U.S. County level probabilistic population projections by age, sex, and race

Mathew E. Hauer

April 25, 2018

County-level Population Projections

- No rigourous set of U.S. sub-national projections by age, sex, and race presently exists.
- 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.
- Present just the projection methodology, South Dakota keeps tripping me up.

Cohort-Component

Very familiar with the Demographic Accounting Equation:

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

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**.

Cohort-Component Problems

- There is no comprehensive dataset of:
  - Gross migration estimates by age, sex, and race for all U.S. counties.
  - Birth data are suppressed by NCHS for $<$100k pop.
  - Death data are only available for cells with more than 10 deaths.
- Birth/Death data must come from uneven state-level vital reporting agencies.

A universal county-level population projection is difficult, if not impossible, to complete using publicly available datasets.

CCRs or Hamilton-Perry

Hamilton-Perry or Cohort-Change Ratios offers a more parsimonious solution

$$CCR_t = \frac{{}_nP_{x,t}}{{}_nP_{x-y,t-1}} \quad (2)$$

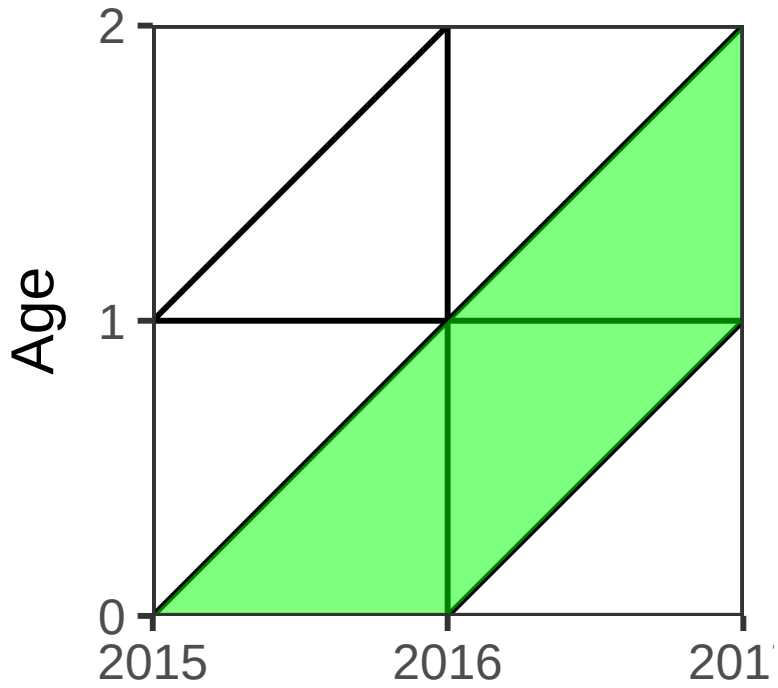
$${}_nP_{x+t} = CCR_t \cdot {}_nP_{x-y,t} \quad (3)$$

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

CCR Advantages

CCRs offer several advantages over the use of cohort-component:

1. Very parsimonious.
2. Low data requirements.
   - Just age-sex distributions at two time periods.
3. Easy to implement.
4. Low barrier of entry.



CCR Disadvantages

CCRs suffer from one major disadvantage over the use of cohort-component: 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):

- 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.

CCR Disadvantages

This problem of impossible over-projection has lead to general "guidelines" surrounding CCRs.

1. Projection horizons should typically be small, typically 10- to 20-years.
2. Dimensionality should typically be limited due to the possibility of massive ratios (ie, 2 -> 4 persons)

I *think* these problems can be resolved using a slight change to the CCR formulation.

CCRs? Try CCDs

The implementation of CCRs naturally implies a multiplicative model.

However it is possible to implement an **additive** model by using the *difference* in population rather than the *ratio* of population.

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

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 Advantages/Disadvantages

CCDs are just as parimonious as CCRs but have the additional advantage of producing *linear* growth rather than *exponential* growth.

However, CCDs have the potential of creating impossible negative populations through linear decline.

A blended approach using CCDs in areas projected to grow and CCRs in areas projected to decline would rectify the possibility of negative populations.

Projecting CCRs and CCDs

I employ the use of Unobserved Components Models (UCMs) to forecast CCRs/CCDs.

UCMS are a dynamic time series forecasting and in the family of Bayesian Structural Time Series models.

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.

Very easy to implement! library(RUCM) is all you need!

Projecting CWRs

I projected the CWRs within a constrained forecast interval.

CWRs are constrained to lie between (α, β), where $\alpha = 0.14$ ($TFR \approx 1$) and β is the maximum CWR observed in the time series.

I then transform the data using a scaled logit transformation to map (α, β) to the whole real line:

$$y = log(\frac{x - \alpha}{\beta - x}) \tag{5}$$

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

Leslie Matrices CCRs

Simple Leslie Matrix for CCRs.

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

Leslie Matrices CCDs

$$\mathbf{T} = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ D_0 & 0 & 0 & \dots & 0 & 0 \\ 0 & D_1 & 0 & \dots & 0 & 0 \\ 0 & 0 & D_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 & 0 \\ 0 & 0 & 0 & \dots & D_{16} & D_{17} \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ n_0 & 0 & 0 & \dots & 0 & 0 \\ 0 & n_1 & 0 & \dots & 0 & 0 \\ 0 & 0 & n_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 & 0 \\ 0 & 0 & 0 & \dots & n_{16} & n_{17} \end{bmatrix}$$

$$P_{t+1} \equiv \begin{bmatrix} \sum_{i=1}^{n} \mathbf{T_{ij}} \\ \sum_{i=1}^{n} \mathbf{T_{ij}} \\ \vdots \\ \sum_{i=1}^{n} \mathbf{T_{ij}} \end{bmatrix}$$

DATA

A single primary data source: the National Vital Statistics System U.S. Census Populations with Bridged Race Categories data set.

- All Racial classifications are harmonized across space and time.
- All county boundaries have been rectified.

I use the 1969-2016 datafile that utilizes three race groups (White, Black, and Other), 18 age groups, and 2 sex groups.

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 $t + 1$.

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**).

Results: Overall

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

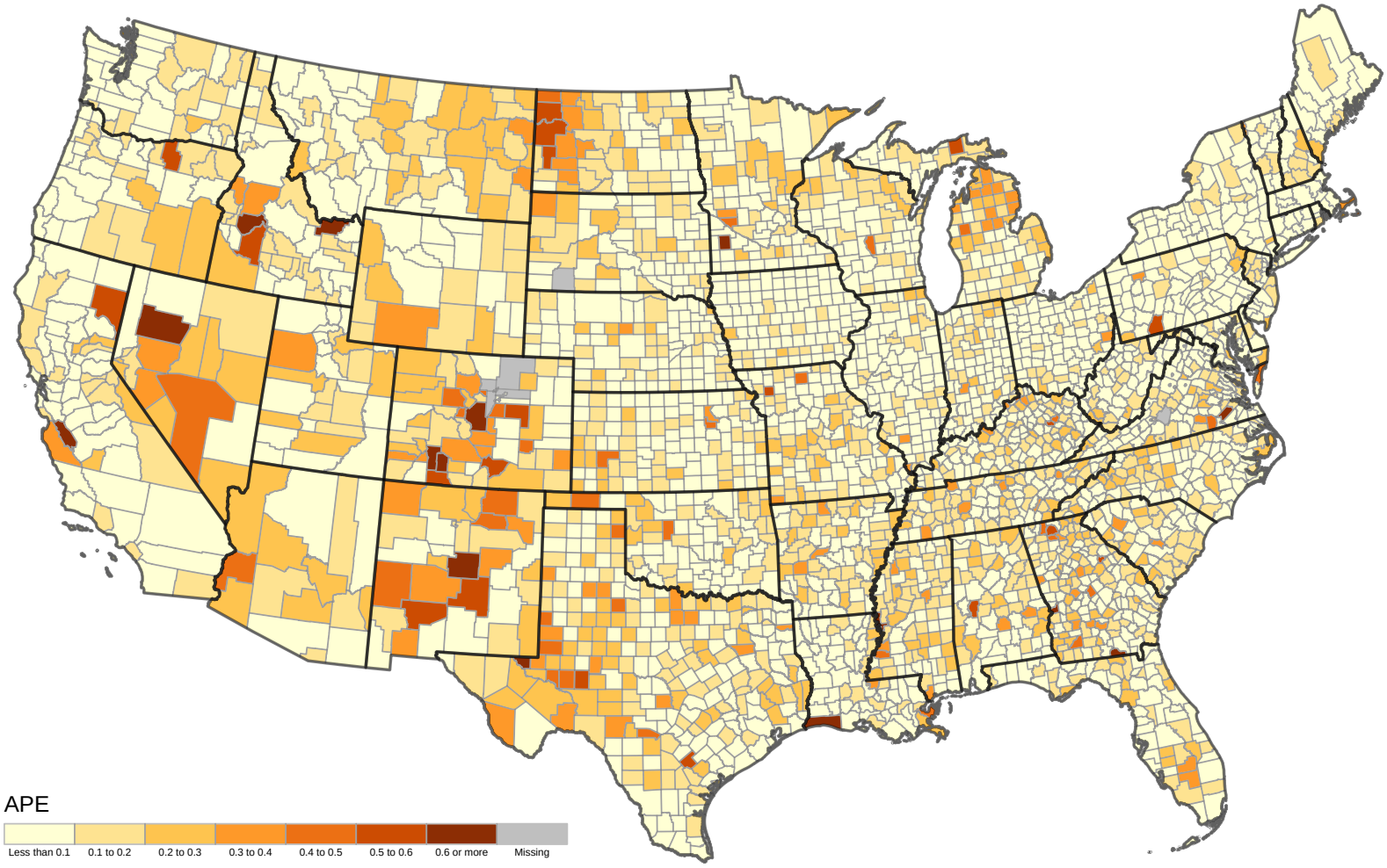
TYPE	YEAR	POPULATION	PRED	APE
ADD	2005	292,554,817	297,219,886	1.59%
ADD	2010	306,396,740	313,983,491	2.48%
ADD	2015	317,729,565	331,173,497	4.23%
ADDMULT	2005	292,554,817	297,516,199	1.70%
ADDMULT	2010	306,396,740	314,705,975	2.71%
ADDMULT	2015	317,729,565	332,461,792	4.64%
Mult	2005	292,554,817	299,142,100	2.25%
Mult	2010	306,396,740	320,917,411	4.74%
Mult	2015	317,729,565	351,252,666	10.55%

Results: Counties

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%

Results: Counties

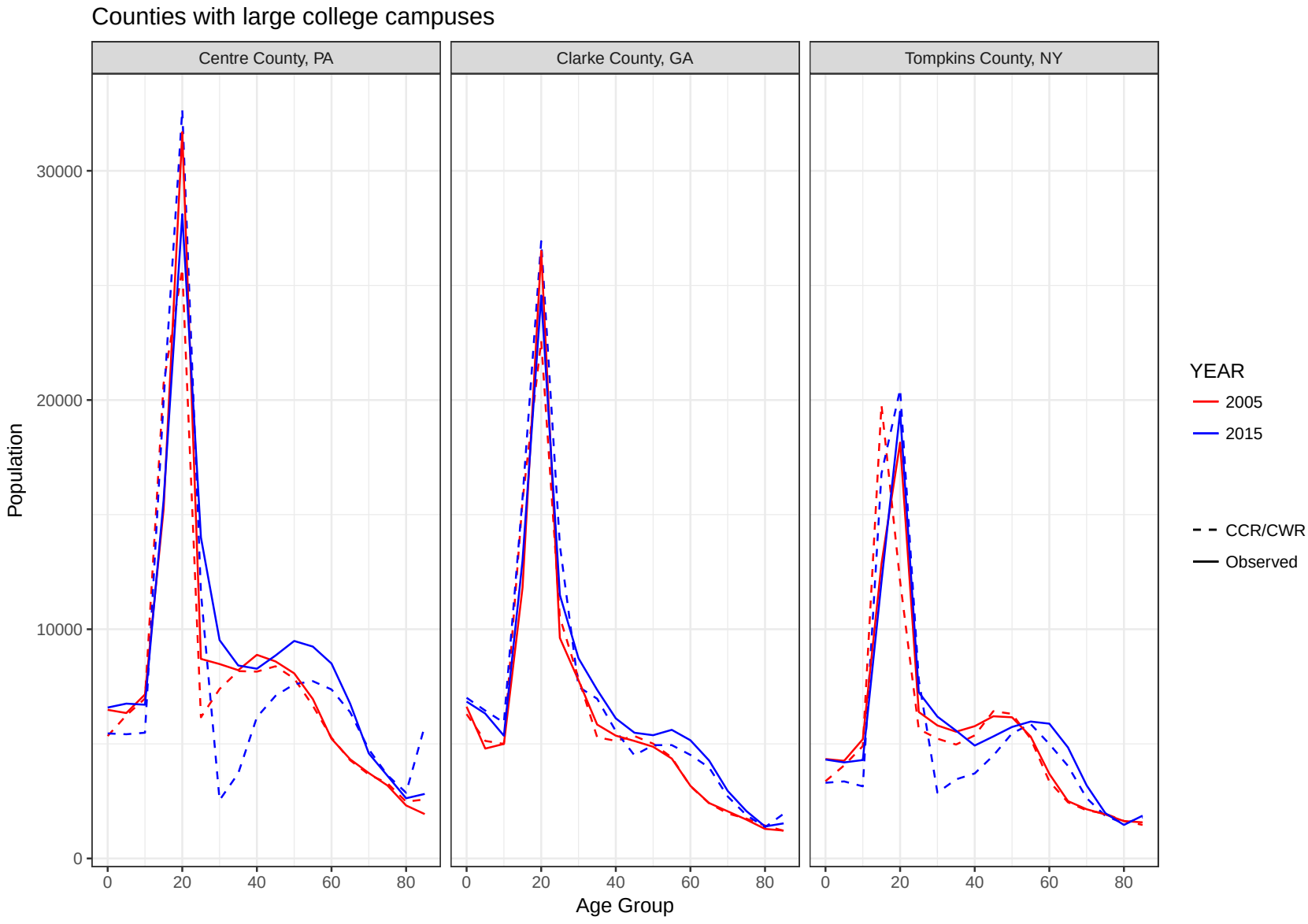


Age Structure Error

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%

Results: Age Structures: College Counties



S

This is an R Markdown presentation. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document.

Slide with Bullets

- Bullet 1
- Bullet 2
- Bullet 3

Slide with R Output

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Slide with Plot

```
plot(pressure)
```