Modeling methodology for the 2016 baseline California population projections.

Ethan Sharygin

Demographic Research Unit, California State Department of Finance[*]

January 20, 2018

Release notes.

2018.1.20	Updated projections to accord to county population totals for July 2017. Consistency is imposed reweighting the age 0 population (by the difference between projected and estimated FY2016 births) and then reweighting the age 25 and older population. The difference between estimated and projected total population by county as of July 2017 is used to reweight total population in all future periods. Reassigned race/ethnicity of projected births using new algorithm and data [5]. Corrected undercount of population age 15-17 in households and overcount of total population age 18-19 (an error introduced in the previous revision).
2017.6.20	Corrected for miscount of population age 15-19 in group quarters: overcount of population 15-17 and undercount of population age 18-19.
2017.2.2	Initial public data release.

Summary.

The California Department of Finance (DOF) produces population projections for the state and counties of California on a regular basis. The Demographic Research Unit is responsible by statue for maintaining up-to-date postcensal population estimates and projections, which are both calculated using the identity known as the demographic balancing equation:

$$N_{t+1}=N_t+\left(B_{t,t+1}\ -D_{t,t+1}\ \right)+\left(I_{t,t+1}\text{-}O_{t,t+1}\right)$$

This identity decomposes the population in the next year into the population at the start of the current year, plus births (*B*) and less deaths (*D*) that occurred during the current year, plus migration in (*I*) and less migration out (*O*). The births, deaths, and migration anticipated during the current year are called the components of change. To generate these components at the county level, different approaches are used for births, deaths, and migrants. Birth and death counts are first converted to birth and death rates using population estimates; then, the log rates are modeled, forecast, and converted back into counts. Estimated net migration flows to and from each county are modeled and forecast from the historical net migration data as counts.

The 2016 baseline projections incorporate the latest population and birth, death, and migration estimates as of July 1, 2016 (i.e., through the end of the 2015 fiscal year). County populations by age, sex, and race/ethnicity are projected to 2060— 50 years ahead from the last Census. The following sections summarize the methods used for each of the components, introduce the data used to define the starting population, define the algorithms for combining the components, and explain the public use data products derived from the projections.

1. Estimation and projection of vital rates (births and deaths).

The State of California has kept records of the deaths and births within its borders since 1905, although coverage was incomplete until 1906 and 1919, respectively [4]. At present, records of births and deaths, including selected characteristics of children, parents, and the deceased, are obtained under an agreement with the California Department of Public Health (CDPH). Mortality and fertility rates are calculated by the number of events (deaths or births) during the year divided by the person-years of exposure (approximated by the population at the midpoint of the year, notated *N*).

[*] 915 L Street, Sacramento, CA 95814, tel.: (916) 323-4086. Email: pop.projections@dof.ca.gov.

The calculated rates are specific for each sex and age group in each county. In many counties, the rates thus estimated are unstable or undefined, either due to volatility in the events or population counts, or due to zero population counts for some cells. To address these issues, we first fit the following Poisson model for mortality:

$$ln\,(\mu_{ijt}) = ln\,(N_{ijt}) + (\beta_1 + \eta_j) T_t + BX_{it} + \rho_i^{geo} + \eta_j + \eta_i + \varepsilon_{ijt}\ ,$$

where subscript *i* refers to a single county and *t* to a single year. The subscript *j* refers to a single demographic group, which is a unique combination of age and sex. Mortality is shown in this example, but the fertility model (or equation) is very similar. The Poisson model above defines the log mortality rate μ as a function of the log population (N_{ijt}) plus a fixed coefficient (β_1) on time (T_t) and a county-level random coefficient (η_j) on time, a vector of fixed effects *B* on county-year specific covariates *X*,[*] a spatial residual ρ, a county and group-specific random intercepts η_i and η_j, and an error term ε. The spatial residual ρ is calculated by initially running the model with all other terms, predicting the county level random intercept, and taking the mean of those predicted intercepts for the neighboring counties. This model is based on the approach by Kulkarni et al. for the U.S. as a whole [9], except for the use of different covariates and the exclusion of data from counties outside California. Mortality to age 100 is estimated empirically, and mortality rates above age 100 are generated by an extrapolation of the mortality rate above age 30 using the following logit specification [8]:

$$logit\ \left(\mu_{ist}\right) = ln(\alpha) + \beta x,$$

where *x* is a continuous age vector, *s* is an id variable indicating sex, and α and β are the parameters to be estimated.

Final results are obtained by combining the results of independent model run for each calendar year 1990-2015, in moving panels of up to 5 years of available historical data and 1 year of data from the following year are included in each regression. Because the data series begin in 1990, the model for 1990 cannot look further back, and relies on data from only 1990 and 1991. The model for 1991 uses 1990-1992; the 1992 model uses data from 1990-1993, and so on, with the final model for 2015 using data from 2010-2015. An advantage of this method is the immutability of past results: when new data are added for calendar year 2016, the results for 2014 and earlier will not change. This method predicted the observed number of events (births and deaths) with great precision. During 2010-2015, the mean absolute percentage error (MAPE) of the mortality model was 0.915 percent; of the fertility model, 0.868. In other words, on average, the models were less than 1 percent off from the observed number of events recorded statewide.

Another advantage of this method is that it produces multiple sets of coefficients derived from the observed relationships over different time periods, rather than a single average relationship over the entire time period. There are many methods for generating mortality or fertility forecasts. The approach chosen in these population projections is a meta-forecast: coefficients from each model are used to generate a set of predictions after each regression, and the final set of all regression results is collapsed into percentiles. By default, the median result is used as the forecast rate. During the projections review process, decisions are made that may override the median projected rate in favor of higher or lower rates, e.g. to achieve a county-level target for the total fertility rate (TFR) or life expectancy at birth ($\dot{e}_0$).

2. Estimation and projection of migration.

Migration projections are based on the estimated county net migration totals from the July 1 components of change published by DOF (E2 and E6 series). These counts are converted into a crude rate of net migration (CRNM). Each county-level migration vector is fitted to four time series forecast models: three Box-Jenkins models (mean, naive, and drift specifications), and also a reversion model in which the rate of migration

[*] The county level covariates that are used for fertility and mortality have no research significance, since the model is not causal. Rather, covariates are used to aid in the estimation of rates for counties which have few events.

returns from its last estimated level to its long-run mean by the end of the forecast. In a final step, the four models are combined to produce a single central projection for net migration. The mean of the models is used as the default, and adjustments are made during review that override the default in cases deemed appropriate or necessary. DOF evaluates the initial projections in consultation with county planning agencies, Councils of Government, other affiliates of the State Census Data Center, and other members of the DOF demographic research network. Reviewers contribute independent assessments of future migration and notable developments or expected developments within their jurisdictions. When such input is not available, the migration model results are evaluated and revised internally by DOF.

The final net migration series is separated into net migration by age using household population records from the American Community Survey (ACS) Public Use Microdata Sample (PUMS). Annual gross migrants by age to and from each county are tallied using the variables for age and place of residence last year (excluding the population that moved in to GQ in California).[*] Net migrants are calculated for each five-year age group (0-4, 5-9, etc. to 75+). The ACS data are used to define an age pattern of migration that is used to divide net migration into net in- or out-migrants by age. The result each year is that the net migration figure is converted into a vector with 16 age categories and either a negative or positive number corresponding to the net migration in each of the 16 five-year age groups (the assignment of traits to migrants is discussed in further detail in the section, 'Method of population projection').

3. Special populations.

Special treatment is required for the population living in group quarters (GQ), including prisons, dormitories, military barracks, residential hospitals or nursing homes, monasteries, and other group accommodations. In the 2010 Census, this included 819,816 persons. These populations are not subject to the same mortality, fertility, or migration hazards as those living in households. DOF tracks changes in the size of these populations each year on a per-facility basis, but does not maintain data on changes in population characteristics. To address this gap, the 2010 Census SF2 file is used, which includes a breakdown of the population by age, race/ethnicity, and sex for each county GQ population. For each year during 2010-2016, the size of the GQ population is expanded or contracted by reweighting records to accord to the DOF-estimated total GQ population at that date.[†] After 2016, the GQ population is held constant at 2016 levels.

University students are a special case, because many do not reside in group quarters (dormitories) despite exhibiting similar population dynamics to the GQ population (i.e., these populations maintain a stable age structure, as outgoing students are generally replaced by incoming students). For California State University (CSU) and University of California (UC) campuses, an additional number of population records are set aside. The number of records set aside is equal to the difference between the published enrollment records for each CSU and UC campus for fall of the academic year and the DOF-estimated population in GQ for each campus. Published data and estimates are used for 2010-2016, and the size of the set-aside population is held constant in years after 2016. The set-aside records are treated the same as special population records for the purposes of the population projection; e.g., in order to simulate the dynamic of replenishment through graduation and new enrollment, they are not aged forward.

4. Starting population.

The basis of the DOF population projection series is the most recent decennial census of population, which as of 2016 is the April 1, 2010 population count of 37,253,956 for California. To produce the base population for initializing the projections, three Census Bureau datasets are used, and one DOF dataset. The 2010 Census Summary Files 1 and 2 (SF1 and SF2) are used to identify the population by age, sex, race/ethnicity, and GQ status (in a GQ or household). In the 2010 Census, the Census Bureau definitions and enumerations of group quarters (GQ) differed slightly from those used by the DOF in its published population estimates. Therefore,

[*] The smallest geography available in the ACS PUMS is the Public Use Microdata Area (PUMA), representing an area with a minimum population of 100,000. Counties with populations failing to meet the threshold of a PUMA were assigned a proportion of the PUMA's migrants equal to their proportion of the PUMA's population from the 2010 Census.

[†] The GQ population for each facility as of July 1 is used when available from DRU estimates; for many facilities, only January 1 estimates are recorded and these are used when July 1 estimates are unavailable.

the GQ population is resized to accord to DOF estimates. The data are then adjusted using the Modified Race Summary File (MRSF), which contains additional detail on the multiracial population. The MRSF is a table with totals of the population of each U.S. county who self-identified as Hispanic or not and as one of 31 single or multiple races in the 2010 Census.

Thus, the starting dataset for the population projection is a modified version of the SF1 that is re-weighted to match the county-level distribution of population by age, sex, GQ status, and race/ethnicity (according to the definitions reflected in the MRSF).

As an additional step, the full 31 race categories in the MRSF under the 1997 Office of Management and Budget (OMB) guidelines are collapsed into the 'bridged race' summary measure under the 1977 OMB guidelines that specify four race categories [7]. This step is necessary to preserve continuity with data collected in California before the year 2000, as well as to minimize mismatch between race as recorded on death certificates (where the determination may be made by a coroner, physician, or family member) and as recorded in a Census (where race is recoded usually by the respondent or a parent).

5. Method of population projection.

The population projection is run using the logic of the cohort component method [15], implemented as a discrete time microsimulation. In a traditional cohort component projection, the components of change would be calculated by the rate multiplied by the person-years of exposure, usually approximated by the mid-year population. A microsimulation adds an element of stochasticity, interpreting the estimated mortality or fertility rates as a probability of transition between states of nature (from alive to dead, no birth to birth, or staying to migrating). It allows for different outcomes with each projection iteration, reflecting the inherent uncertainty of life events (this logic can also be applied to population-level life tables [11]).

In a first step, the starting dataset is projected from April 1, 2010 to July 1, 2016. DOF projections employ birth and death vital registration records (VR) together with fiscal year net migration by county generated by the population estimates team. Missing data in vital records (county of residence, age, education, etc.) are handled by a Multiple Imputation by Chained Equations (MICE) model [1]. Births, deaths, migration, and changes in the size of the GQ population during the period from the last decennial census to the latest July 1 population estimates are made using empirical data, and after the latest estimates the forecasted rates or counts are used. Events are simulated for the population in households in the following order: births, deaths, in-migration, out-migration, and finally aging and residuals. In the final step, all individuals remaining in the register except new births and special populations are aged forward one year to simulate the process of aging. The GQ population is randomly resampled upward or downward to accord with DOF's estimated change in each county-level GQ population.

The next sections explain how births, deaths, and migrants are determined during the estimates window (2010-2016) and the projections window (2016-2060). The projection is run multiple times; by design, the results differ slightly with each iteration. The results provided in the public use dataset are the median of repeated simulation runs with a predetermined sequence of pseudorandom number seeds.

5.1. Birth projections.

The number of births during 2010-2016 is determined by the actual count of events reported in the state vital registration system. The California standard birth certificate does not include fields for identifying the race/ethnicity of children; in order to project the future race/ethnic distribution of the population, these traits must be inferred. The approach taken in the DOF projections is to relate child's race and ethnicity to the race and ethnicity of the child's parents, using the distribution of race/ethnicity of parents and their own children under 18 within primary families in the 2010 US Census PUMS (the 'parent link file').

When mother's and father's race and ethnicity are known, random draws from a uniform distribution are used to assign a race and ethnicity to a child, where the probability of each assignment is determined by the proportion of children with each unique combination of mother's and father's traits that are reported with each race or ethnicity. The proportions are calculated from the 2010 US Census PUMS using records from primary families in households where the mother, father, and children under 18 can be linked, based on the 'kid link file' approach used by the Census Bureau [5]. When mother's traits are missing in the vital statistics, they are imputed according to the procedure described above. When father's traits are missing, they are imputed using the proportion of mothers with a given set of traits who partner with men of each race/ethnicity in the parent link file.

For example, children from a mother age identified as Black, Non-Hispanic and a father who is White, Hispanic are not all assigned to Black, White, or multi-racial.[*] Such children might be assigned a 10% probability of identifying as White, a 60% probability of identifying as Black, and a 30% probability of identifying as multi-racial (White and Black). In addition, their child might have a 60% probability of being labeled Hispanic.[†]When the father's traits are unknown, we impute a prospective father by observing in the parent link file that 70% of such mothers partnered with a Black, Non-Hispanic father, 10% probability of a White, Non-Hispanic father, and so on until 100% of combinations are accounted for. Draws from a uniform distribution are used to assign missing father's details, and then subsequent draws are used to assign child traits given complete information about parents.

The number of births after July 1, 2016 is determined by repeated draws from a uniform distribution for all women age 15-49. A fertility rate is merged into the population dataset from the fertility projections (since the projections are run using fiscal years, adjoining calendar years are averaged). Draws below the age-specific fertility rate (ASFR) associated with each individual result in a simulated birth. For births after 2015, the fertility model only uses the traits of mothers to assign the traits of children. Characteristics of the father (race and ethnicity) are imputed using the proportion of mothers by race/ethnicity who partnered with fathers of each race/ethnicity. Random numbers are used to impute father's traits, and with complete parents' details additional random draws are made to assign child's traits, very similar to how traits are assigned to children from vital statistics records with missing father's data. A similar approach is used by the Census Bureau in its projections by race/ethnicity [5]. Children are assigned a slightly higher probability of male sex (51.2%) based on the assumption that the empirical sex ratio at birth continues indefinitely [14].

5.2. Death projections.

Deaths during 2010-2015 are subtracted from the population register by identifying records that closely match the recorded traits of the actual deceased, using data on the decedent's county of residence, age, sex, and bridged race.[‡] While death certificates contain multiple possible race responses, bridged race is used to reduce the likelihood of mismatch between deaths and exposure to risk (population size by race/ethnicity). After July 1, 2015, random draws from a uniform distribution are used to simulate mortality in the population using projected age-specific death rates (ASDR).

5.3. Migration projections.

Net migration is calculated by the DOF population estimates team using a variety of survey and administrative data sources. The DOF method of calculation does not generate granular age detail of net migrants. Instead, net migration is disaggregated into 5-year age groups using the methods described above ('Estimation and projection of migration'). For age groups with positive net migrants, observations are added to the dataset. For age groups with negative net migrants, some observations are dropped in order to simulate out-migration. For example, 6,000 net migrants might correspond to roughly 8,000 expected in-migrants and 2,000 expected out-migrants. Within these groups, there might be 1,200 net 35-39 year olds moving in, and 500 net 40-44 year olds who leave. The model would add 1,200 records randomly selected (without replacement) from the records

[*] Proportions described in this paragraph are hypothetical and do not refer to proportions observed in the parent link file.

[†] It is important to note that determinations of child race or ethnicity made in the Census are made by the survey respondent-- usually the household head-- on behalf of children in the household, and may not be entirely predictive of how children will self-identify in adulthood.

[‡] In cases where multiple individuals are matched, one is randomly selected; in cases where the deceased cannot be identified in the population register (often due to misspecification of age or race/ethnicity), an individual with the closest matching traits (e.g. matching sex and age within 5 years) is selected.

of past respondents from the ACS who were age 35-39 and reported moving in to the geography, and randomly drop 500 records of pre-existing residents age 40-44.

For age groups with net in-migration, traits of net migrants are generated by randomly replicating records of in-movers captured in the ACS PUMS. For future years where no ACS PUMS data are available, records are randomly drawn from a pooled sample of ACS PUMS movers since 2005. For groups with negative net migration, individuals of the same age are randomly moved out from the population register when the projection is run. Because the traits of immigrants and net in-migrants are determined by random draws from the cumulative ACS records of all in-migrants, the distribution of migrants by domestic or foreign origin is implicitly held constant at the same weighted average proportion observed during the decade 2005-2015.

6. Assumptions and limitations.

The projection models rely heavily on trends and relationships observed in the past. Although the repeated overlapping regression model design is inherently more robust than a single pooled model, it does rely on taking a measurement from a distribution of possible outcomes defined by a variety of relationships observed in the recent past, but none from the distant past (or the future, which may change in ways unanticipated by the past). Implicitly, this means that the results assume no radical change in the economic, policy, or natural environments. The projection is based on a model that has persistent below-replacement fertility and continued improvement in survival rates, typical of a post-demographic-transition society [10].

The projection assumes sufficient resources to support population growth (or the development of more efficient/productive technology). Changes in immigration, education, or transportation policy would have significant effects that are not considered here; likewise there is a risk of unforeseen changes in technology (especially reproductive technology and healthcare). The model is subject to several sources of bias in addition to those mentioned above: among them, bias from the use of discrete time scale, and bias from not modeling gross migration flows distinctly (i.e., foreign and domestic arrivals and departures separately), which could affect the age, race, ethnic, geographic, and gender distributions of the population. There is additionally the likelihood that the 2020 Census will use different race/ethnic categories than those used here, in which case the future race/ethnic distribution may not be directly comparable.

Migration is projected to increase in the near future but then to remain stable, whereas the previous projections saw a more significant rise and decline cycle that was projected to peak in the late 2030s. Our view is that California's strong economic performance and attractive climate means that the state will continue to attract net positive migration when international and domestic migration are combined, but high prices relative to median income will keep net migration from returning to the high rates witnessed in past decades. For quantitative comparisons with the previous projections [3], see the Appendix.

7. Public use datasets.

The resulting public use dataset (P-3) contains counts of the population for each California county for July 1 of every year from 2010 through 2060, by age (0-100+), sex, and race or Hispanic ethnicity. Summarized data are published as P-1 series (statewide) and P-2 series (county) projections. Additional public use data available on the Department of Finance website as part of the 2016 Baseline release include county total population and components of change (births, deaths, and net migration).

8. Authority.

The population projections were prepared under the mandate of the California Government Code (Cal. Gov't Code § 13073, 13073.5). It is state policy that all state plans make use of the "*. . . population projections and demographic data that is provided by the State's Demographic Research Unit*" (Cal. State Admin. Manual § 1100).

9. Acknowledgements

Research design, data collection, analysis and interpretation, technical report and dataset by Ethan Sharygin.
Reviewers: Walter Schwarm, Julie Hoang.

10. Suggested citation.

California Department of Finance. Demographic Research Unit. 2018. *State and county population projections 2010-2060* [computer file]. Sacramento: California Department of Finance. January 2018.

Methodology report: Sharygin, Ethan. 2018. *Modeling methodology for the 2016 baseline California population projections.* Sacramento: California Department of Finance. January 2018.

References

[1] Azur, M.J., E.A. Stuart, C. Frangakis, P.J. Leaf. 2011. "Multiple Imputation by Chained Equations: What is it and how does it work?" *Int J Methods Psychiatr Res* vol. 20 no. 1 pp. 40–49. doi: 10.1002/mpr.329

[2] Box, G., G. Jenkins. 1976. *Time Series Analysis: Forecasting and Control (Rev. Ed.)*. San Francisco: Holden-Day.

[3] California Department of Finance (DOF). Demographic Research Unit. 2014. *Forecasts of population, births and public school enrollment at the state and county level produced by the Demographic Research Unit (P-3).* [computer file]. Sacramento: California Department of Finance. December 2014. Retrieved 12/1/2016 from http://www.dof.ca.gov/Forecasting/Demographics/Projections/

[4] Dunn, H., ed. 1954. "History and organization of the vital statistics system," in *Vital Statistics of the United States, 1950 Volume I.* Washington, D.C.: U. S. Department of Health, Education, and Welfare. Public Health Service.

[5] Guarneri, C.E. and C. Dick. 2012. "Methods of Assigning Race and Hispanic Origin to Births from Vital Statistics Data." Paper presented at the Federal Committee on Statistical Methodology Annual Meeting in Washington, D.C. on January 12, 2012.

[6] Harvey, A. 1989. *Forecasting, structural time series models and the Kalman filter*. New York: Cambridge University Press.

[7] Ingram, D.D., J.D. Parker, N. Schenker, J.A. Weed, B. Hamilton, E. Arias, J.H. Madans. 2003. "United States Census 2000 population with bridged race categories." Vital Health Statistics vol. 2 no. 135.

[8] Kannisto, V. 1994. *Development of Oldest-Old Mortality, 1950-1990: Evidence from 28 Developed Countries.* Odense: Odense University Press.

[9] Kulkarni, S., A. Levin-Rector, M. Ezzati, C.J.L. Murray. 2011. "Falling behind: life expectancy in US counties from 2000 to 2007 in an international context." *Population Health Metrics* vol. 9 no. 16. doi: 10.1186/1478-7954-9-16

[10] Lesthaeghe, R. 2014. "The second demographic transition: A concise overview of its development." *Proc. Nat. Acad. Sci.* vol. 111 no. 51, pp. 18112–18115. doi/10.1073/pnas.1420441111

[11] Li, N. and S. Tuljapurkar. "The Probabilistic Life Table and Its Applications." Paper presented at 2013 Annual Meeting of the Population Association of America. Retrieved from paa2013.princeton.edu/papers/130118.

[12] National Center for Health Statistics (NCHS). 2016a. "Vintage 2015 postcensal estimates of the U.S. resident population 85 years and over (April 1, 2010, July 1, 2010-July 1, 2015), by year, single-year of age, bridged race, Hispanic origin, and sex" [computer file]. Retrieved from: http://www.cdc.gov/nchs/nvss/bridged_race.htm.

[13] National Center for Health Statistics (NCHS). 2016b. "Bridged-race intercensal estimates of the U.S. resident population 85 years and over for July 1, 2000-July 1, 2009, by year, single-year of age, bridged race, Hispanic origin, and sex" [computer file]. Retrieved from: http://www.cdc.gov/nchs/nvss/bridged_race.htm.

[14] Parazzini, F., C. La Vecchia, F. Levi, S. Franceschi. 1998. "Trends in male: female ratio among newborn infants in 29 countries from five continents." *Human Reproduction* vol.13 no.5 pp.1394–1396.

[15] Preston, S.H., P. Heuveline, M. Guillot. 2001. *Demography*. Malden, MA: Blackwell.