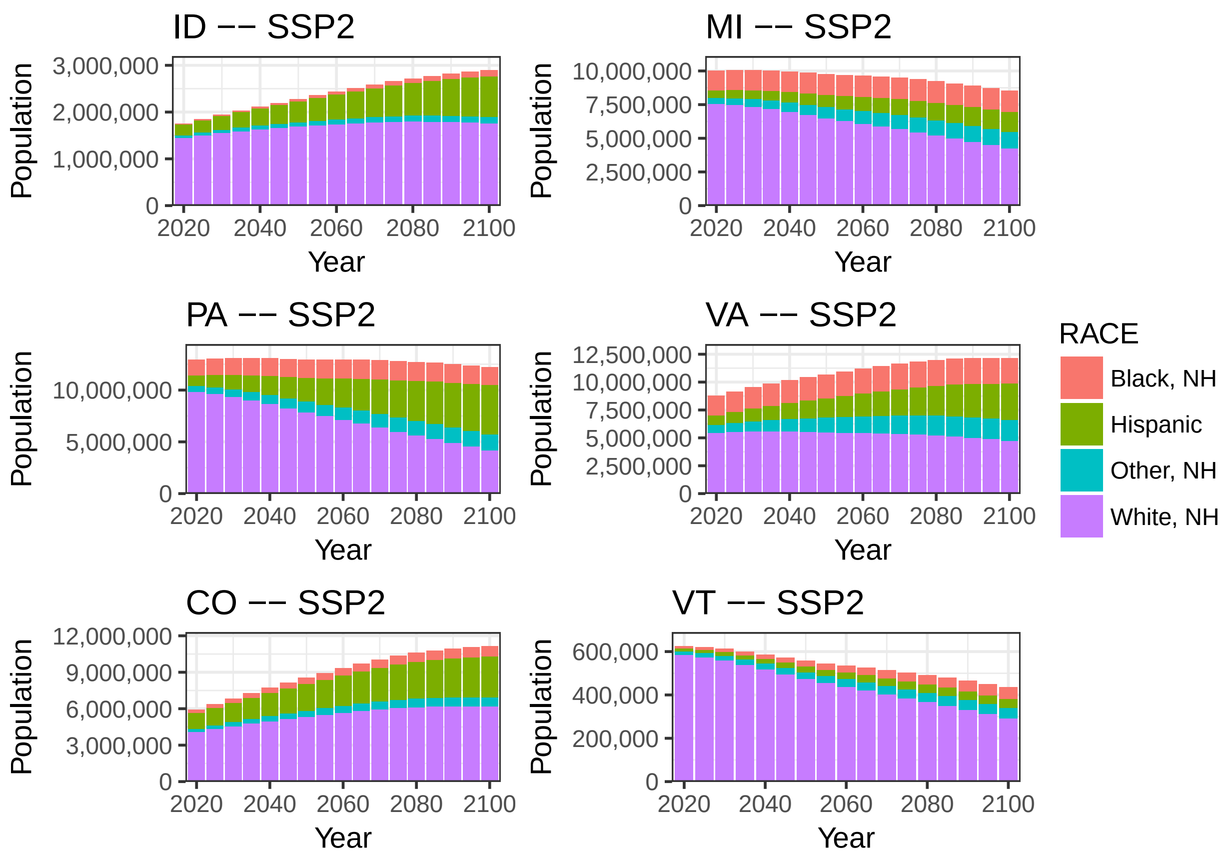

Figure 7: **Sample projected populations for six random states by race.** Projections are aggregated by race for SSP2.

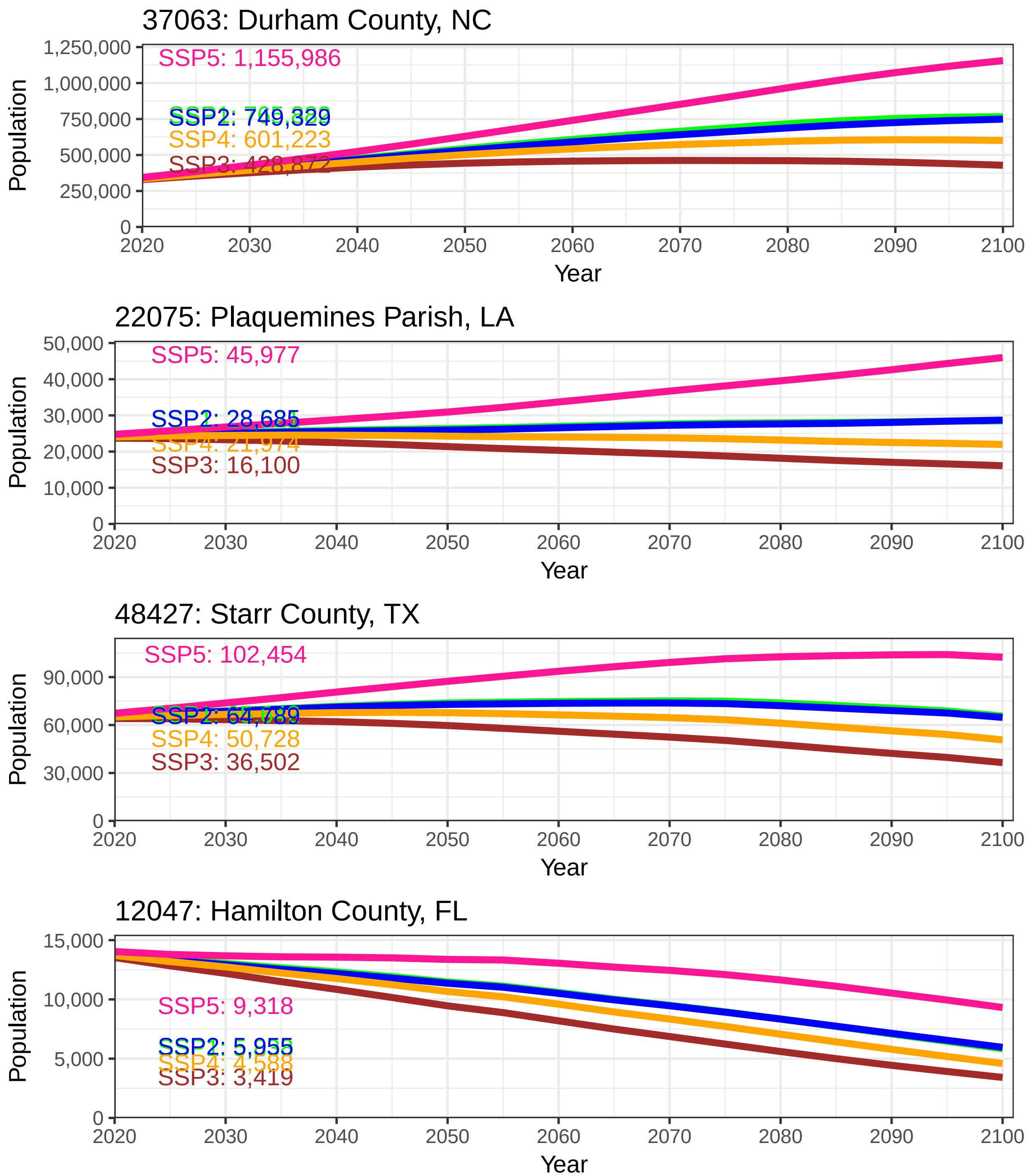

Figure 8: **Sample projected populations for four random counties.** Projections are aggregated to the total population for each SSP.

References

Alexander, Monica, Emilio Zagheni and Magali Barbieri. 2017. “A Flexible Bayesian model for estimating subnational mortality.” *Demography* 54(6):2025–2041.

Baker, Jack, David A Swanson, Jeff Tayman and Lucky M Tedrow. 2017. *Cohort change ratios and their applications.* Springer.

Booth, Heather. 2006. “Demographic forecasting: 1980 to 2005 in review.” *International Journal of Forecasting* 22(3):547–581.

Boyle, James P, Theodore J Thompson, Edward W Gregg, Lawrence E Barker and David F Williamson. 2010. “Projection of the year 2050 burden of diabetes in the US adult population: dynamic modeling of incidence, mortality, and prediabetes prevalence.” *Population health metrics* 8(1):29.

Caswell, Hal. 2001. *Matrix population models.* Wiley Online Library.

Chi, Guangqing. 2009. “Can knowledge improve population forecasts at subcounty levels?” *Demography* 46(2):405–427.

Colby, Sandra L and Jennifer M Ortman. 2017. “Projections of the size and composition of the US population: 2014 to 2060: Population estimates and projections.”.

Daponte, Beth Osborne, Joseph B Kadane and Lara J Wolfson. 1997. “Bayesian demography: projecting the Iraqi Kurdish population, 1977–1990.” *Journal of the American Statistical Association* 92(440):1256–1267.

Gerland, Patrick, Adrian E Raftery, Hana Ševčíková, Nan Li, Danan Gu, Thomas Spoorenberg, Leontine Alkema, Bailey K Fosdick, Jennifer Chunn and Nevena Lalic. 2014. “World population stabilization unlikely this century.” *Science* 346(6206):234–237.

Hales, Simon, Neil De Wet, John Maindonald and Alistair Woodward. 2002. “Potential effect of population and climate changes on global distribution of dengue fever: an empirical model.” *The Lancet* 360(9336):830–834.

Hamilton, C Horace and Josef Perry. 1962. “A short method for projecting population by age from one decennial census to another.” *Social Forces* 41(2):163–170.

Harvey, Andrew C. 1990. *Forecasting, structural time series models and the Kalman filter.* Cambridge university press.

Hauer, Mathew E, Jason M Evans and Deepak R Mishra. 2016. “Millions projected to be at risk from sea-level rise in the continental United States.” *Nature Climate Change* 6(7):691–695.

Hauer, Matt, Jack Baker and Warren Brown. 2013. “Indirect Estimates of Total Fertility Rate Using Child Woman/Ratio: A Comparison with the Bogue-Palmore Method.” *PloS one* 8(6):e67226.

Hebert, Liesi E, Paul A Scherr, Julia L Bienias, David A Bennett and Denis A Evans. 2003. “Alzheimer disease in the US population: prevalence estimates using the 2000 census.” *Archives of neurology* 60(8):1119–1122.

Jones, Bryan and BC O’Neill. 2016. “Spatially explicit global population scenarios consistent with the Shared Socioeconomic Pathways.” *Environmental Research Letters* 11(8):084003.

Lutz, Wolfgang, Warren C Sanderson and Sergei Scherbov. 1996. "Probabilistic world population projections based on expert opinion.".

Martin, Joyce A, Brady E Hamilton, Michelle JK Osterman, Anne K Driscoll and Patrick Drake. 2018. "Births: final data for 2016.".

O'Neill, Brian C, Elmar Kriegler, Keywan Riahi, Kristie L Ebi, Stephane Hallegatte, Timothy R Carter, Ritu Mathur and Detlef P van Vuuren. 2014. "A new scenario framework for climate change research: the concept of shared socioeconomic pathways." *Climatic Change* 122(3):387–400.

O'Neill, Brian C, Elmar Kriegler, Kristie L Ebi, Eric Kemp-Benedict, Keywan Riahi, Dale S Rothman, Bas J van Ruijven, Detlef P van Vuuren, Joern Birkmann, Kasper Kok et al. 2017. "The roads ahead: narratives for shared socioeconomic pathways describing world futures in the 21st century." *Global Environmental Change* 42:169–180.

Passel, Jeffrey S and DVUS Cohn. 2008. "US population projections: 2005-2050.".

Preston, Samuel, Patrick Heuveline and Michel Guillot. 2000. "Demography: measuring and modeling population processes.".

Raftery, Adrian E, Nan Li, Hana Ševčíková, Patrick Gerland and Gerhard K Heilig. 2012. "Bayesian probabilistic population projections for all countries." *Proceedings of the National Academy of Sciences* 109(35):13915–13921.

Rayer, Stefan. 2008. "Population forecast errors: A primer for planners." *Journal of Planning Education and Research* 27(4):417–430.

Raymer, James, Guy J Abel and Andrei Rogers. 2012. "Does specification matter? Experiments with simple multiregional probabilistic population projections." *Environment and Planning A* 44(11):2664–2686.

Samir, KC and Wolfgang Lutz. 2017. "The human core of the shared socioeconomic pathways: Population scenarios by age, sex and level of education for all countries to 2100." *Global Environmental Change* 42:181–192.

Shcherbakov, Maxim Vladimirovich, Adriaan Brebels, Nataliya Lvovna Shcherbakova, Anton Pavlovich Tyukov, Timur Alexandrovich Janovsky and Valeriy Anatol'evich Kamaev. 2013. "A survey of forecast error measures." *World Applied Sciences Journal* 24:171–176.

Smith, Stanley K. 1997. "Further thoughts on simplicity and complexity in population projection models." *International journal of forecasting* 13(4):557–565.

Smith, Stanley K and Jeff Tayman. 2003. "An evaluation of population projections by age." *Demography* 40(4):741–757.

Smith, Stanley K, Jeff Tayman and David A Swanson. 2006. *State and local population projections: Methodology and analysis.* Springer Science & Business Media.

Smith, Stanley K, Jeff Tayman and David A Swanson. 2013. *A Practitioner's Guide to State and Local Population Projections.* Springer.

Sprague, W Webb. 2012. "Automatic parametrization of age/sex Leslie matrices for human populations." *arXiv preprint arXiv:1203.2313* .

Swanson, David A, Alan Schlottmann and Bob Schmidt. 2010. "Forecasting the population of census tracts by age and sex: An example of the Hamilton–Perry method in action." *Population Research and Policy Review* 29(1):47–63.

Tatem, Andrew J, Susana Adamo, Nita Bharti, Clara R Burgert, Marcia Castro, Audrey Dorelien, Gunter Fink, Catherine Linard, Mendelsohn John, Livia Montana et al. 2012. "Mapping populations at risk: improving spatial demographic data for infectious disease modeling and metric derivation." *Population health metrics* 10(1):8.

Tiwari, Chetan, Kirsten Beyer and Gerard Rushton. 2014. "The impact of data suppression on local mortality rates: The case of CDC WONDER." *American journal of public health* 104(8):1386–1388.

West, Mike. 1996. *Bayesian forecasting.* Wiley Online Library.

Wilson, Tom. 2012. "Forecast accuracy and uncertainty of Australian Bureau of Statistics state and territory population projections." *International Journal of Population Research* 2012.

Wilson, Tom. 2016. "Evaluation of alternative cohort-component models for local area population forecasts." *Population Research and Policy Review* 35(2):241–261.

Wilson, Tom and Phil Rees. 2005. "Recent developments in population projection methodology: A review." *Population, Space and Place* 11(5):337–360.

Wonder, C. 2016. "Bridged-race population estimates." *Atlanta, GA: Centers for Disease Control and Prevention* .