U.S. County level population projections by age, sex, and race for the period 2016-2066 *

Mathew E. Hauer [1*] *University of Georgia*

I provide county-level population projections by age, sex, and race in five-year intervals for the period 2016-2066 for 3,136 counties. Using historic U.S. census data in temporally rectified county boundaries and race groups for the period 1990-2016, I calculate cohort-change ratios (CCRs) and cohort-change differences (CCDs) for eighteen five-year age groups (0-85+), two sex groups (Male and Female), and four race groups (White NH, Black NH, Other NH, Hispanic). I then project these CCRs/CCDs using Unobserved Components Models as inputs into leslie matrix population projection models for a blended CCD/CCR population projection. My ex-post facto evaluations using three race groups (White, Black, Other) on the 1980-2000 base period evaluated at 2005, 2010, and 2015 demonstrate confidence in the accuracy of the projections. These data have numerous potential uses and can serve as inputs for addressing questions involving sub-national population change in the United States.

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

[1] Carl Vinson Institute of Government, University of Georgia. 201 N. Milledge Ave. Athens, GA USA 30602.

*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 [CITES].

Producing high-quality, highly-detailed population projections is a challenging endeavor and no rigourous set of U.S. national projections currently exists. With such a large need for sub-national projections and to better understand the changing demographics of the U.S. population, I sought to produce such a set of high-quality, highly-detailed projections and make both the R code and subsequent projections available for dissemination with a wide audience. Here, I present age-sex-race specific population projections for all U.S. counties and their uncertainty, an ex-post facto evaluation of the projection methodology, and details on the calculations of these projections. I generate these projections using a historic time series of population estimates for the period 1990-2016 in temporally rectified county boundaries and race groupings using leslie matrices populated by cohort-change ratios (CCRs) and cohort-change differences (CCDs) projected through the use of Unobserved Component Models (UCMs) in a combined additive/multiplicative model.

To ensure quality projections, I employ the use of ex-post-facto evaluations of the projection accuracy for three variant models: purely additive with CCDs, purely multiplicative with CCRs, and a blended model with CCDs in areas projected to grow and CCRs in areas projected to decline. I report the accuracy, bias, and uncertainties associated with these variants using absolute percent error, algebraic percent error, and the number of observations where the observed population is within the 80th percentile projection interval. Overall, the errors reported here are on par with deterministic cohort-component population projection models undertaken at the county level in individual states [CITES] and with Bayesian cohort-component projection models undertaken at the national scale [CITES].

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. 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 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 dataset 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. Birth data, however, are only available for counties with populations greater than 100k. These limitations surrounding fertility and migration render a universal county-

level population projection difficult, if not impossible, to complete using publicly available datasets.

An alternative to cohort-component is the Hamilton-Perry method [CITES], which uses cohort-change ratios (CCRs) in place of components to project populations. The basic CCR equation is found in equation 3.

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

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 ($125/100 \cdot 90 = 112.5$).

Two age groups must have special consideration: the population aged 0-4 (${}_5P_0$) and the population comprising the open-ended interval (${}_\infty P_{85}$). The population aged 0-4 (${}_5P_0$) 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)

$$
\begin{aligned}
CWR_t &= \frac{{}_5P_{0,t}}{{}_{45}W_{15,t}} \\
{}_nP_{x+t} &= CWR_t \cdot {}_{45}W_{15,t+1}
\end{aligned}
\tag{3}
$$

Where ${}_{45}W_{15}$ is the population of women in child-bearing ages 15-50.

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

$$
\begin{aligned}
{}_\infty CCR_{85,t} &= \frac{{}_\infty P_{85,t}}{{}_\infty P_{85-y,t-1}} \\
{}_\infty P_{85+t} &= {}_\infty CCR_{85,t} \cdot {}_\infty P_{85-y,t}
\end{aligned}
\tag{4}
$$

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 long-range projections due to the natural compounding of the ratios. 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.

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.

$$\begin{aligned} CCD_t &= {}_nP_{x,t} - {}_nP_{x-y,t-1} \\ {}_nP_{x+t} &= CCD_t + {}_nP_{x-y,t} \end{aligned} \tag{5}$$

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 differene 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 parimonious 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 at the cost of some accuracy (see **??**).

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 (Harrison and West 1999). 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:

$$y_t = \mu_t + \sum_{j=1}^{m} \beta_j x_{jt} + \epsilon_t \qquad (6)$$
$$\varepsilon_t \sim i.i.d.\ N(0, \theta_\epsilon^2)$$

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:

$$\mu_t = \mu_{t-1} + \xi_t \qquad (7)$$
$$\xi_t \sim i.i.d.\ N(0, \theta_\xi^2)$$

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_{iasr}) over all series are modelled (n=16) in individual UCMs.

Rather than use the prediction intervals output from the UCMs, I set the upper and lower bounds as the projected UCM plus or minus the 80th percentile based on the standard deviation of the original time series.

For the CWRs, I projected the CWRs within a constrained forecast interval. CWRs are constrained to lie between (a, b). I limit CWRs such that each age/race/county combination are be constrained within the maximum/minimum of the time series such that $a = 0.14$ for all projections and $b = max(CWR_{arc})$. I then transform the data using a scaled logit transformation to map (a, b) to the whole real line:

$$y = log\Big(\frac{x-a}{b-x}\Big) \tag{8}$$

Where x is the original data and y is the transformed data.

The projected CCRs and CCDs are then input into Leslie matrices to create projected populations.

Equation 9 describes the leslie matrices for CCRs and equation 10 describes the leslie matrices for CCDs.

$$\begin{bmatrix} n_0 \\ n_1 \\ \vdots \\ n_{18} \end{bmatrix}_{t+1} = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ CCR_0 & 0 & 0 & \dots & 0 & 0 \\ 0 & CCR_1 & 0 & \dots & 0 & 0 \\ 0 & 0 & CCR_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 & 0 \\ 0 & 0 & 0 & \dots & CCR_{16} & CCR_{17} \end{bmatrix} \cdot \begin{bmatrix} n_0 \\ n_1 \\ \vdots \\ n_{17} \end{bmatrix}_t \tag{9}$$

$$\mathbf{T} = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ CCD_0 & 0 & 0 & \dots & 0 & 0 \\ 0 & CCD_1 & 0 & \dots & 0 & 0 \\ 0 & 0 & CCD_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 & 0 \\ 0 & 0 & 0 & \dots & CCD_{16} & CCD_{17} \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ n_0 & 0 & 0 & \dots & 0 & 0 \\ 0 & n_1 & 0 & \dots & 0 & 0 \\ 0 & 0 & n_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 & 0 \\ 0 & 0 & 0 & \dots & n_{16} & n_{17} \end{bmatrix} \tag{10}$$

$$P_{t+1} \equiv \begin{bmatrix} \sum_{j=1}^{n} \mathbf{T_{1,j}} \\ \sum_{j=1}^{n} \mathbf{T_{2,j}} \\ \vdots \\ \sum_{j=1}^{n} \mathbf{T_{17,j}} \end{bmatrix}$$

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 $t+1$.

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. These data harmoize 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 1990-2016 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 dataset 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.

EVALUATIONS

To evaluate the projection accuracy, I use the base period 1980-2000 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 (named ADD), a pure CCR model (named MULT), and blended model (named ADDMULT). The ADDMULT model utilizes CCDs if a county is projected to grow and CCRs if it is projected to decline.

In keeping with demographic tradition [CITES], I evaluate the projections using three primary statistics. To determine the overall accuray 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 subsituted a Symmetric Absolute Percent Error (SAPE).

Equations 11 describe the equations used to evaluate errors. P_i refers to the projected value and A_i refers to the actual, observed value.

$$APE = |\frac{P_i}{A_i}| \tag{11}$$

$$ALPE = \frac{P_i}{A_i} \tag{12}$$

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

Overall Errors

Table **1** reports the overall errors for the sum of the population in each of the subsequent states and counties. Overall the purely ADDITIVE model outperformed the purely MULTIPLICATIVE model, suggesting CCDs in this model could produce more accurate results compared to CCRs. It should also be noted that all model variants (ADD, MULT, and ADDMULT) 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	LOW	HIGH	APE
ADD	2005	292,554,817	297,219,886	276,264,815	318,659,413	1.59%
ADD	2010	306,396,740	313,983,491	271,208,688	358,388,724	2.48%
ADD	2015	317,729,565	331,173,497	265,990,045	400,103,409	4.23%
ADDMULT	2005	292,554,817	297,516,199	277,008,471	318,552,479	1.70%
ADDMULT	2010	306,396,740	314,705,975	273,358,793	358,991,660	2.71%
ADDMULT	2015	317,729,565	332,461,792	269,649,687	404,532,006	4.64%
Mult	2005	292,554,817	299,142,100	272,199,340	327,496,956	2.25%
Mult	2010	306,396,740	320,917,411	266,253,047	395,742,885	4.74%

TYPE	YEAR	POPULATION	PRED	LOW	HIGH	APE
Mult	2015	317,729,565	351,252,666	260,203,811	530,377,085	10.55%

Table **2** 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 ADDITIVE and ADDMULT models produce similar APEs but the ADDMULT model tends to produce slightly lower APEs when compared to the purely ADD model. In all cases, the errors associated with the MULT model are greater than the ADD or ADDMULT varieties.

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

TYPE	n	EVAL	2005	2010	2015
ADD	3134	Median APE	2.617%	5.276%	8.63%
ADDMULT	3134	Median APE	2.583%	5.201%	8.29%
Mult	3134	Median APE	2.709%	5.540%	9.40%
ADD	3134	Median ALPE	1.190%	1.687%	4.13%
ADDMULT	3134	Median ALPE	1.287%	1.870%	4.76%
Mult	3134	Median ALPE	1.457%	2.553%	5.85%
ADD	3134	In 80th percentile	95.98%	96.04%	94.77%
ADDMULT	3134	In 80th percentile	95.92%	96.04%	94.80%
Mult	3134	In 80th percentile	97.38%	97.35%	96.65%

Figure Figure 1 shows the absolute percent errors associated with the total population for the ADDMULT model. 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

error map-1.pdf

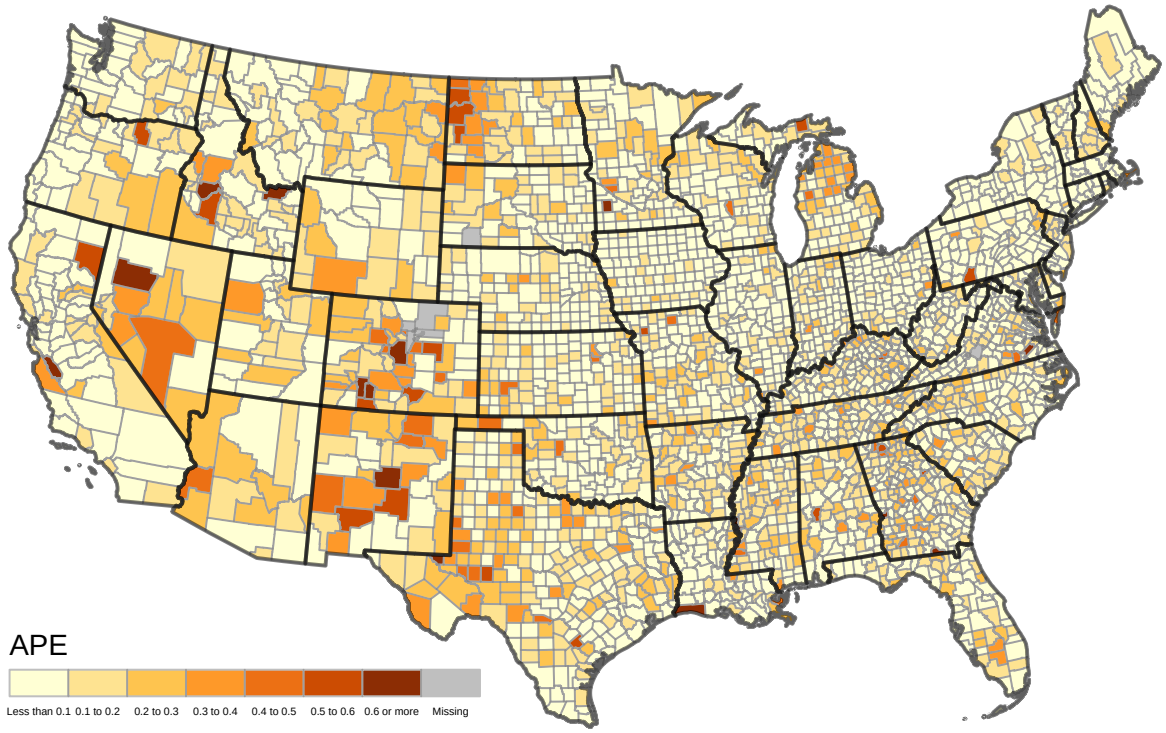

Figure 1: **Map of county errors of the total population in 2015 using the ADDMULT model.** Here I show the geographic distribution of absolute percent errors. Most states and counties have low error rates of the total population with isolated pockets of large errors.

population totals within the 80th percentile prediction interval with the ADDMULT 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

Table **3** 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.

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

TYPE	n	EVAL	2005	2010	2015
ADD	56412	Median APE	5.3550%	8.247%	11.602%
ADDMULT	56412	Median APE	5.2963%	8.014%	11.024%
Mult	56412	Median APE	5.3644%	8.152%	11.389%
ADD	56412	Median ALPE	0.982%	1.179%	3.389%
ADDMULT	56412	Median ALPE	1.119%	1.460%	3.697%
Mult	56412	Median ALPE	1.327%	1.687%	3.682%
ADD	56412	In 80th percentile	69.99%	81.79%	84.90%
ADDMULT	56412	In 80th percentile	68.82%	81.07%	84.68%
Mult	56412	In 80th percentile	73.88%	85.96%	89.74%

Figure 2 shows projected age structures in nine 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.

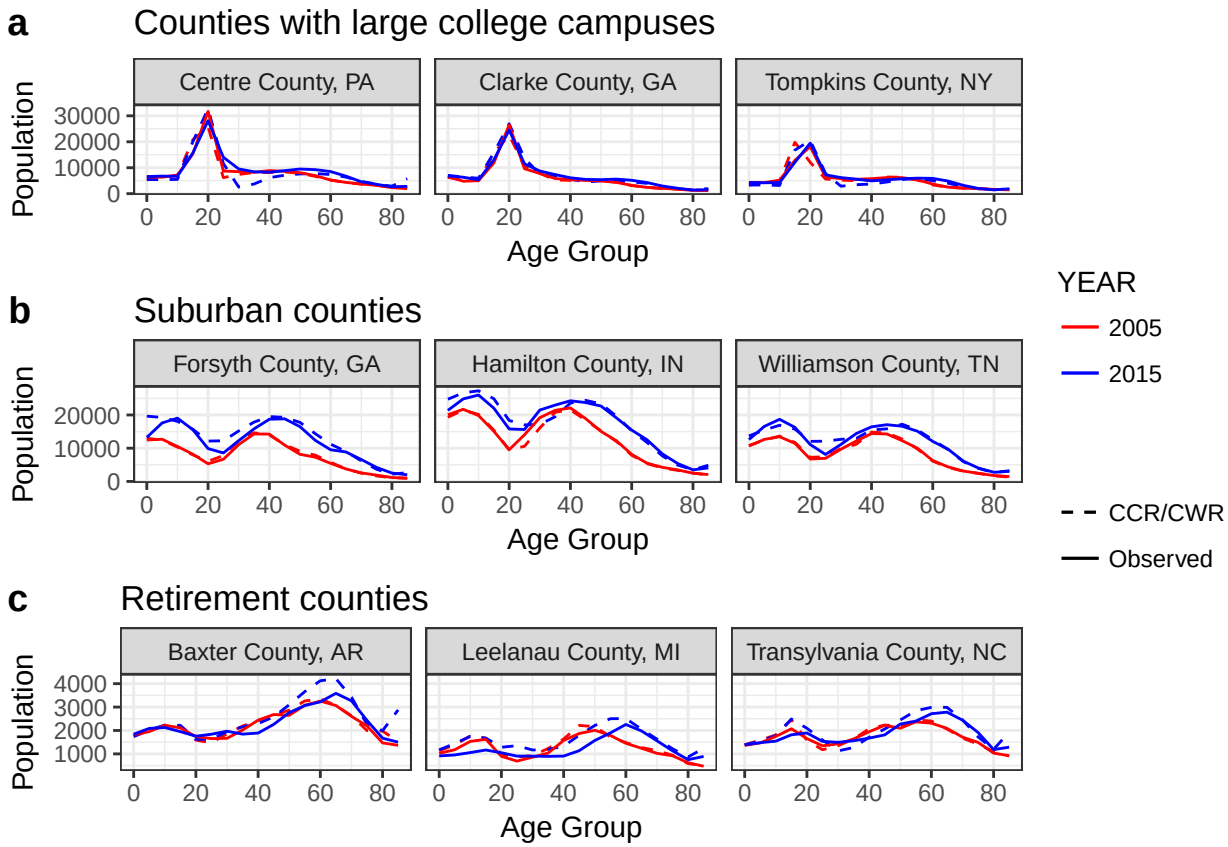

Figure 2: **Age structures of various county types.** I compare the projected age structures to the observed age structures in nine counties across three county types using the ADD/MULT model. (a) demonstrates counties with major universities, (b) demonstrates sample suburban counties, and (c) demonstrates sample retirement counties. All three county types have age structures largely preserved despite widely different age structures.

errors-1.pdf

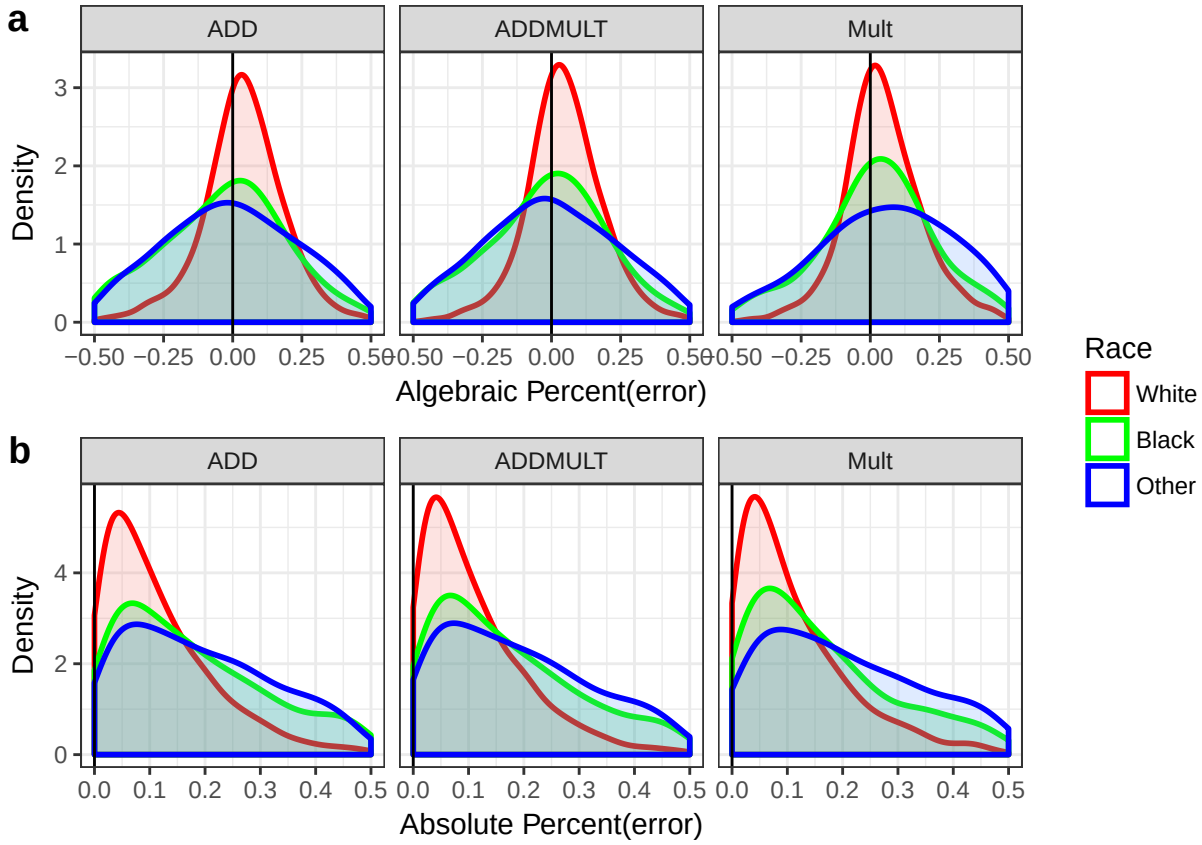

Figure 3: **Race group errors. (a) shows the Algebraic Percent Errors for all three methods and (b) shows the APE distribution of errors.**

Race Errors

Figure 3 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 proejctions, 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 to the confidence bounds being too wide when discussing the overal

total populations in counties, when examining any given ASRC combination it appeasrs 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
ADD	334872	Median SAPE	6.195%	8.709%	11.39%
ADDMULT	334872	Median SAPE	6.031%	8.466%	11.18%
Mult	334872	Median SAPE	6.145%	8.844%	12.64%
ADD	334872	In 80th percentile	49.00%	61.62%	66.41%
ADDMULT	334872	In 80th percentile	50.34%	62.11%	66.34%
Mult	334872	In 80th percentile	58.49%	70.04%	73.11%

PROJECTIONS

Projections will go here.

References