

# Household Projections Methodology

California State Department of Finance  
Demographic Research Unit

## Release notes

- Vintage 2025 (2025-11-7): initial public data release

The Demographic Research Unit (DRU) within the California Department of Finance (DOF) produces household projections for California and its 58 counties. These projections, reflecting the latest population projection series (2024 Baseline, Vintage 2025), combine machine learning and exponential smoothing techniques to forecast future households. This blended approach integrates official demographic projections with statistically derived trend continuations, yielding results that balance empirical accuracy with long-term stability.

Overall, the number of California households is expected to grow at an annualized rate of 0.6 percent between 2020-30 compared to 0.5 percent for the previous series, yielding approximately 50,000 more households in 2030 than previously projected. The increase is driven by an upgrade in the base year 2020 households of 15,000 and improved household formation rates over the last 4 years. While household growth is projected to slow somewhat between 2030-40 due to declines in 25-40 year-old population growth, it remains around an annualized 0.4 percent per year (~75,000 households).

## Modeling Framework

### 1. Machine Learning

Machine learning (ML) derives household structures directly from observed data, enabling the integration of diverse factors that influence household formation while reducing reliance on distributional assumptions. Previous literature demonstrates that ML methods improve accuracy [4], as they are designed to identify complex, nonlinear patterns that traditional techniques often overlook.

The household projections employ four widely used ML approaches: Random Forest (RF), Gradient Boosting (GB), Extreme Gradient Boosting (XGBoost), and Light Gradient-Boosting Machine (LightGBM). All four are tree-based, non-parametric approaches, meaning they do not assume any predefined functional relationship between inputs and outcomes.

Random Forest (RF) operates by randomly selecting subsets of explanatory variables at each split, constructing multiple independent decision trees. The underlying assumption is that averaging the predictions of uncorrelated trees reduces variance more

effectively than averaging similar models. For further theoretical and applied details, see Liaw & Wiener [5].

In contrast, boosting methods, including GB, XGBoost, and LightGBM, build trees sequentially, with each new tree focusing on correcting the errors of its predecessors. This iterative refinement allows boosting models to achieve higher predictive accuracy, particularly for complex datasets. For a comprehensive discussion of GB and XGBoost, refer to Chen & Guestrin [6].

LightGBM, a more recent advancement in boosting, improves efficiency and scalability by introducing novel techniques such as histogram-based splitting, leaf-wise tree growth, and gradient-based sampling [7]. These optimizations allow LightGBM to train faster and handle large datasets more effectively than traditional boosting methods, making it suitable for high-dimensional and computationally intensive tasks.

## 2. Exponential Smoothing

Exponential smoothing (ES) models were used to extend the projection horizon of ML models and extrapolate household counts from 2031 to 2040. This hybrid approach combines the short-term precision of ML methods with the long-term stability of time-series forecasting.

Specifically, double exponential smoothing or Holt's two-parameter method with linear and damped trends was utilized to expand ML projections [8,9]. ES damped trend models were explored because they have shown improved accuracy in empirical applications [9]. Forecasts from this hybrid ML-ES damped trend model were used for most of the counties. Since ES models rely heavily on historical time series data, adjustments were made for several counties to account for recent trend changes or atypical time series. For example, the projections for Los Angeles County were adjusted to reflect the impacts of the January 2025 Palisades and Eaton fires. In addition to ES point forecast, 80 percent prediction intervals were explored for model considerations.

## Data Sources

The data used for the machine learning portion of the projection series consists of historical datasets for model training and forecasted input datasets for generating predictions. The training data covers the period from 2010 to 2024, comprising 870 observations (58 California counties × 15 years).

Historical demographic and housing data form the foundation of the ML modeling approach. It incorporates DRU estimates for core variables including household counts, total population, fertility rates, death rates, net migration rates, and the percentage of new housing completions. Other demographic data — marriage rates, divorce rates, old-age dependency ratios, young-age dependency ratios, and group quarters population percentages — were calculated from the U.S. Census Bureau's American

Community Survey (ACS). In addition, the model incorporates unemployment rates from the Bureau of Labor Statistics (BLS) and the Housing Price Index (HPI) from the Federal Housing Finance Agency (FHFA) as key socioeconomic predictive features. The model also integrates indicators including median household income, educational attainment (measured as the proportion of population holding at least a bachelor's degree), and renter-to-owner ratio from the ACS.

For forecasting, multiple projection sources were employed to ensure comprehensive future scenario modeling. The latest population projection series provided the foundation for future total population estimates along with fertility, mortality, and net migration trends. The California Department of Transportation's Long-Term Socio-Economic Forecasts by County served as the source for projected unemployment rates.

ES model inputs included DRU historical household data as of Jan. 1 for 1990-2025 and ML final household projections for 2026-2030.

### **Context and Assumptions**

These household projections are produced for informational and planning purposes and do not fall under the mandate of Government Code, Sections 13073 and 13073.5, or State Administrative Manual, Section 1100. Forecasts of future housing and housing needs are the responsibility of the Department of Housing and Community Development and may use different assumptions and methods.

These projections are not a prediction of housing growth. They represent one modeled scenario derived from current demographic and socioeconomic assumptions. No conclusions are made about whether those assumptions will hold true and thus these projections should not be interpreted as the most likely outcome but rather one outcome given a set of assumptions. The modeling framework does not account for either the existence of pent-up demand in today's market nor potential further changes that may occur if economic and social conditions differ from those of the past. Because the projections reflect the impact of recent long-run trends and do not allow for any adjustments either upward or downward in response to changing economic conditions, actual household growth could deviate dramatically from the projected figures.

In determining household population, a distinction is made between the portion of the population living in households and those living in group quarters (i.e. military barracks, college dormitories, old-age institutions or prisons). The various components of the group quarters population are projected based on past trends, and where available, information about future facility growth or decline.

## References:

- [1] Mason, A., & Racelis, R. (1992). A comparison of four methods for projecting households. *International Journal of Forecasting*, 8(3), 509-527.
- [2] Spicer, K., Diamond, I., & Bhrolchain, M. N. (1992). Into the twenty-first century with British households. *International Journal of Forecasting*, 8(3), 529-539.
- [3] Wilson, T. (2013). The sequential propensity household projection model. *Demographic research*, 28, 681-712.
- [4] Şahinarslan, F. V., Tekin, A. T., & Çebi, F. (2021). Application of machine learning algorithms for population forecasting. *International Journal of Data Science*, 6(4), 257-270.
- [5] Liaw, A., & Wiener, M. (2002). Classification and regression by RandomForest. *R news*, 2(3), 18-22.
- [6] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).
- [7] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.
- [8] Hyndman, R.J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*, 2nd edition, OTexts: Melbourne, Australia. <https://otexts.com/fpp2/>
- [9] McKenzie, E. & Gardner, E. S. Jr. (2010). Damped trend exponential smoothing: A modelling viewpoint. *International Journal of Forecasting* 26 (4) pp. 661-665. <https://www.sciencedirect.com/science/article/abs/pii/S0169207009001277>