Evaluation of Population Projection Errors

Mathew E. Hauer

February 27, 2018

Overview

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 national coverage. Birth and death data are typically obtained through the National Center of Health Statistics (NCHS) vital events registration databases. These data, however, are only available for counties with populations greater than 100k and are suppressed in populations with fewer than 1k (I think) members rendering 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, which uses cohort-change ratios (CCRs) in place of components to project populations. The basic CCR equation is found in equation 2.

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

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 calculate the CCR for the open-ended age group,

$$\begin{aligned} {}_\infty CCR_{85,t} &= \frac{{}_\infty P_{85,t}}{{}_\infty P_{85-y,t-1}} \\ {}_\infty P_{85+t} &= {}_\infty CCR_{85,t} \cdot {}_\infty P_{85-y,t} \end{aligned} \tag{3}$$

Where y is the time difference between time periods.

For the population aged 0-4, we use the ratio of the population aged 0-4 to the number of women of reproductive age. Here we define women of reproductive age as the ages $[15, 50)$.

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) driving 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 population 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.

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

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 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).

Projecting CCRs and CCDs

It is unlikely that CCRs will remain unchanged over the projection horizon. To account for possible changes in CCRs, I employed 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:

$$y_t = \mu_t + \varepsilon_t \tag{5}$$

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:

$$\begin{aligned}\mu_t &= \mu_{t-1} + \beta_{t-1} + \eta_t \\ \beta_t &= \beta_{t-1} + \xi_t \\ \eta_t &\sim N(0, \sigma_\eta^2) \\ \xi_t &\sim N(0, \sigma_\xi^2)\end{aligned}$$

These equations specify a trend where the level μ_t and the slope β_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 the series were modelled (n=339,444) in individual UCM models.

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.

We forecast these UCMs for each CWR within a constrained forecast interval. CWRs are constrained to lie between (a, b). We limited CWRs such that each age/race/county combination would be constrained within the maximum/minimum of the time series such that $a = 0$ for all projections. and $b = max(CWR_{arc})$. We then transform the data using a scaled logit transformation to map (a, b) to the whole real line

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

Where x is the original data and y is the transformed data. The prediction intervals from these transformations have the same coverage probability as on the transformed scale, because quantiles are preserved under monotonically increasing transformations.

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

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

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

$$P_{t+1} \equiv \begin{bmatrix} \sum \mathbf{T_{1i}} \\ \sum \mathbf{T_{2i}} \\ \vdots \\ \sum \mathbf{T_{17i}} \end{bmatrix}$$

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

Extra considerations

These projections were carried out with 18 age groups (0,85,5), 2 sex groups, and 3 race groups (White, Black, Other).

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 $t + 1$.

Several county boundaries have also shifted since 1980:

- FIPS 12025 was changed to 12086.
- FIPS 15005 was absorbed by FIPS 15009.
- FIPS 51780 was merged into 51083.
- FIPS 51560 was merged into 51005.
- FIPS 30113 was split into 30031 and 30067. All three have been merged into 30031 – the larger county.
- FIPS 08014 was created out of parts of 08013, 08123, 08001, and 08059. Over 90% of the created population came out of 08013 so it is remerged.
- FIPS 02105 was created from 02105, 02230, and 02232 were all created out of the same 02230. 02230 was changed in 1992 from 02231.
- FIPS 02130, 02195, 02198, 02201, 02275, and 02280 were carved out of 02130.
- FIPS 02270 was recoded to 02158.
- FIPS 46113 was recoded to 46102.

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. Additionally, any infinite, NA, or NAN CCR, CCD, or CWR was set to 0.

States included in this analysis: AL, AK, AZ, AR, CA, CO, CT, DE, DC, FL, GA, HI, ID, IL, IN, IA, KS, KY, LA, ME, MD, MA, MI, MN, MS, MO, MT, NE, NV, NH, NJ, NM, NY, NC, ND, OH, OK, OR, PA, RI, SC, SD, TN, TX, UT, VT

Total number of counties: 2813

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 could produce more accurate results compared to CCRs.

Table 1: Evaluation of TOTAL Errors. MAPE refers to MEDIAN Absolute Percent Error

TYPE	YEAR	POPULATION	PRED	LOW	HIGH	MAPE
ADD	2005	276,406,913	281,533,136	264,654,992	298,642,664	1.85%
ADD	2010	286,469,560	299,534,542	264,726,511	335,047,364	4.56%
ADD	2015	297,171,625	318,089,664	264,264,490	373,347,568	7.04%
Mult	2005	276,406,913	287,629,227	267,385,087	308,167,323	4.1%
Mult	2010	286,463,331	322,367,229	276,082,191	380,277,495	12.5%
Mult	2015	297,163,134	386,119,499	277,605,581	570,420,926	29.9%

The total error for any given county is also small and only marginally larger than the nationwide total.

Table 2: Evaluation of TOTAL Errors for counties. MAPE refers to MEDIAN Absolute Percent Error

COUNTYnum	TYPE	VAR	2005	2010	2015
2813	ADD	MAPE	2.78%	6.22%	10.5%
2813	Mult	MAPE	3.61%	9.70%	21.0%
2813	ADD	in 80th percentile	91.86%	89.94%	87.20%
2813	Mult	in 80th percentile	90.33%	83.43%	77.89%

Errors by Age

The errors for age groups are also relatively low with the average age group having an overall error of 13%.

Table 3: Evaluation of Age Group Errors. MAPE refers to MEDIAN Absolute Percent Error

num	TYPE	VAR	2005	2010	2015
50634	ADD	MAPE	0.0549	0.0953	0.1417
50634	Mult	MAPE	0.0619	0.1134	0.1841
50634	ADD	in 80th percentile	0.6367	0.7198	0.7390
50634	Mult	in 80th percentile	0.6173	0.6984	0.7197

Errors by Sex

Table 4: Evaluation of Sex Errors. MAPE refers to MEDIAN Absolute Percent Error

num	SEX	TYPE	YEAR	MAPE	in80percentile
2813	FEMALE	ADD	2005	2.641%	92.53%
2813	FEMALE	Mult	2005	3.358%	90.72%
2813	MALE	ADD	2005	3.040%	89.87%
2813	MALE	Mult	2005	4.005%	88.52%
2813	FEMALE	ADD	2010	5.85%	90.72%
2813	FEMALE	Mult	2010	8.97%	84.57%
2813	MALE	ADD	2010	6.79%	87.84%
2813	MALE	Mult	2010	10.43%	81.66%
2813	FEMALE	ADD	2015	9.71%	88.20%
2813	FEMALE	Mult	2015	19.16%	78.49%
2813	MALE	ADD	2015	11.3%	85.35%
2813	MALE	Mult	2015	21.5%	76.68%

Errors by Race

Table 5: Evaluation of Race Errors. MAPE refers to MEDIAN Absolute Percent Error

num	RACE	TYPE	YEAR	MAPE	in80percentile
2813	BLACK	ADD	2005	10.67%	79.4%

num	RACE	TYPE	YEAR	MAPE	in80percentile
2813	BLACK	ADD	2010	17.96%	82.55%
2813	BLACK	ADD	2015	23.15%	84.18%
2813	BLACK	Mult	2005	13.07%	67.5%
2541	BLACK	Mult	2010	19.11%	75.56%
2541	BLACK	Mult	2015	27.86%	75.13%
2813	OTHER	ADD	2005	26.0%	40.42%
2813	OTHER	ADD	2010	54.1%	31.92%
2813	OTHER	ADD	2015	71%	34.23%
2813	OTHER	Mult	2005	65.6%	30.75%
2788	OTHER	Mult	2010	144.9%	29.77%
2788	OTHER	Mult	2015	278%	30.09%
2813	WHITE	ADD	2005	2.8587%	89.406%
2813	WHITE	ADD	2010	6.706%	86.314%
2813	WHITE	ADD	2015	11.034%	83.576%
2813	WHITE	Mult	2005	2.8775%	88.660%
2813	WHITE	Mult	2010	6.351%	85.674%
2813	WHITE	Mult	2015	10.364%	83.221%

Errors for all joint combinations

Table 6: Evaluation of Age/Sex/Race Errors. MAPE refers to MEDIAN Absolute Percent Error

num	TYPE	YEAR	MAPE	in80percentile
303210	ADD	2005	14.285%	44.256%
303210	Mult	2005	15.131%	44.898%
302616	ADD	2010	23.36%	52.835%
293112	Mult	2010	25.32%	52.335%
302022	ADD	2015	32.48%	55.81%
293112	Mult	2015	37.82%	54.41%