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 a trend component μ_t and a random error component ε_t and it can be described as:

$$y_t = \mu_t + \varepsilon_t \tag{6}$$

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. The trend component is modeled using the following equations:

$$\mu_t = \mu_{t-1} + \eta_t$$
$$\eta_t \sim N(0, \sigma_\eta^2)$$

These equations specify a trend where the level μ_t vary over time, governed by the variance of the disturbance terms η_t and ξ_t in their equations. Here all individual CCRs/CCDs (CCR_{iasr}) over all series are modelled (n=339,444) 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:

$$y = log\Big(\frac{x-a}{b-x}\Big) \tag{7}$$

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 8 describes the leslie matrices for CCRs and equation 9 describes the leslie matrices for CCDs.

$$\begin{bmatrix} n_0 \\ n_1 \\ \vdots \\ n_{18} \end{bmatrix}_{t+1} = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ CCR_0 & 0 & 0 & \dots & 0 & 0 \\ 0 & CCR_1 & 0 & \dots & 0 & 0 \\ 0 & 0 & CCR_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 & 0 \\ 0 & 0 & 0 & \dots & CCR_{16} & CCR_{17} \end{bmatrix} \cdot \begin{bmatrix} n_0 \\ n_1 \\ \vdots \\ n_{17} \end{bmatrix}_t \tag{8}$$

$$\mathbf{T} = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ CCD_0 & 0 & 0 & \dots & 0 & 0 \\ 0 & CCD_1 & 0 & \dots & 0 & 0 \\ 0 & 0 & CCD_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 & 0 \\ 0 & 0 & 0 & \dots & CCD_{16} & CCD_{17} \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ n_0 & 0 & 0 & \dots & 0 & 0 \\ 0 & n_1 & 0 & \dots & 0 & 0 \\ 0 & 0 & n_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 & 0 \\ 0 & 0 & 0 & \dots & n_{16} & n_{17} \end{bmatrix} \quad (9)$$

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

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 $t+1$.

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 $t+1$.

County boundary changes: Several county boundaries or names have also shifted since 1980 and I have accounted for these changes in both the historic time series and the projections.

- FIPS 02105 Hoonah-Angoon Census Area AK was created from FIPS 02105, 02230 Skagway Municipality, and 02232 Skagway-Hoonah-Angoon Census Area were all cre-

ated out of the same 02230 Skagway Municipality. FIPS 02230 was changed in 1992 from FIPS 02231.

- FIPS 02130 Ketchikan Gateway Borough AK, 02195 Petersburg Census Area, 02198 Prince of Wales- Hyder Census Area, 02201 Prince of Wales-Outer Ketchikan Census Area, 02275, and 02280 Wrangell-Petersburg Census Area were carved out of 02130 Ketchikan Gateway Borough.
- FIPS 02270 Wade Hampton, AK was recoded to FIPS 02158.
- FIPS 08014 Broomfield County CO was created out of parts of FIPS 08001 Adams, 08013 Boulder, 08059 Jefferson, and 08123 Weld. Over 90% of the created population came out of 08013 Boulder so it is remerged there.
- FIPS 12025 Dade County FL changed it's name to Miami-Dade FL to FIPS 12086.
- FIPS 15005 Kalawao County HI was absorbed into FIPS 15009 Maui County HI.
- FIPS 30113 Yellowstone National Park MT was split into FIPS 30031 Gallatin and FIPS 30067 Park. All three have been merged into 30031 Park MT – the larger county.
- FIPS 46113 Shannon SD was recoded to FIPS 46102.
- FIPS 51780 South Boston VA was merged into FIPS 51083 Halifax County VA.
- FIPS 51560 Clifton Forge VA was merged into 51005 Allegheny County VA.

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 3136 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.

Overall Errors

Table **??**tab:TOTALeval) 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	298,379,612	303,663,147	285,340,185	322,272,611	1.77%
ADD	2010	309,347,527	322,957,965	285,381,737	361,545,453	4.40%
ADD	2015	320,894,895	342,957,970	285,495,854	402,884,822	6.88%
ADDMULT	2005	298,373,465	304,674,178	286,505,559	323,211,601	2.11%
ADDMULT	2010	309,344,478	326,521,302	289,545,501	367,786,796	5.55%
ADDMULT	2015	320,890,305	353,678,081	292,224,571	444,354,649	10.22%
Mult	2005	298,379,612	310,264,056	288,517,257	332,581,349	4.0%
Mult	2010	309,341,025	348,226,543	298,099,857	411,654,046	12.6%

TYPE	YEAR	POPULATION	PRED	LOW	HIGH	APE
Mult	2015	320,885,858	421,672,155	313,302,023	630,712,200	31.4%

Table **??** 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 higher average projections 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	3136	Median APE	2.780%	6.17%	10.4%
ADDMULT	3136	Median APE	2.767%	6.18%	11.4%
Mult	3136	Median APE	3.553%	9.55%	20.6%
ADD	3136	Median ALPE	0.70%	2.02%	4.5%
ADDMULT	3136	Median ALPE	1.27%	3.74%	8.3%
Mult	3136	Median ALPE	2.44%	7.74%	18.6%
ADD	3136	In 80th percentile	91.87%	89.92%	87.4%
ADDMULT	3136	In 80th percentile	92.03%	89.76%	86.5%
Mult	3136	In 80th percentile	89.99%	83.51%	77.2%

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 11.4% by 2015, however some isolated pockets of high errors do exist randomly distributed throughout the United States. Additionally, 86.5% of counties had observed

error map-1.pdf

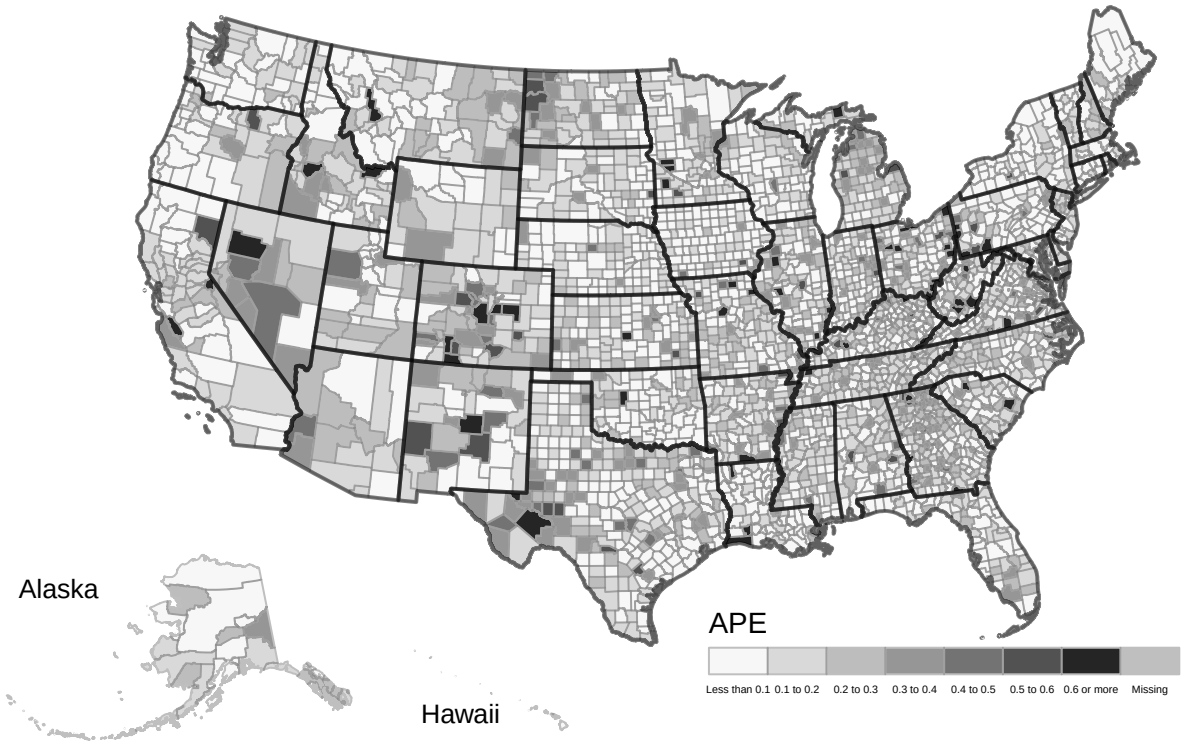

Figure 1: **Map of county errors of the total population in 2015 using the ADDMULT model.** Here I show the geographic distribution of absolute percent errors. Most states and counties have low error rates of the total population with isolated pockets of large errors.

population totals within the 80th percentile prediction interval with the ADDMULT model by 2015.

Age Structure Error

Table **??** reports the overall errors for age groups at the county level. Both the ADD and ADDMULT models produce similar APEs. For any given county, the average error is approximately 14%. Similar to the overall errors, the bias tends to be to over-project the populations as all of the ALPEs are positive.

Table 3: Evaluation of Age Group Errors.

TYPE	n	EVAL	2005	2010	2015
ADD	56448	Median APE	5.460%	9.46%	13.99%
ADDMULT	56448	Median APE	5.521%	9.40%	13.96%
Mult	56448	Median APE	6.135%	11.20%	18.09%
ADD	56448	Median ALPE	0.96%	2.91%	5.79%
ADDMULT	56448	Median ALPE	1.37%	3.93%	7.79%
Mult	56448	Median ALPE	2.47%	6.30%	12.39%
ADD	56448	In 80th percentile	63.53%	71.81%	73.93%
ADDMULT	56448	In 80th percentile	62.03%	70.56%	72.89%
Mult	56448	In 80th percentile	61.64%	69.85%	71.97%

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.

```
## Warning: Removed 10232 rows containing non-finite values (stat_density).
```

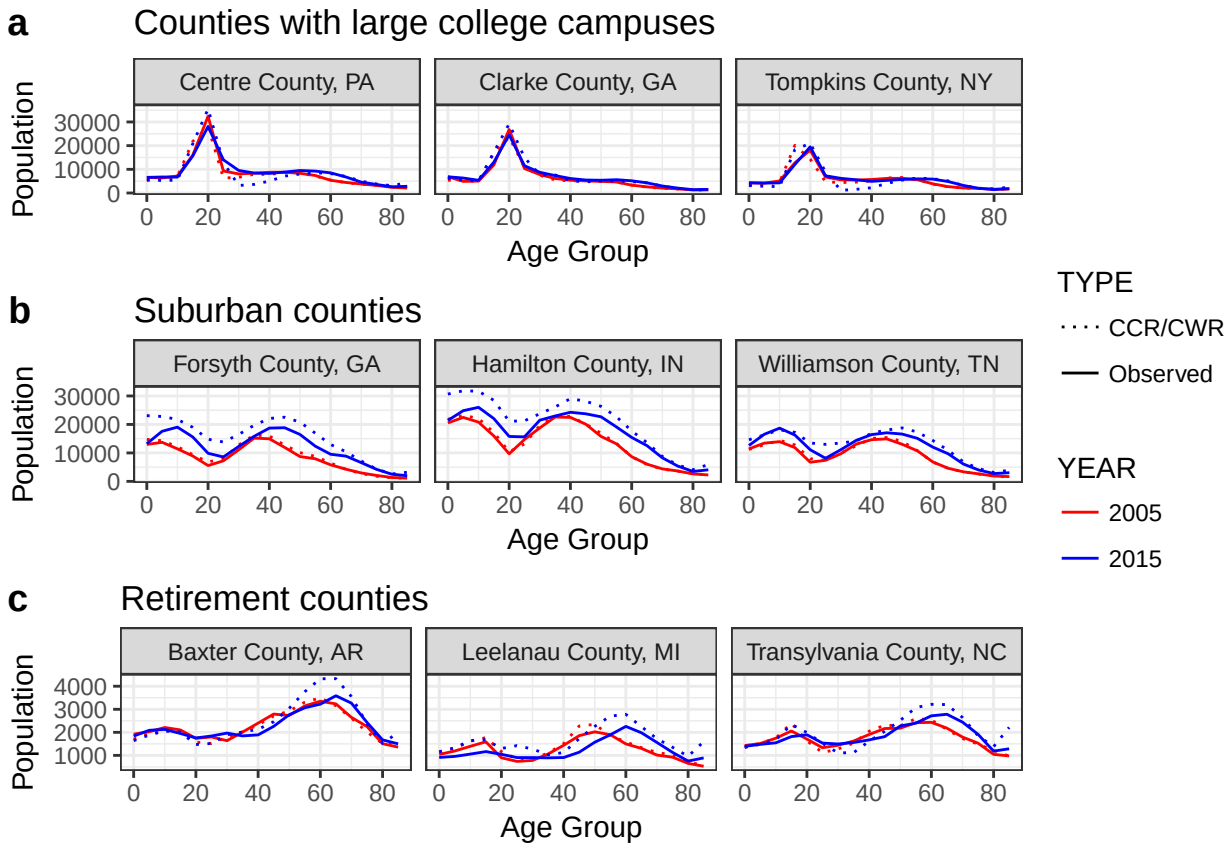

Figure 2: **Age structures of various county types.** I compare the projected age structures to the observed age structures in nine counties across three county types using the ADD/MULT model. (a) demonstrates counties with major universities, (b) demonstrates sample suburban counties, and (c) demonstrates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

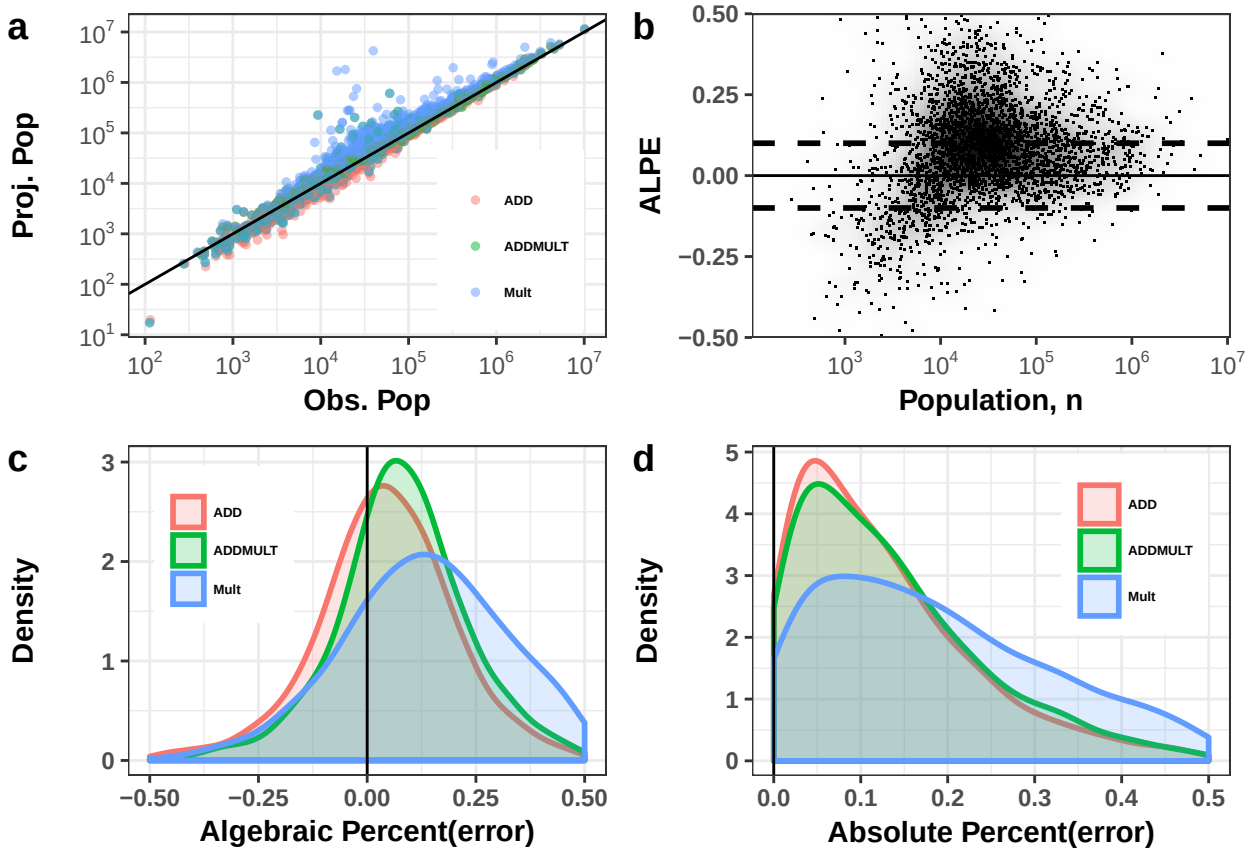

Figure 3: **Total population errors.** I compare the projected total popualtions to the observed total populations for the year 2015. (a) plots the observed populations against the predicted populations. (b) plots the density of algebraic percent errors for the ADDMULT method in 2015. The dashed lines represent +/- 10%. (c) plots the algebraic percent error for all three methods in 2015.(d) plots the absolute percent error for all three methods in 2015.

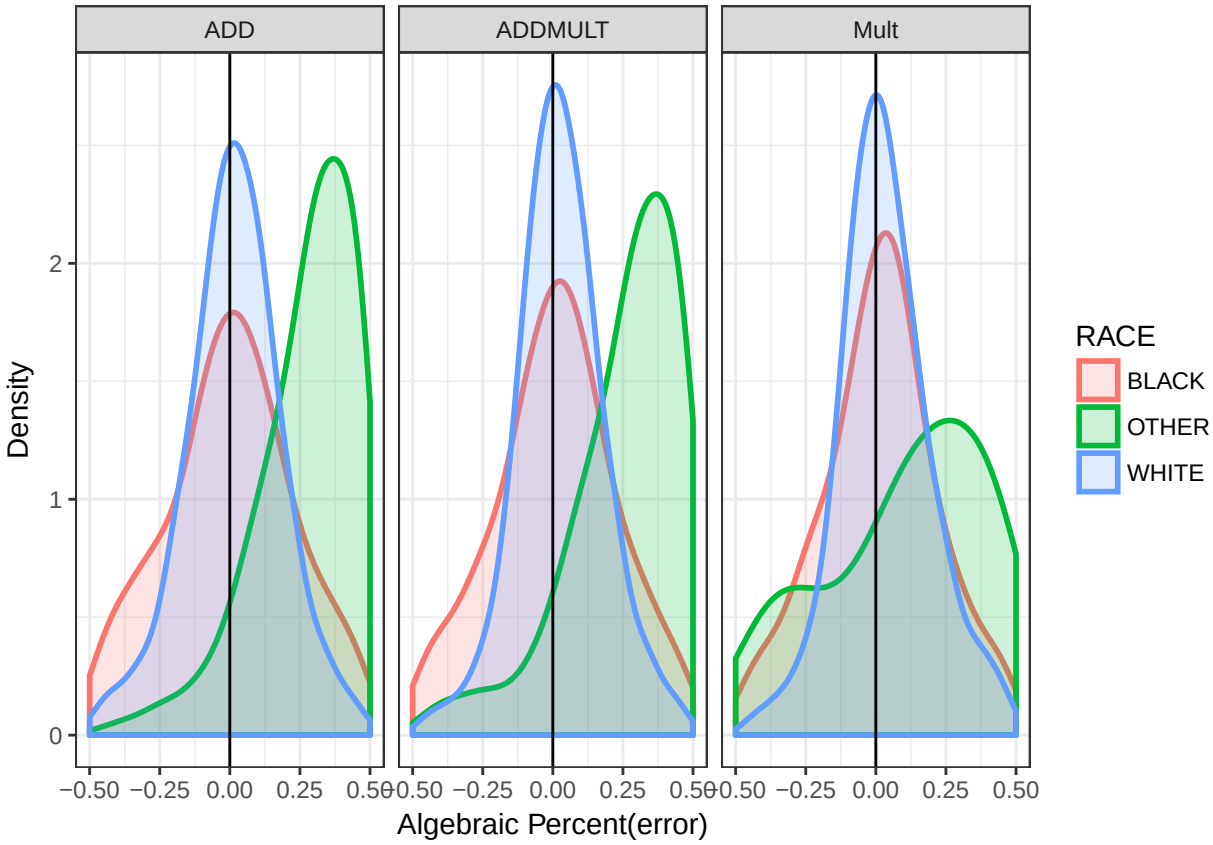

Figure 4: **Race group errors.**

Race error

Table

Table 4: Evaluation of Age Group Errors.

RACE	TYPE	n	EVAL	2005	2010	2015
BLACK	ADD	9408	Median APE	10.48%	17.31%	22.64%
BLACK	ADDMULT	9408	Median APE	10.93%	16.73%	22.47%
BLACK	Mult	9408	Median APE	12.78%	19.34%	27.90%
OTHER	ADD	9408	Median APE	25.7%	53.5%	70%
OTHER	ADDMULT	9408	Median APE	29.9%	62.5%	83%
OTHER	Mult	9408	Median APE	64.6%	145.8%	279%
WHITE	ADD	9408	Median APE	2.865%	6.613%	10.86%
WHITE	ADDMULT	9408	Median APE	2.639%	5.889%	9.73%
WHITE	Mult	9408	Median APE	2.850%	6.282%	10.20%
BLACK	ADD	9408	Median ALPE	1.49%	5.66%	3.92%
BLACK	ADDMULT	9408	Median ALPE	-0.37%	2.64%	1.49%
BLACK	Mult	9408	Median ALPE	-1.95%	-0.30%	-1.79%
OTHER	ADD	9408	Median ALPE	25.4%	53.5%	70%
OTHER	ADDMULT	9408	Median ALPE	23.7%	51.1%	69%
OTHER	Mult	9408	Median ALPE	47.7%	145.8%	279%
WHITE	ADD	9408	Median ALPE	-0.072%	0.08%	1.57%
WHITE	ADDMULT	9408	Median ALPE	0.268%	0.68%	2.38%
WHITE	Mult	9408	Median ALPE	0.749%	1.26%	2.61%
BLACK	ADD	9408	In 80th percentile	76.08%	77.89%	81.75%
BLACK	ADDMULT	9408	In 80th percentile	70.97%	76.36%	79.05%
BLACK	Mult	9408	In 80th percentile	67.98%	75.13%	74.68%

RACE	TYPE	n	EVAL	2005	2010	2015
OTHER	ADD	9408	In 80th percentile	38.6%	29.98%	31.90%
OTHER	ADDMULT	9408	In 80th percentile	36.7%	30.03%	30.22%
OTHER	Mult	9408	In 80th percentile	26.9%	24.96%	25.41%
WHITE	ADD	9408	In 80th percentile	89.349%	86.252%	83.764%
WHITE	ADDMULT	9408	In 80th percentile	89.442%	86.507%	83.477%
WHITE	Mult	9408	In 80th percentile	88.680%	85.933%	83.700%

Age, Sex, Race joint errors

References